UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

"ANÁLISIS Y VISUALIZACIÓN DE RESULTADOS OLAP"

TESIS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN INGENIERÍA (COMPUTACIÓN)

PRESENTA:

EDGAR DURÁN MUÑOZ

TUTOR DE TESIS:

DR. JAVIER GARCÍA GARCÍA

FACULTAD DE CIENCIAS - UNAM

MÉXICO, D. F. JUNIO 2013

A todas las personas que de alguna u otra manera se vieron involucradas e hicieron posible que iniciara y concluyera esta etapa de mi formación culminando todos los esfuerzos en este trabajo, pero, principalmente le agradezco a mis padres **Remedios Muñoz Sarabia** y **Rosendo Durán Ramírez** por su apoyo incondicional.

G R A C I A S

Índice general

Capítulo 1

Introducción

Hoy en día se realizan diversas actividades encausadas al análisis de datos, para que estas tareas se lleven a cabo de manera más sencilla se cuenta con algunos métodos y herramientas. Hacer uso de técnicas OLAP (On-Line Analytical Processing o Procesamiento Analítico en Línea) es muy útil para explorar y analizar bases de datos, durante un análisis OLAP generalmente se exploran grandes cantidades de registros con múltiples niveles de granularidad y de esta manera se obtienen resultados interesantes.

El trabajar con OLAP en bases de datos implica manejar un modelo multidimensional, al pensar en una representación gráfica de los datos es común imaginar la información organizada en una estructura geométrica tal como un cubo. El objetivo principal del presente trabajo es mostrar una propuesta para generar una visualización en 3D que incorpore los resultados obtenidos, un cubo es la figura base para este modelo visual, además se proporcionan métodos para poder interactuar y manipularlo, de esta manera, para el usuario puede ser mucho más fácil detectar patrones en los resultados que se consideran significativos.

En aplicaciones actuales referentes a la visualización de resultados OLAP se trabaja con 1, 2 o 3 dimensiones, además dichas herramientas están enfocadas a resolver un problema en específico o solamente dentro de un campo de estudio. La herramienta aquí presentada brinda la posibilidad de trabajar con más de tres dimensiones, y adicionalmente, es posible utilizarla para realizar un análisis con diferentes datos almacenados no importando el campo de estudio al que pertenezcan, lo que se re-

quiere de dichos datos es que cumplan ciertos requisitos que se mencionarán en este trabajo.

Un punto más a tratar involucra a las funciones de agregación que se utilizan para la creación de cubos OLAP. De las aplicaciones que realizan este análisis de datos, la mayoría emplean operaciones que existen comúnmente en un manejador, tales como sumatorias, conteos, promedios, etc.. Por otro lado, se sabe que un análisis de tipo estadístico brinda evidencias y un sustento para los resultados que se puedan generar. Es por ello que en la presente investigación se proponen un grupo de 7 pruebas estadísticas para la tarea de exploración de datos, el motivo de usar específicamente las pruebas elegidas en este trabajo se debe a que se involucran ecuaciones matemáticas que pueden ser evaluadas eficientemente mediante consultas SQL utilizando algunas de las funciones de agregación antes mencionadas.

1.1. Antecedentes

En el campo de estudio de la visualización de resultados OLAP, existen algunas aplicaciones, las cuales se describen más adelante, que presentan propuestas interesantes para el análisis y como plasmar de los datos gráficamente. En particular, en [28] se abordó un caso de estudio donde se involucra el campo de la medicina, para realizar este artículo se contaba con registros acerca de la historia clínica de un número determinado de pacientes, pudiendo observar en ellos sus enfermedades y el número de personas que se encontraban en algún supuesto médico, con estos datos se llevaba a cabo un análisis estadístico para determinar si existía alguna relación entre los factores de riesgo y el problema médico.

En el caso mencionado se generaban consultas SQL hacia la base de datos para obtener la información médica. A través de dichas consultas se generaban resultados OLAP dependiendo de la selección de dimensiones y medidas de interés sobre los datos. El análisis fue realizado utilizando una prueba estadística en particular (diferencia entre medias poblacionales) para un caso de estudio (predicción de estenosis), para mostrar los resultados finales se creó una representación en 2D. Este desarrollo fue presentado en el decimocuarto taller internacional sobre DataWarehousing y OLAP (DOLAP 2011) con el título Interactive Exploration and Visualization of

OLAP Cubes en el mes de Octubre durante la conferencia número 20 sobre la información y gestión de conocimiento de ACM (CIKM 2011) presentada en Galsgow, Escocia, UK.

1.2. Objetivo de la tesis

El objetivo fundamental de la presente investigación es proponer innovaciones para generar un sistema de visualización de cubos OLAP tomando como base el prototipo presentado en [28], además que pueda tener aplicabilidad en cualquier ámbito de la vida humana diaria y de esta manera no enfocarlo a un caso aislado.

Con el objetivo anterior, se proponen procedimientos basados en diferentes pruebas estadísticas, con el fin de complementar el análisis de datos OLAP, obteniendo además una visualización en 3D de los resultados esperados, de igual forma se proponen métodos de apoyo para la fácil interpretación e interacción con los datos que se obtengan. Con lo cual se podrá tener un uso general de la herramienta OLAP para aplicarse en distintos campos y contextos.

1.3. Aportaciones

En particular se aporta lo siguiente al campo de herramientas enfocadas al análisis y visualización de resultados OLAP.

- Generación de una aplicación que pueda ser empleada en diversos campos del conocimiento donde se vea involucrado un análisis de datos y se requiera una visualización clara de resultados. Cabe aclarar que el prototipo inicial presentado en [28] tenía un enfoque médico únicamente para la predicción de estenosis (En medicina, estenosis o estegnosis es un término utilizado para denotar la constricción o estrechamiento de un orificio o conducto corporal. Puede ser de origen congénito o adquirido por tumores, engrosamiento o hipertrofia, o por infiltración y fibrosis de las paredes o bordes luminales o valvulares.).

- La visualización de los datos normalmente se genera en un formato de 2D, con la presente investigación se logra una aportación a un modelo en 3D desarrollando

mecanismos adicionales de navegación entre los datos.

- La base del análisis cuenta con un fundamento estadístico, inicialmente se contaba con sólo una prueba de hipótesis, por lo que se implementaron 6 pruebas adicionales.

- Se realizó la comparación de los resultados obtenidos con el sistema contra los resultados publicados de un problema real tratado en una investigación nacional de salud relacionada con personas que padecen de diabetes y tuberculosis.

1.4. Metodología

Para lograr este cometido se desarrolló la presente investigación. El trabajo se estructuró de la siguiente forma:

En el primer capítulo se presenta una breve introducción para dar a conocer los objetivos y aportaciones así como algunos trabajos relacionados con el tema tratado.

Posteriormente en el segundo capítulo se tocarán los conceptos sobre estadística, los elementos que la componen y las operaciones que se realizan dentro de ella, así como los objetivos que se persiguen al utilizarla como medio de obtención de información sobre alguna circunstancia de la vida.

En el tercer capítulo trataremos las cuestiones referentes a almacenes de datos relacionadas con la tecnología OLAP.

En el cuarto capítulo se revisarán métodos utilizados en investigaciones relacionadas con la visualización OLAP y se mostrará la metodología utilizada empleando SQL para la generación de cubos OLAP.

En el quinto y último capítulo establecemos concretamente las innovaciones que se proponen a la herramienta OLAP a través de dos experimentos, comenzando por la constitución de un cubo, las dimensiones que lo conforman, así como el análisis de sus datos, y la presentación final de resultados, esto es, mejoras en cuanto a la visualización para una mejor apreciación y entendimiento de lo que los datos nos quieren decir, en este mismo capítulo se mostrará el uso y aplicación que puede tener y los beneficios al utilizar las nuevas implementaciones.

Dentro de esta investigación se puede visualizar las innovaciones que se proponen en la herramienta OLAP, utilizando un sistema que puede tener aplicación en diferentes problemas de la vida diaria, se piensa que una herramienta OLAP puede aplicarse en cualquier campo en donde existan datos estadísticos, de ahí que la presente investigación tenga como propósito analizar y visualizar estos datos a través de la tecnología OLAP.

1.5. Trabajos Relacionados

Actualmente se necesita contar con información que sea oportuna, la cual no sea estática, que esté centralizada, además de poder acceder a ella para realizar análisis y emplearla para la toma de decisiones cuando se requiera. Los almacenes de datos en la mayoría de los casos son implementados con el modelo multidimensional OLAP y es a través del uso de ellos que se puede facilitar la exploración, el punto fundamental de este método es la selección de medidas y dimensiones.

En cuanto a la presentación de la información es imprescindible el uso de tecnologías, mismas que exhiban los datos en diferentes tipos de visualizaciones a través de gráficos o mapas para que el investigador cuente con una visión global, detallada y filtrada de los datos y en consecuencia tome mejores decisiones. La manera en que la información se presenta para ser analizada es de gran importancia al momento de obtener resultados. Para realizar una investigación de la mejor manera posible, el usuario debe observar los datos, explorarlos y que le sean claros, por esta razón es importante que las herramientas que se utilicen, además de mostrar los datos, permitan manipularlos para posteriormente llegar a una conclusión.

Existen variedad de trabajos relacionados al tema de OLAP, enfocados al manejo y análisis de los datos en las etapas de almacenamiento o del tratamiento de estos mismos, de igual forma se habla sobre técnicas para mejorar el rendimiento en las consultas o al momento de manejar grandes volúmenes de datos, como se mencionó anteriormente, el objetivo está involucrado un tanto con la visualización de los resultados, es por ello que a continuación se mencionan algunos artículos relacionados con técnicas de visualización o de presentación de resultados OLAP.

En [40] se emplea un software llamado CommonGIS, se trata de una herramienta

gráfica que está enfocada a la visualización y exploración de datos de acuerdo al espacio y tiempo. Utiliza gráficas interactivas, métodos para clasificaciones dinámicas, criterios de decisión y análisis de series. En cuanto a la visualización de resultados, se enfoca en un ejemplo acerca de la venta de productos, toma en cuenta tres dimensiones (tipo de producto, período de tiempo, lugar) y una medida (ventas). Como resultado, muestra una gráfica similar a una gráfica de barras en donde cada barra representa un lugar de localización, a su vez, cada barra está dividida en pequeñas secciones dependiendo de la cantidad de productos, cada uno se identifica dentro de la barra por un color diferente, el tamaño de la división en la barra es proporcional al total de ventas del producto, como período de tiempo para la gráfica se toma un mes. Adicionalmente, se muestra otra gráfica de líneas donde se muestran datos similares a los mencionados en la gráfica anterior, la diferencia es que se muestran diversas líneas y cada línea representa un mes diferente, de esta manera se muestran 13 meses diferentes. Esta herramienta esta enfocada a proveer información útil para los negocios. En cuestión de los datos, toma en cuenta únicamente tres dimensiones y una medida, la información que se obtiene con las medidas es a través de funciones de agregación comúnmente utilizadas (conteos, sumatorias, promedios). Se presentan gráficas de barras y de líneas en 2D representando las tres dimensiones y una medida. En particular el software está enfocado al análisis de ventas en diferentes regiones, por lo que se muestra un mapa para identificar la localización de los distintos puntos de venta. En contraste, en el trabajo que aquí se presenta, la herramienta no es para un problema en específico, esta abierta a más posibilidades, las dimensiones no están limitadas a 3 como en este caso.

Los autores de [16] proponen un método para crear una visualización 3D de cubos de datos. Extrae de la base de datos tanto dimensiones como medidas, estas son utilizadas para construir una tabla en 2D en la cual, la primera fila y la primera columna contienen las dimensiones con las que se trabaja, al centro de la tabla, en los cruces de filas y columnas son ubicadas las medidas. Cada cruce puede ser seleccionado, y a partir de estos datos es como se crea el cubo. Al seleccionar un cruce en la tabla se tienen dos dimensiones y una medida, a partir de estas se crea un cubo en 3D, un eje se la dimensión 1, otro eje es la dimensión 2, y el último eje es la medida, el cubo analizado está formado por muchos cubos pequeños llamados dados. El pro-

grama brinda dos operaciones para manipular el cubo, una consiste en seleccionar un plano completo, y la otra en elegir un renglón en específico, en cualquiera de los casos posteriormente es posible escoger un solo dado, con el dado seleccionado los resultados pueden ser mostrados como una gráfica de barras, de líneas o de pastel. Las diferencia con la herramienta que presentamos en este trabajo está en que las funciones de agregación utilizadas son las comunes en un sistema manejador de bases de datos y en el cubo creado, dos ejes son utilizados para representar dos dimensiones y el tercer eje involucra la medida.

En [22] se describe un modelo para la presentación de los cubos de datos denominado CPM (Cube Presentation Model), este modelo es una propuesta de presentación para herramientas OLAP la cual se compone de una capa lógica y una de presentación. En particular se centra en como mapear la capa de presentación a partir de técnicas alternativas para la visualización. Adicionalmente se muestra cómo combinar CPM con un modelo para la visualización denominado Table Lens, una técnica que permite visualizar y manipular grandes tablas en 2D. Se crea una tabla que asume dos ejes por lo que se utilizan filas y columnas. Se construye un espacio en 2D a partir de los dos ejes, se trata de una matriz de celdas. Cada punto es ubicado en los ejes con un valor que indica el nivel de interés. Cada celda tiene asociada una función de transferencia para poderla ubicar físicamente en el cubo que se había planteado. La ventaja que presenta este modelo respecto a la capa lógica y de presentación, es la flexibilidad de poder cambiar alguna de las capas por alguna propuesta alternativa en algún momento. El modelo no especifica un límite para las dimensiones y cada una de las celdas representa el cálculo o la función de agregación que se aplica para obtener su contenido. Se hace una abstracción de los datos a manera de un cubo, además de proyectar capas y cruces entre filas y columnas.Para la visualización se toma como modelo una técnica que emplea dos dimensiones, filas y columnas, el ingrediente extra en esta matriz es que mediante una función de transferencia es como se puede obtener la posición que tomaría una celda en un espacio de más dimensiones. Este es un trabajo que aborda la manera en cómo poder formar un cubo a partir de una tabla, sin embargo, al ser la tabla una estructura plana, sólo se pueden representar dos dimensiones.

En el trabajo [35] se presenta la dificultad de leer grandes volúmenes de datos,

es por ello que emplean data mining, data warehousing y técnicas de visualización espacial para descubrir patrones implícitos y las relaciones en los datos. Se propone el sistema de visualización denominado CubeView. En particular se trata la diferencia entre un data warehouse convencional y un data warehouse espacial, la cual consiste en que el primero muestra tablas de resumen con texto y números, y en el segundo pueden ser colecciones de mapas en los cuales se pueden identificar patrones. Se describe un ejemplo de aplicación utilizando como medida el número de autos que fluyen, y como dimensiones, la estación del año, el día de la semana y la hora del día. Los resultados son mostrados en 3 planos, para cada uno se muestra una gráfica de la relación que existe entre las dos dimensiones involucradas, además se observan comparaciones de mapas seleccionando una estación, o día u hora diferentes. En este caso, la recopilación de datos se hace a través de imágenes satelitales durante períodos de tiempo definidos, se puede decir que la función de agregación tiene que ver con la identificación de patrones en las imágenes, en este caso se identifican patrones respecto al flujo del tráfico, se obtiene su comportamiento a través del tiempo, horas, días de la semana y estaciones del año. Para la visualización se consideran tres dimensiones, aunque de acuerdo con las definiciones en algunos otros artículos, se puede ver como una sola dimensión(tiempo) para la cual se hace una jerarquización (estaciones del año, días de la semana, período de tiempo en minutos). Se visualiza un cubo a la mitad, ya que sólo son tres planos, cada eje representa una escala diferente, los patrones identificados son presentados para cada uno de los planos formados en el cruce de los ejes, y es similar a una gráfica de un espectro. El uso de colores a través del espectro representa los niveles de actividad o complicación en dichas zonas. Así como en los casos anteriores, el uso de tres planos involucra solamente tres dimensiones, además de estar enfocado explícitamente en un problema donde trabaja con reconocimiento de patrones.

En el artículo [23] se presenta la exploración de cubos multidimensionales utilizando técnicas de visualización jerárquicas, se hace énfasis en el uso de la función en cubos OLAP drill-down. Se propone el uso de estructuras denominadas árboles de descomposición, en donde cada nivel es creado por una desagregación a lo largo de una dimensión. A diferencia de otros trabajos, al no generarse un cubo aparentemente no se realizan cruces entre dimensiones, sin embargo al formar las jerarquías

se puede decir que se obtienen las combinaciones entre los niveles, ya que en el nivel más bajo se tienen los mismos valores para cada desagregación, de acuerdo a las jerarquías de los datos se obtienen gráficas dependiendo del nivel en que se encuentren. Los datos en la base son descompuestos en dimensiones y medidas, y a su vez en cada dimensión se identifican los diversos niveles que pueden encontrarse. La desagregación se forma en una estructura de árbol, para el análisis deseado con las medidas pueden usarse funciones de agregación conocidas tales como la suma, promedio, etc.. En cuanto a la visualización, al ser acomodados los datos de una manera jerárquica, los datos se representan de manera similar a como se muestra la navegación dentro de las carpetas en algún sistema operativo con entorno gráfico, podemos tener una carpeta a nivel general que represente un período de tiempo, adentrándonos en la carpeta se puede desglosar en años y así sucesivamente en semestres, meses, semanas, etc.. La forma de representación de los datos es diferente, dado que se emplean estructuras de árbol y no se forma el cubo como se ha venido planteando, de igual forma al momento de formar las desagregaciones se utilizan funciones de agregación provistas por el manejador de bases de datos.

Finalmente en [41] también se presenta una visualización interactiva donde se hace uso de jerarquías en una estructura de árbol, además se representan los cubos de datos utilizando dos dimensiones. Mediante operaciones de drill-down y roll-up se pueden explorar de manera interactiva los niveles jerárquicos y explorar todos los datos. Una característica adicional es que se muestran distintos cubos de datos al mismo tiempo. Los cubos de datos son definidos como operaciones que pueden ser aplicadas a un grupo de valores obtenidos de una base de datos. Se puede visualizar la agrupación de los datos como una tabla con N columnas, el número de columnas depende de las características o dimensiones que se tengan en la base de datos. Si la dimensión 1 $D1$ contiene 3 posibles valores, es decir, es de cardinalidad 2, $D2$ es de cardinalidad 3 y $D3$ de cardinalidad 4 el número de valores en el cubo de datos será $2 \cdot 3 \cdot 4 = 24$. Tres atributos para un cubo de datos generarán 23 grupos de agregación, es decir que el número de grupos de agregación será 2 elevado a una potencia igual al número de atributos. La tabla mencionada anteriormente se descompone en un grafo con estructura de árbol donde se forman nuevas tablas agrupando los atributos y generando valores con los datos y las funciones de agregación. Para la muestra

visual de los resultados, se utilizan páneles, un panel de control, un panel de cubos, panel de nodos y panel de datos. En el panel de control se selecciona la base de datos, de igual manera se selecciona la forma para visualizar los nodos. El panel de cubo muestra los cubos de datos en la estructura de árbol, cada nodo puede ser representado por una gráfica de pastel o de barras, cada barra o en su caso cada sector del pastel representa el número de atributos en el cuboide, el tamaño de un sector o el alto de una barra es proporcional al porcentaje del valor que representa. Los nodos pueden ser seleccionados individualmente, al elegir alguno, en el panel de nodos se muestra un acercamiento de la gráfica, debajo de la gráfica se muestra un resumen de los valores obtenidos con las funciones de agregación. El panel de datos se utiliza para mostrar los valores de las funciones de agregación de acuerdo a una agrupación seleccionada, esto es, al momento de seleccionar un nodo en específico se hace una agrupación de los datos con esas características, se obtienen los resultados aplicando la función de agregación y es lo que se muestra en esta sección. En este trabajo se mencionan dos operaciones involucradas con OLAP (drill-down y roll-up) debido al manejo de jerarquías es la manera en como se descompone una dimensión, podría pensarse que en el presente trabajo al manejar un cubo solamente se tiene la opción de 3 dimensiones, sin embargo, se trabajaron opciones para poder revisar las diferentes dimensiones en caso de seleccionar más de 3, así mismo, existe la posibilidad de extender el tamaño del cubo hacia los 3 ejes dependiendo del número de elementos que conformen la dimensión en cuestión dificultando la visión completa de los cuboides de la parte central, es por ello que existe la posibilidad de manipular estos cuboides con movimientos horizontales y verticales.

Capítulo 2

Definiciones y conceptos de estadística

En este capítulo se tratarán conceptos relacionados con estadística, debido a que para aplicar una prueba de hipótesis hubo que seguir pasos determinados, debemos conocer los fundamentos que existen detrás de este procedimiento para así poder planear adecuadamente dicha prueba y aplicarla debidamente.

2.1. Tipos de estadísticas

La estadística no es completamente lineal, pues de acuerdo con J. Hurtado [13], es posible clasificarla en varias categorías. La primera de ellas se refiere a la Estadística Descriptiva o Deductiva, que se ocupa de organizar, resumir y presentar los datos en una forma conveniente e informativa. Este tipo de estadística se utiliza con el propósito de recolectar, describir y resumir en un conjunto de datos obtenidos, que es posible visualizarlos de manera numérica y gráfica; sin embargo, su uso se limita sólo al uso de la información obtenida. Sus resultados no son de gran utilidad para realizar ningún tipo de generalización.

La estadística inferencial o inductiva procede de manera contraria a la anterior, ya que tiene la particularidad de que a partir de datos muestrales es posible llegar a conclusiones y predicciones que abarquen toda una población. Son más precisos sus resultados y por lo tanto, son de gran utilidad para ser extrapolados y, de esta forma

realizar un pronóstico inclusivo. En este caso, las inferencias pueden presentarse a través de respuestas a preguntas simples del tipo sí/no, relaciones entre una serie de variables, estimaciones numéricas, etc. Es un método comúnmente utilizado para obtener conclusiones o inferencias acerca de determinada característica de una población, fundamentados en una muestra que permite interpretar información simple para llegar a conclusiones. Puede volverse necesaria cuando se requiere afirmación sobre otros elementos que se están midiendo, aunque no ofrecerá seguridad absoluta puede ofrecer una respuesta probabilística. Normalmente estima las características de un grupo total (población), basados en datos de un conjunto pequeño (muestra), cuyo propósito es extraer conclusiones acerca de la naturaleza de una población, situación u objeto.

La estadística aplicada, que está conformada por las dos clases estadísticas mencionadas anteriormente, la conjunción de estas dos permite alcanzar mayores objetivos que consisten en deducir resultados sobre un universo, partiendo de una muestra determinada, es decir, es más extensivo. Este tipo de estadística puede ser aplicada a cualquier área aunque no pertenezca a ella como historia, psicología, educación, medicina, etc.

La estadística matemática se refiere al empleo de ésta desde un punto de vista formal, a través del uso de distintas ramas propias de la matemática, vinculadas con la teoría de la probabilidad, que las vuelven más sustentables y garantizan con mayor efectividad la investigación. Su uso se hace necesario debido a que los datos manejados en la estadística matemática suelen ser aleatorios e inciertos.

Estos tipos de estadística sirven para solucionar o resolver diferentes problemas, utilizando técnicas de estimación o prueba de hipótesis. En ambos casos se busca generalizar la información obtenida de la muestra de una población cuya exigencia de técnicas sean aleatorias.

En este sentido B. C. Martínez [24] refiere que en los estudios estadísticos existen dos conceptos que son fundamentales, la población o universo y la muestra. La primera se refiere al proceso de medir todos y cada uno de los integrantes de un problema o situación particular considerada, generalmente es muy grande y en la mayoría de las veces resulta prácticamente imposible cuantificarla por sus dimensiones. Lo cierto es que una medida descriptiva sobre una característica de la población se le conoce

con el nombre de parámetro.

La muestra se refiere al subconjunto de la población total, es decir, es cuando se observa solamente una parte de la población. Las muestras representativas de ésta última son útiles pues facilitan el manejo de los datos, una muestra puede ser representativa de la población si al escogerla cada elemento tiene la misma probabilidad de salir o de ser escogido. Por lo que se considera que una propiedad importante de la muestra es su representatividad, ya que se entiende que posee las características importantes de la población en la misma proporción en la muestra. En este mismo contexto, debe ser mencionado que al proceso de estimar una característica de la población, basada solamente en los resultados de una muestra se le conoce como inferencia estadística.

Por las características de los datos recolectados y almacenados en las bases de datos, para el análisis realizado en este trabajo se hace uso de la estadística inferencial, ya que a partir de los datos muestrales se obtienen conclusiones para una población.

2.2. Elementos de la estadística

Los elementos que normalmente se requieren para llevar a cabo la tarea en el proceso de una investigación con carácter científico de estadística, suelen estar relacionados con las necesidades que exige ésta para su elaboración, teniendo en cuenta que cada elemento es una unidad utilizada para el estudio, Howard B. Chistensen [5] refiere que entre los principales están:

Datos numéricos: Son obtenidos mediante la medida de una característica o propiedad que tienen los objetos de interés (personas o cosas); sobre los cuales se realizan las medidas denominadas unidades experimentales u observaciones.

Unidades experimentales u observaciones: Es representada por las características de la población o muestra y varía de una a otra, normalmente son consideradas variables en el contexto estadístico.

Variables: Son valores que hacen posible observaciones para poder establecer rangos que normalmente varían de 0 a 100 puntos.

Estimado: Es un dato estadístico que se encuentra en función con la muestra, empleado para estimar un parámetro desconocido de la población. Se utiliza normalmente para establecer la media.

Estimación puntual: Es el valor estadístico calculado a partir de la información obtenida en la muestra, que se usa para estimar el parámetro poblacional.

Parámetro: Se utiliza cuando existen varios estimados diferentes, tomando en cuenta el que posean mejores propiedades como insesgadez, eficiencia, convergencia y robustez.

Intervalo de confianza: Es un valor que se establece comúnmente con el objetivo de localizar el parámetro poblacional.

2.2.1. Hipótesis

Una hipótesis es una proposición que se establece acerca de una o más poblaciones. Estas proposiciones generalmente se generan en los supuestos siguientes[25]:

- Supuestos acerca de las características de los datos con los que se va a trabajar: independencia de las observaciones, homogeneidad de los datos, etc.
- Supuestos acerca de la forma de la distribución de partida: binomial, normal, etc.
- Supuestos acerca de los parámetros de la distribución ya que se conoce su forma.

En este trabajo trataremos con dos tipos de hipótesis:

H_0: Hipótesis nula, se refiere siempre a un valor específico del parámetro de la población, no a una estadística de muestra. La letra H significa hipótesis y el subíndice cero no hay diferencia, aunque normalmente algunos estadísticos utilizan una "n", que indica"no hay cambio". La conformación de la hipótesis nula representa una afirmación que no es rechazada a menos que los datos muestrales proporcionen evidencia convincente de que es falsa. El planteamiento de

la hipótesis nula siempre contiene un signo de igualdad con respecto al valor especificado del parámetro.

H_a: Hipótesis alternativa, es cualquier hipótesis que difiera de la nula, es una afirmación que se acepta si los datos muestrales proporcionan evidencia suficiente para determinar que la hipótesis nula es falsa. El planteamiento de la hipótesis alternativa nunca contiene un signo de igualdad con respecto al valor especificado del parámetro.

2.2.2. Prueba de hipótesis

En muchos problemas, la finalidad es decidir si se acepta o se rechaza un enunciado acerca de algún parámetro. Al enunciado se le llama hipótesis o prueba estadística, y al procedimiento que se realiza para tomar decisiones acerca de la hipótesis se le llama prueba de hipótesis. Este es uno de los temas más útiles de la inferencia estadística, ya que muchos problemas en donde esta involucrada la toma de decisiones, pruebas o experimentos pueden tratarse como problemas de prueba de hipótesis.

El propósito principal es lograr hacer una elección adecuada entre dos hipótesis que se refieren a valores de parámetros poblacionales y haciendo la suposición de que la distribución de los datos es de algún tipo conocido, en este caso se dice que se hace una prueba entre parámetros.

Toda prueba implica la comparación entre dos hipótesis. Como resultado, el investigador tomará una decisión. Para realizar las pruebas, se establecen dos hipótesis mutuamente excluyentes y complementarias, estas son, hipótesis nula e hipótesis alternativa. Desde el punto de vista estadístico, el establecer una hipótesis como nula y la otra como alternativa no es lo mismo que elegirlas al contrario. La hipótesis nula muchas veces es utilizada como estrategia o un método del investigador para poder probar la alternativa. Al momento de establecer la hipótesis, conviene que la hipótesis nula sea aquella que contenga el signo de igualdad, si es que este signo aparece en alguna de ellas. Una hipótesis alternativa, se consigue demostrando (con los datos de la muestra como evidencia) que la hipótesis nula, es falsa. Por lo tanto, una teoría se comprueba al demostrar que no hay evidencia que sustente la teoría opuesta, se realiza una prueba por contradicción. Aunque con frecuencia se habla de probar una

hipótesis nula, el objetivo de la investigación por lo general es el de presentar evidencias, que se puedan justificar, a favor de la hipótesis alternativa. En el proceso se plantea la comparación de dos estadísticos entre sí, o entre un estadístico y un valor poblacional ya determinado. Los métodos de prueba de hipótesis permitirán decidir entre las dos hipótesis formuladas previamente, con un determinado nivel de error.

La prueba de hipótesis esta basada en la información que proporciona la muestra. La terminología de rechazar o aceptar una hipótesis debe quedar clara, es decir, el término de rechazar una hipótesis significa concluir que es falsa, mientras que aceptar una hipótesis solamente nos dice que no se tiene suficiente información como para rechazarla. Al establecer las hipótesis que se refieren a los valores que puede tomar un parámetro poblacional, se debe distinguir entre hipótesis simples y compuestas. En una hipótesis simple sólo se especifica un valor para el parámetro poblacional, las hipótesis compuestas, por el contrario, establecen un rango de valores que puede tomar el parámetro poblacional [25].

El proceso completo para llevar a cabo una prueba de hipótesis puede resumirse de la siguiente manera:

1. Exposición de la hipótesis nula y alternativa. Por conveniencia estadística, se toma como hipótesis nula la que contenga el signo igual, si es que este signo aparece en alguna.

2. Establecimiento de un nivel de significación α. Usualmente se toma el valor $\alpha = 0.05$.

3. Cálculo de la probabilidad de que nuestros resultados puedan haberse obtenido bajo la hipótesis nula, es decir, cálculo de la significación muestral p de la hipótesis nula. Esto implica una selección de la prueba estadística adecuada y la ejecución de la misma prueba.

4. Toma de decisiones, para lo cual se procede como sigue:

   - Se acepta la hipótesis nula si $p > \alpha$.
   - Se acepta la hipótesis alternativa si $p \leq \alpha$.

5. Conclusiones de tipo estadístico.

6. Conclusiones de naturaleza no estadística, propias de la aplicación en estudio.

2.2.3. Nivel de significación

Es la probabilidad de rechazar la hipótesis nula cuando es verdadera. ésta es denotada mediante la letra griega α, también conocida como nivel de riesgo, un término más apropiado, ya que se corre el riesgo de rechazar la hipótesis nula cuando pueda resultar verdadera. Para lograr el nivel de significación, la distribución de muestreo de la estadística de prueba se divide en dos regiones, la primera de rechazo (región crítica) y una más de no rechazo (región de aceptación).

El primer problema que surge es determinar en qué umbral se fija la probabilidad α, a partir de la cual se decide si las diferencias son significativas o no. Usualmente este valor, conocido como nivel de significación, y que se representa por α, se toma en $\alpha = 0.05$, esto es en el 5 %, significa que en el caso de ser verdadera la hipótesis H_0 sólo en un 5 % de las veces, o menos, las diferencias observadas serán lo suficientemente grandes como para poder llegar erróneamente a la conclusión H_a. En ocasiones, se toman niveles de significación más exigentes, como del 1 % ($\alpha = 0.01$) o incluso del uno por mil ($\alpha = 0.001$), con el objetivo de tener más confianza en los descubrimientos que se hagan con el nivel de significación elegido. La probabilidad p de que las diferencias observadas sean debidas exclusivamente al azar se conoce con el nombre de significación muestral de la hipótesis nula [17].

2.2.4. Regiones de aceptación y crítica

Al llegar a este nivel de la investigación, es necesario establecer las condiciones específicas en la que se rechaza la hipótesis nula y las condiciones en que no se rechaza la misma. Esto se debe a que la región de rechazo define la ubicación de todos los valores que son tan grandes o tan pequeños que servirán para definir la probabilidad de confirmar si la hipótesis nula es verdadera.

Para aceptar o rechazar una hipótesis, se necesitan reglas de decisión. Estas reglas se determinan por una región de aceptación y región de rechazo (región crítica) esto se refiere a la división del espacio muestral de valores del estadístico en dos zonas exhaustivas y excluyentes. El conjunto de resultados que llevan a la aceptación de la

hipótesis nula H_0 se denomina región de aceptación, y el conjunto complementario de resultados que llevan a rechazar la hipótesis nula, y por lo tanto a aceptar la alternativa H_a, se conoce como región crítica. Es muy frecuente que la decisión se fundamente en la observación de un estadístico, de tal manera que según los valores que tome se acepte o rechace la hipótesis nula.

Hay dos modos, que son equivalentes entre sí, para tomar la decisión respecto de las hipótesis formuladas, una es a través de la significación muestral de la hipótesis nula, comparando p con el nivel de significación α, y la otra utilizando tablas estadísticas para comprobar si su valor observado cae dentro de la región de aceptación o de la región crítica [17].

2.2.5. Tipos de errores

Durante el desarrollo de este proceso pueden presentarse distintos tipos de errores, sin embargo, J. R. Jiménez [17] señala que los errores fundamentales que pueden afectar totalmente el desarrollo del proceso son los siguientes:

Error tipo I: Se presenta si la hipótesis nula H_0 es rechazada cuando es verdadera y debía ser aceptada. La probabilidad de cometer un error tipo I se denomina con la letra alfa α.

Error tipo II: Se denota con la letra griega β, se presenta si la hipótesis nula es aceptada cuando de hecho es falsa y debía ser rechazada.

Al efectuar una prueba, nunca se puede estar completamente seguro acerca de la decisión que se tomó. Si se trabaja con un nivel de significación $\alpha = 0.05$, esto significa que se rechaza la hipótesis nula H_0 siempre que su probabilidad de explicar las diferencias observadas en los resultados sean de un 5 % o menos; por lo que la mayor parte de las veces se toma la decisión correcta, pero, en ocasiones, se puede rechazar la hipótesis nula H_0 siendo verdadera (Error tipo I). La probabilidad de cometer este error es a lo más del 5 % o el nivel α que se haya fijado. Para atacar los errores de tipo I se puede elegir un nivel de significación más riguroso, por ejemplo $\alpha = 0.01$ o $\alpha = 0.001$, pero, como ahora será más fácil aceptar la hipótesis nula, esta se aceptará más veces aunque sea falsa, (Error tipo II). Por lo tanto, si se desea reducir

el error de tipo I, entonces se incrementa el error de tipo II, y viceversa, la única forma de reducir ambos simultáneamente es tomando muestras de tamaño cada vez mayor.

Existe también el error de estimación, que es la diferencia entre la estadística muestral y el parámetro poblacional.

2.3. Pruebas de estadística

Son métodos para obtener un valor determinado a partir de la información muestral que se utiliza para determinar si se rechaza la hipótesis nula. Es necesario mencionar que existen muchos estadísticos de prueba, que la elección de uno de estos depende de la cantidad de muestras que se toman, si las muestras son iguales a 25 o más se utiliza el estadístico z, en caso contrario se utiliza el que resulta apropiado a la muestra.

Cabe señalar que para trabajar con algunas de las pruebas estadísticas presentadas a continuación se hace la suposición de que los datos presentan una distribución normal, sin embargo, es necesario realizar el análisis correspondiente para asegurar que la muestra utilizada presenta la distribución supuesta, en el Apéndice A se muestra un ejemplo de cómo realizar esta comprobación.

2.3.1. Prueba de hipótesis para la media de una sola población

ésta comienza con una suposición (hipótesis) que se hace acerca de un parámetro de población, posteriormente se recolectan datos de muestra con los que se producirán estadísticas muestrales, empleando ésta información para decidir qué tan probable es que el parámetro poblacional referencia sea el correcto.

Durante el procedimiento se determina un cierto valor estableciendo una media de población. Una vez recolectados los datos de muestra se procederá a determinar la diferencia entre el valor hipotético y el valor real de la media de la muestra, obtenidos estos resultados, es posible establecer si la diferencia obtenida es significativa o no. Mientras más pequeña es la diferencia, mayor es la probabilidad de que el valor

hipotético declarado al inicio para la media sea el correcto; en cambio mientras mayor sea la diferencia, más reducida será la probabilidad de que sea verdadera.

En estos casos, se debe tener en cuenta que la diferencia entre el parámetro de población hipotético y la estadística real, rara vez es tan grande que permita tomar la decisión inmediata de rechazar automáticamente la hipótesis inicial, esto es muy importante, ya que en las pruebas de hipótesis, como en la mayoría de las decisiones importantes de la vida real, las soluciones claras o bien definidas son la excepción, no la regla.

Ejemplo:

Un grupo de investigadores está interesado en conocer la edad media de cierta población. Se plantean lo siguiente: ¿Se puede concluir que la edad media de la población es diferente de 30 años?.

Con una muestra aleatoria simple de 10 individuos se tiene que la media $\bar{\mu} = 27$.

La población tiene una varianza conocida de $\sigma^2 = 20$.

Las hipótesis quedan planteadas de la siguiente manera:

H_0: $\mu = 30$
H_a: $\mu \neq 30$

La regla de decisión indica que H_0 se rechaza si el valor calculado de la estadística de prueba cae en la región de rechazo.

Se quiere que la probabilidad de rechazar la hipótesis nula sea $\alpha = 0.05$. Dado que es una prueba de dos colas, el valor α se divide en 2 partes iguales $\alpha/2 = 0.025$, una mitad se asocia con valores pequeños y la otra mitad con valores grandes.

Se plantea la pregunta: ¿Cuál es el valor de z a la derecha del cual está 0.025 del área bajo la distribución normal estándar?.

El valor de z a la derecha del cual está 0.025 del área es el mismo valor que tiene 0.975 del área entre este valor y -∞. Para este ejemplo z = 1.96 y -1.96. La región de rechazo consiste en todos los valores de la estadística de prueba mayores o iguales que 1.96 o menores o iguales que -1.96.

Al realizar los cálculos se tiene:

$$z = \frac{27 - 30}{\frac{\sqrt{20}}{\sqrt{10}}}$$

$$z = -2{,}12$$

Con base en la regla de decisión, se puede rechazar la hipótesis nula porque -2.12 está en la región de rechazo y el valor calculado de la prueba estadística tiene un nivel de significación de 0.05.

Se concluye que μ no es igual que 30.

2.3.2. Prueba de hipótesis para la diferencia entre medias de dos poblaciones

Cuando se tienen dos poblaciones y se toman muestras aleatorias independientes de tamaños n_1 y n_2 , se puede comparar el comportamiento de dichas poblaciones a traves de los promedios. Para la prueba de hipótesis se puede plantear lo siguiente:

Hipótesis:

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

Prueba estadística:

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{x}$ = media de la muestra
σ = desviación estándar de la población
n = tamaño de la muestra

Ejemplo:

Un equipo de investigadores quiere saber si los datos que recolectaron proporcionan evidencia suficiente para indicar una diferencia entre las concentraciones medias de ácido úrico en el suero de personas normales y personas con síndrome de Down.

Los datos fueron recopilados de las lecturas de ácido úrico en el suero de 12 personas con síndrome de Down y 15 personas sanas.

Las medias son $\bar{x_1} = 4.5$ mg / ml y $\bar{x_2} = 3.4$ mg / ml.

Varianza de la población de síndrome de Down $\sigma_1^2 = 1$, y de la población sana $\sigma_2^2 = 1{,}5$

Hipótesis:

H_0: μ_1 - $\mu_2 = 0$

H_a: μ_1 - $\mu_2 \neq 0$

Cálculos:

$$z = \frac{(4{,}5 - 3{,}4) - 0}{\sqrt{\frac{1}{12} + \frac{1{,}5}{15}}}$$

$$z = \frac{1{,}1}{0{,}4282}$$

$$z = 2{,}57$$

Se rechaza H_0 debido a que $2.57 > 1.96$

2.3.3. Prueba de hipótesis para la proporción de una sola población

Cuando se requiere probar una hipótesis acerca de la proporción de una población de valores ubicados dentro de una categoría específica, en vez de probar la media poblacional, se selecciona una muestra aleatoria y se calcula la proporción de la muestra con la siguiente fórmula:

$$p = \frac{n}{N}$$

Una vez obtenida la muestra aleatoria, a continuación se compara el valor del parámetro establecido en la hipótesis, con el fin de decidirse a rechazar la hipótesis nula.

En tanto el número de éxitos (p) como el de fracasos (p - q) sea por lo menos cinco de cada uno, la distribución muestra de una proporción tiene una distribución muestral estandarizada aproximadamente normal. Para efectuar la prueba de hipótesis de la diferencia que existe entre la proporción muestra de p y la proporción poblacional establecida en la hipótesis, se utiliza la prueba Z para la proporción que muestra la ecuación siguiente:

Hipótesis:

H_0: $p = p_0$
H_a: $p \neq p_0$

Prueba estadística:

$$z = \frac{p - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

$$q_0 = 1 - p_0$$

p = proporción observada
p_0 = proporción asumida
n = tamaño de la muestra

Ejemplo:

En una investigación de consumidores de drogas intravenosas en una ciudad, se encontraron 18 de 423 individuos con VIH positivo. Se quiere saber si es posible concluir que el 5 % de consumidores de drogas intravenosas en la población muestreada tienen VIH positivo.

Hipótesis:

H_0: $p = 0{,}05$
H_a: $p \neq 0{,}05$

Cálculos:

$$z = \frac{\frac{18}{423} - 0{,}05}{\sqrt{\frac{(0{,}05)(0{,}95)}{423}}}$$
$$z = -0{,}70$$

No se rechaza H_0 porque -0.70 no está en la región de rechazo.

2.3.4. Prueba de hipótesis para la diferencia entre las proporciones de dos poblaciones

El objetivo de esta prueba es determinar si las dos muestras independientes fueron tomadas de dos poblaciones, estas presentan la misma proporción de elementos con una característica en especial. La prueba se enfoca en la diferencia relativa, es decir, la diferencia dividida entre la desviación estándar de la distribución de muestreo, entre las dos proporciones muestrales. La existencia de diferencias pequeñas únicamente muestran la variación casual por lo que se acepta H_0, por el contrario, grandes diferencias significan que se rechaza. El valor de este estadístico de prueba es comparado con un valor en tablas de la distribución normal, con la intención de decidir si H_0 es aceptada o rechazada. Para efectos de esta investigación, en esta prueba de hipótesis se plantea una prueba unilateral, esto se debe a que en el capítulo 5, en el experimento 2 donde los resultados son comparados con un ejemplo real se hará uso de este tipo de prueba.

Hipótesis:

$$H_0\text{: } p_1 \leq p_2$$
$$H_a\text{: } p_1 > p_2$$

Prueba estadística:

$$z = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}}$$

$$\sigma_{p_1-p_2} = \sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}, \qquad \bar{p} = \frac{x_1 + x_2}{n_1 + n_2}, \qquad \bar{q} = 1 - \bar{p}$$

$p1, p2$ = proporciones de las poblaciones
$\bar{p}$ = estimación ponderada
n_1, n_2 = tamaño de las muestras
x_1, x_2 = elementos en muestras con características de interés
$\sigma_{p_1-p_2}$ = error estándar estimado

Ejemplo:

En un estudio de cuidados nutricionales en asilos, se encuentra que entre 55 pacientes con hipertensión, 24 tenían una dieta con restricción de sodio. De 149 pacientes sin hipertensión, 36 tenían una dieta sin sodio. Se desea saber si la proporción de pacientes con dieta restringida en sodio es mayor con pacientes con hipertensión que entre pacientes sin hipertensión.

Hipótesis:

$$H_0\text{: } p_1 \leq p_2$$
$$H_a\text{: } p_1 > p_2$$

Cálculos:

$$z = \frac{0{,}4364 - 0{,}2416}{\sqrt{\frac{(0{,}2941)(0{,}7059)}{55} + \frac{(0{,}2941)(0{,}7059)}{149}}}$$

$$z = 2{,}71$$

Se rechaza H_0 porque $2{,}71 > 1{,}645$

2.3.5. Prueba de hipótesis para la varianza de una sola población (χ^2)

Esta prueba permite comparar la varianza de una población que tiene una distribución normal.

Hipótesis:

H_0: $\sigma^2 = \sigma_0^2$
H_a: $\sigma^2 \neq \sigma_0^2$

Prueba estadística:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

σ^2 = varianza de la población
s = varianza de la muestra
n = tamaño de las muestras

Ejemplo:

Se examina la respuesta a la inhalación de un alergeno en primates alérgicos. Se estudian 12 monos que cumplían con los criterios para el estudio.

Se reportó un error estándar de 0.4 para uno de ellos.

Se pretende saber si se puede concluir a partir de estos datos que la varianza de la población es diferente de 0.4.

Hipótesis:

H_0: $\sigma^2 = 0{,}4$
H_a: $\sigma^2 \neq 0{,}4$

Cálculos:

$$s^2 = 12(2{,}4)^2 = 1{,}92$$

$$\chi^2 = \frac{(11)(1{,}92)}{4} = 5{,}28$$

No se rechaza H_0 porque 5,28 se encuentra en la región de aceptación $3{,}816 < 5{,}28 < 21{,}920$.

2.3.6. Prueba de hipótesis para la razón de las varianzas de dos poblaciones (χ^2)

Esta es una prueba que al igual que en las anteriores, la hipótesis nula H_0 se asocia con la igualdad entre los estadístico de prueba poblacionales y la hipótesis alternativa H_a.

Hipótesis:

H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 \neq \sigma_2^2$

Prueba estadística:

$$R.V. = \frac{s_1^2}{s_2^2}$$

s = varianza de la muestra

Ejemplo:

Se investiga cierta enfermedad relacionada con el sobrepeso. La desviación estándar de los pesos de una muestra de 12 pacientes fue de 21,4 kg. Los pesos de una muestra de 5 individuos de control produjo una desviación estándar de 12,4 kg.

Se desea saber si es posible concluir que los pesos de la población representada por los pacientes de la muestra ofrecen una variabilidad diferente que los pesos de la población representada por la muestra de control.

Hipótesis:

H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 \neq \sigma_2^2$

Cálculos:

$$R.V. = \frac{(21{,}4)^2}{(12{,}4)^2}$$

$$R.V. = 2{,}98$$

No es posible rechazar H_0 debido a que se encuentra en la región de no rechazo.

2.3.7. Prueba de hipótesis chi-cuadrada (prueba de independencia)

Es una prueba estadística no paramétrica para diferencias entre dos o más muestras donde frecuencias esperadas son comparadas en relación con frecuencias obtenidas. Se utiliza para hacer comparaciones entre frecuencias y no entre valores medios.

Prueba no paramétrica: Procedimiento estadístico que no adopta ningún supuesto acerca de cómo se distribuye la característica bajo estudio en la población, y que sólo requiere datos nominales u ordinales. Estas medidas son importantes porque la mayoría de la información en la investigación social y administrativa es de carácter nominal u ordinal, y porque no siempre se esta seguro que la característica que se desea estudiar se distribuye normalmente en la población.

La prueba de significación χ^2 se refiere esencialmente a la distinción entre frecuencias esperadas y frecuencias obtenidas. Las frecuencias esperadas f_e se refieren a los términos de la hipótesis nula, según la cual la frecuencia relativa (o proporción) se supone es la misma entre los dos grupos. Las frecuencias obtenidas f_o se refieren a los resultados obtenidos en el estudio y que, por consiguiente, pueden variar o no de un grupo a otro. Sólo si la diferencia entre las frecuencias observadas y obtenidas es suficientemente grande, se rechaza la hipótesis nula, y se concluye que existe una diferencia real en la población.

Como resultado, la hipótesis nula para χ^2 señala que las poblaciones o grupos no difieren con respecto a la frecuencia de ocurrencia de una característica dada. Mientras que la hipótesis de investigación señala que las diferencias entre las muestras reflejan diferencias reales en la población con respecto a la frecuencia relativa de una característica dada.

Una vez que tenemos las frecuencias esperadas y obtenidas, el valor de χ^2 se obtiene de la siguiente manera:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Las frecuencias esperadas se obtienen de la siguiente manera:

$$f_{e1,1} = \frac{(TotalColumna)(TotalRenglon)}{TotalTotal}$$

En la prueba de independencia, se toma la decisión acerca de si una variable es independiente de la otra. La hipótesis nula se establece suponiendo que son independientes. Los datos se acomodan en una tabla llamada tabla de contingencia, en la cual existe N clases o categorías de renglón y M clases o categorías de columnas. Al final de cada una de las filas o columnas se escribe los totales marginales de fila R_1 o columna C_x, que es la frecuencia observada.

Ejemplo:

Para estudiar la dependencia entre la práctica de algún deporte y la depresión, se seleccionó una muestra aleatoria simple de 100 jóvenes, con los siguientes resultados:

	Sin depresión	Con depresión
Practica deporte	38	9
No practica deporte	31	22

Figura 2.1:

Se requiere determinar si existe independencia entre realizar deporte y el estado de ánimo con un nivel de confianza de 95 %.

Se obtiene la tabla de frecuencias esperadas:

	Sin depresión	Con depresión	Total
Practica deporte	32.43	14.57	47
No practica deporte	36.57	16.43	53
Total	69	31	100

Figura 2.2:

Cálculos:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(38 - 32{,}43)^2}{32{,}43} + \frac{(9 - 14{,}57)^2}{14{,}57} + \frac{(31 - 36{,}57)^2}{36{,}57} + \frac{(22 - 16{,}43)^2}{16{,}43}$$

$$\chi^2 = 5{,}82$$

Se sabe que:

$$\chi^2_{0{,}95}(1) = 3{,}84$$

Como el valor estadístico es mayor al valor crítico, se rechaza la hipótesis de independencia y se asume que existe relación entre estas dos variables, depresión y hábitos del deporte.

Capítulo 3

Almacén de datos y tecnología OLAP

En el presente capítulo se abordan los fundamentos de los almacenes de datos, además de lo que es una base de datos operacional y de dónde se obtiene la información que integrará un almacén. Del mismo modo se describen sus características y la forma en que se explotan haciendo uso de herramientas OLAP.

3.1. Almacén de datos

Un almacén de datos o Data Warehouse, es un almacén de información temática orientado a cubrir las necesidades de aplicaciones de los Sistemas de Soporte de Decisiones (DSS) y de la Información de Ejecutivos (EIS), que permite acceder a la información corporativa para la gestión, control y apoyo a la toma de decisiones [19].

Se puede decir que un almacén de datos es una plataforma donde se encuentran los datos aislados, pudiendo ser una computadora, un mainframe o muchas máquinas en conjunto, las cuales conforman una plataforma distribuida que permite la toma de decisiones de una manera ágil. Para esto, se necesita que los datos existentes sean duplicados en algún lugar diferente a los originales generando así una redundancia sobre los mismos [32].

En un almacén de datos, se combina el hardware con herramientas de software y gran cantidad de datos. De acuerdo con Bill Inmon [14] un almacén de datos es

una colección de datos orientados a temas, integrados, no volátiles y variantes en el tiempo, organizados para soportar necesidades organizacionales.

El almacén generalmente es una base de datos de sólo lectura, optimizada para el análisis de los datos y para el procesamiento de consultas. Los datos pueden ser extraídos de diversas fuentes para luego ser transformados e integrados antes de ser transferidos al almacén. Los usuarios acceden al almacén a través de herramientas finales o algún software de aplicación para extraer los datos de una manera entendible y utilizable.

La importancia de un almacén de datos radica en:

- Son un soporte para la toma de decisiones tácticas y estratégicas proporcionando un sentido automatizado para la identificación de información valiosa, desde volúmenes de datos generados por procesos ya conocidos o por elementos de software.
- Muestran a los usuarios como manejar sus prioridades hacia decisiones o acciones.
- Permiten la medición de acciones y resultados de una mejor manera.
- Generan modelos descriptivos, es decir, permiten explorar de forma automática, visualizar y comprender los datos e identificar patrones, relaciones y dependencias que influyen en los resultados.

Un almacén de datos se refiere a una base de datos históricos de gran volumen que proporciona la opción de integrar los datos de sistemas distribuidos que contienen información homogénea. Se guardan en él, todos los datos necesarios para un cierto estudio y representa la actividad de un desarrollo durante períodos de tiempo importantes, por lo que puede ser utilizado en el apoyo a la toma de decisiones [30].

La información que contiene un almacén de datos es construida a partir de bases de datos que registran las transacciones de las organizaciones. Una base de datos es un conjunto de información almacenada sistemáticamente por un programa manejador de datos para su uso posterior en forma rápida y racional.

Las aplicaciones más comunes de las bases de datos son en la gestión y almacenamiento de la información experimental en instituciones y entornos científicos,

en documentación y en cualquier campo que trabaje con datos que requieran ser ordenados y estructurados.

Las bases de datos se pueden clasificar en base a su funcionalidad y pueden ser analíticas u operativas. Las bases de datos analíticas son de sólo lectura, son utilizadas principalmente para almacenar datos históricos que en un futuro se pueden utilizar para estudiar el comportamiento de un conjunto de datos a medida que pasa el tiempo, realizar proyecciones y tomar decisiones. Las bases de datos funcionan de una manera dinámica, están orientadas a almacenar información que es modificada con el tiempo, permiten operaciones como actualizar o agregar datos, además de las operaciones fundamentales para la consulta.

En un almacén de datos se utilizan unas cuantas tablas, los datos son producto de los datos operativos, que están integrados, agregados y resumidos para propósitos de soporte de decisiones.

3.1.1. Características de un almacén de datos

Recordando la definición de Bill Inmon [14], la cual dice que un almacén de datos es una colección de datos orientados a temas, integrados, no volátiles y variantes en el tiempo, organizados para soportar necesidades organizacionales. Se extraen de esta definición las siguientes características:

No volátil: En un almacén de datos sólo se realizan dos operaciones, la carga de los datos procedentes de entornos operacionales, que son la carga inicial y las cargas periódicas; y la consulta de los mismos. Una vez introducidos los datos nunca se eliminan y los nuevos siempre se agregan, debido a que representan la historia completa de la organización. Por esta razón un almacén de datos siempre está en proceso de crecimiento. Los datos no son cambiados, una vez que se guardan adecuadamente, no se permiten cambios por lo que los datos son estáticos [31].

Orientados a temas: Los datos son almacenados por materias o temas, se organizan desde el punto de vista del usuario final, de tal forma que puedan responder a preguntas que vengan de diversas áreas, con el objetivo de lograr una mayor

eficiencia en el acceso a los datos. Son organizados y reunidos por temáticas, como ventas, finanzas, etc. Y por cada tema contiene sujetos de interés específicos como productos, clientes, departamentos, regiones, etc. [31].

Integrado: Se refiere a que la información está centralizada y consolidada en un solo lugar; integrando los datos que provienen de toda la organización y traídos de múltiples fuentes con diversos formatos. Esta integración ayuda y mejora la toma de decisiones, y a través de ella se logran entender mejor las operaciones realizadas y reconocer oportunidades. Por otro lado, proporciona una visión de todos los elementos de datos, aunado a una definición y representaciones comunes de todas las unidades de negocios. Los datos en el almacén deben ser ajustados a formatos y estructuras uniformes para evitar conflictos [31].

Variable con el tiempo: Los datos guardados en un almacén representan el flujo de los datos a medida que pasa el tiempo con un punto de vista histórico, además puede contener datos proyectados generados a través de modelos estadísticos o algún otro método. Del mismo modo, es variable en el sentido de que una vez que los datos se encuentran en el almacén, todas las agregaciones que son dependientes del tiempo se vuelven a calcular. Por lo que es importante adicionar una dimensión de tiempo para lograr que el análisis de los datos sea más sencillo así como las comparaciones [31].

Se debe resaltar el carácter histórico, debido a que el tiempo debe estar presente en todos los registros contenidos en un almacén de datos. Por lo que, un almacén se puede ver como una serie de fotos instantáneas en el tiempo tomadas de manera periódica.

William y Chuck Kelley, en 1994 propusieron una lista de 12 reglas que definen un almacén de datos y las cuales son necesarias para que éste sea considerado realmente como tal [15]:

1. El almacén de datos y los ambientes operativos están separados.
2. Los datos guardados en el almacén de datos se encuentran integrados.
3. El almacén de datos contiene datos históricos que abarcan un horizonte amplio en el tiempo.

4. Los datos en el almacén son capturados de forma instantánea en un punto dado del tiempo.

5. Los datos en el almacén de datos están orientados a sujetos.

6. Los datos principalmente son de sólo lectura, con actualizaciones por bloques periódicas a partir de datos operativos. No se permiten actualizaciones en línea.

7. El ciclo de vida del desarrollo de un almacén es diferente al desarrollo de sistemas clásicos. Los datos motivan el desarrollo del almacén de datos; los procesos motivan el método clásico.

8. El almacén de datos contiene datos con distintos niveles de detalle: datos detallados actuales, datos detallados antiguos, datos ligeramente resumidos y datos altamente resumidos.

9. El ambiente del almacén de datos se caracteriza por transacciones de sólo lectura de conjuntos de datos muy grandes. El ambiente operativo se caracteriza por un gran número de transacciones de actualización de unas cuantas entidades de datos al mismo tiempo.

10. El ambiente del almacén de datos dispone de un sistema que rastrea fuentes, transformaciones y almacenamiento de datos.

11. Los metadatos del almacén son un componente crítico de este ambiente. Estos metadatos identifican y definen a los elementos de datos; brindan la fuente, transformación, integración, almacenamiento, uso, relaciones e historial de cada elemento de datos.

12. El almacén de datos contiene un mecanísmo de retroalimentación para el uso de los recursos que exige el uso óptimo de los datos por parte de los usuarios.

De igual forma, el uso correcto de un almacén de datos, un buen diseño e implantación, tienen muchos beneficios tales como:

- Respuestas rápidas a preguntas adecuadas para la toma de decisiones.
- Acceso a datos con el propósito de toma de decisiones.

- Se pueden analizar los datos de manera rápida.
- Se identifican tendencias.
- Errores en el ingreso de datos eliminados.
- El tiempo de respuesta se vuelve más ágil.

3.2. Herramientas OLAP

On Line Analytical Processing (OLAP) es un complejo conjunto de técnicas para realizar análisis de datos exploratorios, estas son generadas por medio de agregaciones y calculadas sobre la base de combinaciones de dimensiones múltiples, lo cual se asemeja a un cubo multidimensional, cuya estructura matemática está representada por una especie de celosía [27].

La explotación de un almacén de datos consiste en hacer consultas, como análisis, manipulación y visualización de la información guardada en el almacén. Una forma de explotar los datos contenidos en un almacén es mediante el uso de aplicaciones OLAP. Este término fue usado en 1993 por EF Codd & Associates, por medio de estas aplicaciones el usuario puede realizar un análisis multidimensional interactivo de los datos.

Las herramientas de DSS (Sistema de Soporte a la Toma de Decisiones) que emplean las técnicas de análisis de datos multidimensionales son conocidas como herramientas de procesamiento analítico en línea. Las herramientas OLAP son herramientas de minería de datos que revelan un problema real, para lograrlo crean un ambiente avanzado de análisis de datos que apoyan a la toma de decisiones, un modelo de negocios y las actividades de investigación de operaciones, esta es la forma más intuitiva de ver la información [32] [31].

Un objetivo de las herramientas OLAP es agilizar las consultas de grandes cantidades de datos. Para lo cual, utilizan estructuras multidimensionales (Cubos OLAP) que contienen datos resumidos de grandes bases de datos o sistemas transaccionales (OLTP, Procesamiento de Transacciones En Línea). Se usan en informes de negocios, informes de dirección, minería de datos, investigación científica y áreas similares. Otro objetivo es que el usuario pueda navegar por la información de manera que vaya

descubriendo los cruces de su interés para el análisis, y de esta forma se enfoque en un punto crítico.

Otra de las características importantes de las aplicaciones OLAP, es la facilidad de uso, la flexibilidad, y la velocidad con la que se obtiene información importante y relevante, el usuario debe realizar la pregunta correcta antes de realizar una consulta.

Las herramientas OLAP ofrecen una de las más importantes funciones de análisis para la generación de información y soporte de decisión, esta es la posibilidad del análisis multidimensional. Del mismo modo, son la principal herramienta de los sistemas de soporte de decisión (DSS). Los sistemas OLAP presentan características como [31]:

- Hacen uso de técnicas multidimensionales para el análisis de datos. Esta es la característica que distingue a las herramientas OLAP. Se refiere al procesamiento de datos tal forma que estos sean vistos como parte de una estructura multidimensional, en consecuencia de que los tomadores de decisiones generalmente visualizan los datos desde una perspectiva del negocio que se relaciona con otros datos.

- Brindan un soporte avanzado para las bases de datos. Poseen funciones de acceso a datos, como acceso a diferentes manejadores de bases de datos, funciones avanzadas de navegación, tiempo de respuesta rápido a consultas, soportan datos muy grandes y proyectan sus propios diccionarios de datos.

- Proporcionan interfaces para los usuarios finales sencillas de utilizar.

- Soportan la arquitectura cliente/servidor. Presentan un marco de referencia en donde pueden disearse, desarrollarse y ejecutarse en sistemas nuevos.

- Funcionan sobre un sistema de información, así sea transaccional o almacén de datos.

- Permiten realizar agregaciones y combinaciones de los datos de formas más complejas, con objetivos de análisis más estratégicos.

- Se basan en sistemas o interfaces multidimensionales.

- Utilizan operadores específicos, además de los ya conocidos como: drill, roll, pivot, slice, dice, etc.
- El resultado se presenta en forma de matriz o híbrida.
- Brindan la facilidad de manejar y transformar los datos.
- Generan otros datos a través de funciones de agregación y combinaciones.
- Ayudan en el análisis de los datos, produciendo diferentes vistas.
- Ayudan a comprender las interrelaciones que existen entre todas las herramientas.

En la actualidad, los almacenes de datos y las herramientas OLAP son las formas más efectivas y tecnológicamente más avanzadas para integrar, transformar y combinar los datos para hacerle más fácil al usuario o a otros sistemas el análisis de la información. La tecnología OLAP la mayoría de las veces es asociada a los almacenes de datos, aunque se pueden tener almacenes de datos sin OLAP y viceversa.

Con OLAP se puede observar el conjunto de datos de muchas y diversas formas sin gran esfuerzo. Los cubos OLAP modelan los datos en dimensiones. Una dimensión se refiere a una clasificación de alguna actividad en alguna organización por la cual se puede medir su éxito.

Los conceptos se agrupan en dimensiones, las dimensiones son diferentes caras de la misma realidad que se encuentran estructuradas jerárquicamente para poder resumir la información y con la posibilidad también de investigar más a detalle. La información organizada de forma multidimensional muchas veces se puede representar con la forma de un cubo.

3.2.1. Operaciones básicas OLAP

En OLAP se utilizan principalmente cuatro tipos de operaciones. Cuando se considera la granularidad de los datos, se puede hacer uso de las operaciones drill down y roll up. Para navegar a lo largo y ancho de todas las dimensiones, se utilizan las operaciones slice and dice.

Drill down y Roll up: Drill down y roll up se refieren a las operaciones para el movimiento de las vistas hacia abajo y hacia arriba a través de los niveles jerárquicos de las dimensiones. Navegando con drill-down, se puede navegar a un nivel más detallado. Con roll-up se puede tener una visualización más amplia y panorámica de los datos.

Slice y Dice: Slice y dice son las operaciones para navegar a través de los datos con el objetivo de visualizar todo el cubo. Haciendo cortes en el cubo es como se pueden tener diferentes perspectivas.

En [2] se mencionan además las operaciones siguientes:

Pivot: Brinda la posibilidad de modificar la posición de las dimensiones que caracterizan a los datos presentados, el punto de vista con el que se están presentando los datos cambia, pero no cambian las dimensiones.

Drill-across: No muestra detalle ni une la información, sino que cambia el modelo actual de consulta, la única condición es que los modelos tengan en común una dimensión.

Drill-through: Es similar a drill-down ya que permite consultar la información del nivel inferior a la dimensión actual a detalle, pero con la diferencia de que drill-through tiene la posibilidad de navegar fuera del modelo multidimensional, crea un enlace entre el modelo multidimensional y el sistema fuente de datos, que es a partir del cual se creó en un principio el modelo multidimensional, para poder consultar los datos del nivel a detallar directamente del sistema fuente.

3.2.2. Cubos OLAP

En todos los sistemas OLAP se hace referencia al concepto de cubo OLAP, también se le puede llamar cubo multidimensional o hiprecubo, este se compone de hechos numéricos llamados medidas y se clasifican por dimensiones. La idea de tener presentes los cubos, es de realizar diferentes consultas buscando información en el almacén desde perspectivas, visiones o puntos de vista diferentes.

En [34] se menciona que un cubo OLAP es una base de datos que presenta múltiples dimensiones para el almacenamiento físico de los datos, el cual se realizará dentro

de una red de múltiples dimensiones según se vaya requiriendo; lo cual, desde su propia perspectiva es como si los datos se fueran guardando dentro de una figura que está compuesta por una cantidad indefinida de lados (dimensiones) según sean necesitados, esto contribuye a mejorar significativamente el análisis y las consultas de cada uno de los datos dentro de la información contenida, que permite manipular gran cantidad de ésta de una forma rápida y precisa, porque se encuentra agrupada dentro del campo correspondiente que mejor la define. Esto permite además proporcionar sistemas más confiables y seguros para la toma de decisiones, lo mismo que para lograr informes más precisos, ya que se puede llegar a entender como un incremento de las dimensiones de una tabla u hoja de cálculo, porque un cubo OLAP amplía las posibilidades de que estas ofrezcan al usuario una mejor visión.

El cubo de metadatos se crea la mayoría de las veces a partir de un esquema en estrella. Este esquema posee una sola tabla de hechos y se encuentra rodeada de un conjunto de tablas de dimensiones. La tabla de hechos es la que contiene la mayor cantidad de información y puede llegar a almacenar millones de registros, a diferencia de cada una de las tablas de dimensión, las cuales contienen muy pocos registros [32].

También puede ser creado a partir de un esquema de copo de nieve aunque esta puede ser una estructura más compleja que el esquema de estrella. Un esquema de copo de nieve se da cuando alguna de las tablas de dimensiones se implementa con más de una tabla de hechos. El objetivo es normalizar las tablas y así reducir el espacio de almacenamiento al quitar redundancia; la desventaja que presenta es la de un rendimiento poco eficiente, debido a que se crean más tablas de dimensiones y más asociaciones entre las tablas.

Un cubo OLAP es una representación multidimensional de los datos detallados y de los datos resumidos, los datos detallados son aquellos datos de filas específicas y los datos resumidos son los datos agregados a dichas filas [32].

Por su parte, Carlos Ordoñez y Chen Zhibo [26] consideran que el punto de vista de la definición técnica del cubo OLAP, resulta ser una representación abstracta de la proyección de una relación de un Sistema Administrador de Bases de Datos Relacionales (RDBMS). Esto se debe a una relación de orden N, donde se considera la posibilidad de una proyección que dispone de los atributos X, Y, Z, como clave

de la relación y de W, como atributo residual. Refieren que los atributos X, Y, Z corresponden con los ejes del cubo, mientras que el valor de W, devuelto por cada tripleta (X, Y, Z) corresponde con el dato o elemento que se rellena en cada celda del cubo. De acuerdo con ellos surge la función siguiente:

$$W : (X, Y, Z) \rightarrow W$$

Los cubos OLAP pueden ser considerados como una compilación de las dos dimensiones de una hoja de cálculo. Los parámetros en función de los cuales se analizan los datos se conocen como dimensiones. Para acceder a los datos sólo es necesario indexarlos a partir de los valores de las dimensiones o ejes.

En un sistema OLAP se pueden encontrar más de tres dimensiones, por lo que los cubos OALP también reciben el nombre de hipercubos. Las herramientas comerciales OLAP tienen distintos métodos de creación y vinculación de estos cubos o hipercubos [10].

3.3. Diseño de un almacén de datos

Este diseño se refiere al proceso de elegir y diseñar la estructura interna, misma que soporta la manipulación y recuperación de los datos, es decir, la forma en que los datos se harán disponibles al usuario.

Para construir un almacén se necesitan herramientas para ayudar a la migración y transformación de los datos hacia el almacén. Una vez creado, se necesitan medios para poder manejar grandes volúmenes de información. Se diseña su arquitectura dependiendo de la estructura interna de los datos del almacén y especialmente del tipo de consultas a realizar.

Los almacenes de datos se han convertido en una tendencia importante por lo que es fundamental que el diseño de los mismos se realice de tal forma que sean eficientes y sencillos de utilizar.

3.3.1. Organización lógica

La organización lógica de un almacén de datos se puede realizar mediante esquemas multidimensionales y relacionales, durante esta investigación se utilizó un esquema multidimensional por lo que es el que se describe a continuación.

Esquema multidimensional

Generalmente para el diseño de un almacén de datos se emplea la representación multidimensional. El esquema multidimensional es una técnica de diseño lógico que tiene el objetivo de representar los datos en un estándar, que permita una recuperación adecuada de estos. Así los datos son almacenados como hechos y dimensiones.

Este esquema brinda una estructura que permite tener un fácil acceso a los datos para explorar y analizar sus relaciones, las cuales pueden ser visualizadas mediante un cubo de datos multidimensional, en donde las variables asociadas se presentan a lo largo de varios ejes o dimensiones y la intersección de las mismas representa la medida, indicador o el hecho que se evalúa.

En un esquema multidimensional, el modelado de datos es un conjunto de medidas definidas por dimensiones adecuado para resumir y organizar los datos. Esta enfocado a trabajar sobre datos numéricos de una forma más simple para lograr visualizar y entender la información de una manera más sencilla que el modelo relacional. El esquema multidimensional está formado por los siguientes elementos:

Cubos: Los datos de un almacén se modelan en cubos de datos o hipercubos. Un cubo organiza los datos en una o más dimensiones que determinan una medida de interés. El esquema de un cubo está determinado por un conjunto de ejes y una o muchas medidas, mientras que una instancia de cubo se define como un conjunto de n número de tuplas o celdas. Cada celda, es una instancia de los niveles y las medidas definidas en el esquema del cubo.

Dimensiones: Una dimensión especifica la forma en que se observan los datos para su análisis. Cada dimensión está formada por un conjunto de atributos, cada uno con su propio dominio. Son un concepto esencial de bases de datos multidimensionales, si los datos se encuentran almacenados dentro de un cubo

se tienen la ventaja de manejar cualquier número de dimensiones, un cubo generalmente soporta un vista de dos o tres dimensiones simultáneamente.

Jerarquías: Las dimensiones se relacionan en jerarquías o niveles, las jerarquías de agregación colocan los atributos a varios niveles y permiten observar los datos a distintas granularidades. Una dimensión puede tener varias jerarquías asociadas, cada una especifica una relación de orden distinta entre sus atributos, El esquema de una dimensión toma la forma de un orden parcial entre los niveles, permitiendo la definición de las jerarquías simples o múltiples. Una instancia de dimensión es una familia de funciones que determinan los elementos de un nivel que se relacionan con los elementos de un nivel superior.

Hechos: Los hechos representan el patrón de interés o el evento que necesita ser analizado, cada hecho está asociado a un miembro de cada dimensión y las dimensiones determinan el contexto de los hechos. Los hechos son implícitamente definidos por la combinación de valores de las dimensiones. Son colecciones de datos relacionados compuestas por medidas y un contexto. Un almacén usualmente contiene tres tipos de hechos:

- Snapshots, son los que modelan entidades en un punto dado en el tiempo.
- Eventos, modelan eventos en el mundo real.
- Snapshots acumulativos, modelan actividades en un punto dado en el tiempo.

Medidas: Una medida es el objeto de análisis, es un valor en un espacio multidimensional definido por dimensiones ortogonales, representa una actividad en específico. Son atributos numéricos asociados a los hechos, es lo que realmente se mide, un dato numérico que representa l agregación de un conjunto de datos.

Dependiendo de sus propiedades se pueden tener medidas:

- Aditivas, pueden ser combinadas a lo largo de cualquier dimensión.
- Semiaditivas, no pueden ser combinadas a lo largo de una o más dimensiones.

- No aditivas, las que no pueden ser combinadas a lo largo de ninguna dimensión.

Capítulo 4

Sistema de visualización de cubos

4.1. Visualización OLAP

El problema de la visualización de datos multidimensionales, el cual es producto de las tareas científicas y estadísticas, es cada vez más difícil de resolver, por esta razón llama la atención de grandes comunidades multidisciplinarias de investigadores y profesionales. El problema fundamental consiste en capturar las dimensiones de los datos para poder lograr una visualización multidimensional del conjunto de datos. Los analistas que interactúan con este tipo de datos de muchas dimensiones, frecuentemente presentan una experiencia de desorientación y sobrecarga de conocimiento. El análisis de datos de muchas dimensiones se puede encontrar en diversos casos reales, como son: Almacenamiento de bases de datos masivas con información biológica de genes y conjuntos de datos de proteínas, Sistemas de vigilancia en tiempo real que acumulan conjuntos de datos producidos por múltiples fuentes y de diferentes velocidades, sistemas avanzados de negocios de inteligencia recolectan datos para la toma de decisiones y sus efectos.

Las herramientas tradicionales para los DBMS son inadecuados para cubrir los requisitos que plantea la exploración interactiva para los conjuntos de muchas dimensiones debido principalmente a 2 razones:

- Los DBMS utilizan el paradigma OLTP que es optimizado para el procesamiento de transacciones y descuida las dimensiones de los datos.

- Los operadores DBMS son muy pobres y no ofrecen nada más allá que las capacidad de las sentencias SQL convencionales, lo que convierte en ineficientes a estas herramientas respecto al objetivo de la visualización y sobre todo de la interacción con datos de un gran número de dimensiones.

A pesar de la relevancia que tiene el problema de la visualización de datos multidimensionales, la literatura para este campo es un tanto escasa debido a que durante muchos años, este problema ha sido relevante sólo para comunidades de investigación en ciencias de la vida y la interacción de esta con la comunidad de investigación en ciencias de la computación ha sido insuficiente. Siguiendo con el gran crecimiento de las disciplinas científicas como la Bio-Informática, este problema se ha convertido en un campo fundamental para las ciencias de la computación así como en la investigación industrial. Al mismo tiempo, aparecen propuestas referentes al problema visualización de datos multidimensionales con la necesidad de fomentar aplicaciones en campos como la visualización de resultados en minería de datos generados por técnicas difíciles como el clustering y el descubrimiento de reglas de asociación.

Los problemas mencionados son con el objetivo de facilitar la comprensión de la relevancia y lo llamativo del problema de la visualización de datos multidimensionales en la actualidad y en el futuro. Se han presentado posibles soluciones para resolver este problema, representadas por técnicas conocidas descritas en [6] [3] [10], enfocadas en obtener una representación eficiente de datos multidimensionales, llamados cubos de datos, así se llega al campo de investigación conocido bajo el término de Visualización OLAP.

De manera formal, dada una Relación R, un cubo de datos L definido en la parte superior de R es una tupla T = (A,D,J,M), tal que:

- A es el dominio de los datos de T contenidos en celdas de datos que almacenan agregaciones de SQL, como las obtenidas con SUM, COUNT, AVG, etc. Obtenidas con registros de R.

- D es el conjunto de atributos de R respecto a la tupla definida, es decir, las dimensiones de T.

- J es el conjunto de jerarquías relacionadas con las dimensiones de T.

- M es el conjunto de atributos de interés de R para el análisis OLAP, esto es, las medidas de T.

De esta manera los cubos de datos OLAP pueden ser usados para visualizar los datos multidimensionales y además permitir la exploración interactiva de los datos usando un conjunto de operadores. [11], drill-down, roll-up, pivot. Además de las bondades de visualización, OLAP brinda soluciones eficientes para el problema de la representación de datos multidimensionales a través de un conjunto de alternativas clasificadas de acuerdo a la manera en que se encuentran almacenados en memoria, ROLAP, MOLAP, HOLAP. Se puede notar que la eficiencia en la representación de los datos tiene un gran impacto en la efectividad al momento de visualizar los datos y en las actividades realizadas para su exploración.

Las herramientas tradicionales OLAP, son diseñadas para presentar reportes y utilizar rutinas para el análisis, el uso de la visualización es únicamente para mostrar una representación de los datos. En OLAP, la visualización juega un papel importante para el análisis interactivo. Para un análisis más exhaustivo, se incluyen una variedad de tareas tales como: Examinar los datos desde múltiples perspectivas, extracción de información útil, verificación de hipótesis, tendencias, patrones relevantes, obtención de ideas y descubrir nuevos conocimientos a partir de grandes y/o complejos volúmenes de datos multidimensionales. Además de las operaciones convencionales para el procesamiento analítico como drill-down, roll-up, slice-and-dice, pivot y ranking, la visualización OLAP permite técnicas de manipulación de datos como el zoom y vista panorámica, filtrado, cepillado, etc.

Las primeras propuestas acerca del uso de la visualización para explorar grandes conjuntos de datos no se adaptaban a las aplicaciones OLAP, pero se trataba el problema genérico de consultas visuales para estos datos en las bases. Las primeras experiencias encontradas para la visualización de datos se encuentran en aplicaciones para escenarios de la vida real, tal es el caso propuesto en [9] donde se propone una interfaz visual inteligente para las bases de datos multidimensionales, o también en fundamentos teóricos, tal como los indicados en [1] donde se habla acerca del problema de eficiencia en la visualización e interacción con datos de grandes dimensiones. Keim y Kriegel, en [18] se propone VisDB, un sistema de visualización basado en un innovador paradigma de consultas. En VisDB, los usuarios indicaban

una primera consulta, posteriormente a través de una guía visual la consulta se podía ajustar dinámicamente especificando rangos sobre simples o múltiples atributos. Los resultados obtenidos eran mapeados sobre pixeles en una zona rectangular diseñada para la visualización, iluminados de acuerdo al grado de relevancia para el conjunto específico que se seleccionó, y en una posición de acuerdo a una agrupación u orden.

Una interfaz común para el análisis de datos OLAP es una tabla dinámica o tabla de cruce, la cual se puede ver como una hoja de cálculo multidimensional producida mediante la especificación de una o más medidas de interés y la selección de dimensiones que servían como ejes verticales. El poder de esta técnica de presentación proviene de la capacidad de resumir detalladamente los datos a lo largo de muchas dimensiones y la organización de las funciones de agregación calculadas para diferentes niveles en una sola vista, manteniendo los resultados parciales. Estas tablas son ineficientes para resolver tareas analíticas no triviales tales como el reconocimiento de patrones, descubrir tendencias, etc. A pesar de esta debilidad, las tablas dinámicas siguen teniendo un gran poder en las técnicas de visualización, ahorran tiempo y reducen errores en el razonamiento analítico a través del uso de capacidades de la visión humana [8].

Las interfaces OLAP actuales mejoran la vista de la tabla dinámica proporcionando un conjunto de técnicas de visualización de negocios, como las gráficas de barras, gráficas de pastel, series en el tiempo, o también algo más sofisticado como diagramas de dispersión, mapas, mapas con estructura de árbol, cartogramas, matrices, etc. Algunas herramientas van más allá de la presentación visual de datos con el fin de proponer mejores enfoques motivados por la búsqueda de información y su visualización para la investigación. Ejemplos de ello son los sistemas visuales avanzados Advizor [8] y Tableau [12]. Advizor utiliza una técnica de organización de datos en tres perspectivas. Una perspectiva es un conjunto de componentes visuales relacionados que se muestran en la misma pantalla. Cada perspectiva se enfoca en un tipo particular de tarea analítica. Una vista para una medida utilizando un diseño en 3D, otra vista para múltiples medidas organizadas en una gráfica de dispersión y una última con las medidas utilizadas presentadas utilizando técnicas de visualización multidimensional. Tableau es una herramienta visual para el análisis multidimensional desarrollado en la universidad de Standford, utiliza la idea básica

de la tabla dinámica que ubica las funciones de agregación en una cuadrícula definida por categorías y dimensiones asignadas a una matriz con filas y columnas. Utiliza marcas gráficas en vez de números en las celdas. Los tipos de gráficos utilizados son rectángulo, círculo, texto, barra, línea, polígono e imágenes.

Continuando con los problemas básicos, en [33] se propone distinguir tres partes del ciclo de vida para las técnicas de visualización, estas son, la maduración, evolución y aparición. Dentro de esta clasificación, la visualización OLAP se clasifica entre las técnicas que aparecen para una interacción avanzada y consultas visuales. En la visualización OLAP la inefectiva presentación de los datos no es la única deficiencia de las herramientas. Otros problemas son la forma engorrosa de utilizarlos y la pobre funcionalidad para explorar. La visualización OLAP se enfoca en el desarrollo de formas nuevas de cómo interactuar con funciones de agregación aplicadas en datos multidimensionales. Una nueva calidad en el análisis visual se obtiene mediante la unión entre la tecnología OLAP y la visualización de información.

La tarea de escoger una técnica de visualización apropiada para resolver un problema en particular no es algo fácil, el problema de ayudar a un analista a identificar una técnica de visualización adecuada para una tarea específica aún es un problema. Generalmente un usuario tiene que encontrar una solución adecuada a través de la experimentación con diferentes opciones para el diseño. Para poder emplear diversas técnicas de visualización y poder cambiar entre una y otra de manera dinámica, se necesita crear una capa donde se especifiquen las relaciones entre los datos y su representación visual.

Con esta idea, el modelo de cinta, propuesto en [9], propone representar una visualización de datos multidimensional utilizando la analogía con cintas y pistas, incluso con la posibilidad de definir estructuras jerárquicas dentro de la cinta. En [22] [21], propone una solución que consiste en capas, llamada Modelo de Presentación del Cubo (CPM), el cual hace la distinción entre 2 capas, una capa lógica la cual se encarga del modelado de datos y recuperación, y una capa de presentación, la cual brinda un modelo genérico para representar los datos, generalmente una pantalla en 2D. Los elementos para la capa de presentación son puntos, ejes, cubos, capas, cintas, uniones y funciones.

Un enfoque común para la visualización en aplicaciones OLAP se basa en un

conjunto de plantillas, ayudas y selecciones para formatos visuales. Hanrahan [12] sostiene que un conjunto abierto de requisitos no puede ser tratado mediante un conjunto limitado de técnicas, y emplea un enfoque diferente para su herramienta de análisis visual, Tableau. Esta nueva herramienta esta caracterizada por el uso de VizQL, un lenguaje visual de consulta. VizQL permite a los usuarios crear una presentación visual propia mediante la combinación de diferentes componentes visuales. Los diseñadores de Tableau restringen el conjunto de elementos de visualización soportados, y proporcionan, tablas, gráficas, mapas, series de tiempo dudando de la utilidad de generar grandes comparaciones visuales. De esta manera, la idea original de Tableau es limitada a la generación de redes en las presentaciones visuales con una granularidad uniforme y con dimensiones limitadas.

Otros investigadores sugieren que la visualización OLAP debe ser enriquecida extendiendo las técnicas básicas del uso de gráficos o mediante el uso de novedosas técnicas para obtener el máximo provecho de las propiedades de los datos multidimensionales y jerárquicos. En [39] [20] [38] [36] se establecieron los requerimientos para la visualización de la información de negocios e hicieron una revisión sobre metáforas visuales avanzadas para datos variantes, tal es el caso de los diagramas de Kiviat y coordenadas paralelas para la visualización de datos multidimensionales, así como las técnicas 3D, gráficas de líneas 3D y mapas 3D basados en gráficas de barras. Una propuesta alternativa es representada por el sistema DIVE-ON(Data mining in an Immersed Visual Enviroment Over a Network) , propuesto en [29]. La idea principal de DIVE-ON es crear un entorno visual donde las fuentes de datos multidimensionales son unificadas y presentadas a los usuarios que pueden interactuar con dichas fuentes. Es así como DIVE-ON utiliza la capacidad natural que tenemos para interactuar con objetos representados en el espacio para mejorar la fase de madurez del conocimiento. Además se explota la tecnología OLAP para el tratamiento de los datos multidimensionales. Considerando lo anterior se puede afirmar que DIVE-ON es una herramienta OLAP útil para la visualización en la investigación.

Otra rama en el campo de la investigación de visualización OLAP está centrada en el desarrollo de niveles dentro de la visualización con el objetivo de presentar los datos en sus diferentes niveles de agregación. [37] describe su implementación de visualización multinivel dentro del entorno de trabajo del sistema Polaris. La

aportación fundamental dentro de lo visual es un gráfico que ayuda a tomar distintos caminos para lograr diversos acercamientos, zoom. [20] proponen una técnica de visualización multidimensional que adopta y modifica el método de coordenadas paralelas para el tratamiento de resultados OLAP. La ventaja principal de dicha técnica es la facilidad de poder usar cualquier número de dimensiones. Cada dimensión se representa por un eje vertical y las agregaciones se alinean a lo largo de cada eje similar a una gráfica de barras. Por otro lado, en el miso eje se pueden generar gráficas de barras aún más detalladas. Las líneas poligonales utilizadas en esta técnica se utilizan para indicar relaciones entre las agregaciones analizadas en distintas dimensiones, existe una relación si los conjuntos de datos en 2 agregaciones se intersecan.

En [23] se concentran en el problema de la pérdida de datos en pasos previos a las consultas al cambiar el nivel de detalle, proponen el uso de diseños jerárquicos para la captura de resultados en las descomposiciones múltiples dentro de una misma vista. Estos dos autores introducen una clase dentro del uso de multiniveles y le denominan árbol de descomposición mejorada. Los niveles dentro de una jerarquía usada para la visualización son creados mediante la descomposición de agregaciones en una dimensión específica, los nodos contienen los resultados del nuevo nivel de agregaciones y se puede observar en una nueva visualización, como una gráfica de barras.

En [36] presenta una técnica de visualización multinivel para OLAP basada en las vistas de las jerarquías de dimensiones. Cada jerarquía con toda su información que representan nodos de nivel inferior se presentan con un diseño de árbol anidado. Las operaciones drill-down y roll-up son utilizadas implícitamente como una especie de zoom dentro de cada vista en las dimensiones. El filtrado de datos se realiza al seleccionar los valores de interés en los niveles de las jerarquías de cada dimensión.

Una técnica similar para la visualización interactiva denominada Visualización Jerárquica Dinámica Dimensional se propone en [38]. Las jerarquías en las dimensiones se muestran de forma alineada en barras. Una barra está dividida en pequeños rectángulos que representan el resultado de la función de agregación para cada miembro en la dimensión. La intensidad en el color utilizado mantiene una relación con el número de registros que existen dentro del rango mostrado. Una técnica para la

representación de los resultados multidimensionales, propuesto por [4] puede ayudar a mejorar el análisis en la visualización. Esta técnica se enfoca en los resultados de las funciones de agregación solicitados a lo largo de los ejes referentes a las dimensiones. De manera predeterminada, el orden para los valores de las medidas se establece alfabéticamente. Para la búsqueda de patrones el usuario debe hacer un reordenamiento manual. El algoritmo propuesto realiza automáticamente el ordenamiento de las medidas de forma que se visualicen mejor patrones de tendencia y similitud.

En [7] propone un entorno de trabajo para apoyar de manera eficiente la visualización OLAP de cubos de datos multidimensionales. Este entorno tiene gran aplicación en la vida real trabajando a partir de la visualización de datos en espacio y tiempo, a diferencia del uso de datos científicos y estadísticos. La novedad consiste en calcular una partición semántica basada en las jerarquías agrupando celdas de datos OLAP que estén semánticamente relacionadas, de esta manera se da origen a los llamados cubos con conciencia semántica. De esta manera se comprimen un poco más los datos y eso genera una visualización más eficiente. Esta representación da origen a un novedoso histograma multidimensional, y es llamada Hierarchy-driven Indexed Quad-Tree Summary (H-IQTS). La principal ventaja del este método es una mejora a las actividades de exploración y visualización de grandes dimensiones permitiendo acceder y navegar en sub particiones basadas en semántica o en algún otro esquema de partición debido a que durante la interacción los usuarios pueden estar interesados sólo en partes específicas del dominio y no en todo.

4.2. Desarrollo del sistema

Este proyecto cuenta con una estructura organizacional de un Sistema de Información (SI), envolviendo una serie de elementos orientados al tratamiento y administración de datos e información para cubrir con las necesidades y objetivos planteados. Los diagramas que forman parte de la realización de este sistema se encuentran en el Apéndice B.

4.2.1. Sistema de información

Un SI es aquel constituido por personas, datos y actividades que procesan datos e información, con un enfoque de acuerdo con un objetivo específico. Con base a lo anterior, un sistema de información logra la automatización de procesos operativos dentro de un sistema informático, estos sistemas se clasifican en:

Sistema de procesamiento de transacciones: (Transactional System, TPS) Sistema de información diseñado para la recopilación, el almacenamiento, la modificación y recuperación de la información que es generada por las transacciones en una organización o en un procedimiento.

Sistemas de soporte a la toma de decisiones: (Decision Support System, DSS) Sistema informático que es utilizado como apoyo, más que para automatizar, un proceso que involucre toma de decisiones.

Sistemas para la toma de decisiones de grupo: (Group Decision Support System, GDSS) Sistema interactivo, el cual facilita la solución de problemas no estructurados por un conjunto de tomadores de decisiones trabajando juntos como un grupo.

Sistemas expertos de soporte a la toma de decisiones: (Expert System Decision Support, EDSS) Sistemas de información basados en el conocimiento, se enfocan sobre un área específica para ser manejado como un consultor experto para los usuarios.

Sistemas de información para ejecutivos: (Executive Information System, EIS) Sistemas de apoyo para funcionarios de alto nivel al momento de dirigir una organización. su objetivo es proporcionar un acceso inmediato y fácil a información específica sobre factores clave que son fundamentales para alcanzar los objetivos estratégicos de una empresa.

Sistemas estratégicos: Sistemas de información considerados en el uso de la tecnología para el soporte de la estrategia competitiva en una organización.

De acuerdo con las características mencionadas anteriormente, los sistemas de información ofrecen variedad de opciones que permiten dar una solución específica

de acuerdo a requerimientos y objetivos planteados por los usuarios. Estos SI deben regirse por el ciclo de desarrollo de los sistemas; un enfoque por etapas de análisis y diseño, el cual especifica que el desarrollo de los sistemas mejora cuando existe un ciclo de actividades del analista y de los usuarios.

4.2.2. Ciclo de desarrollo del sistema

El ciclo de vida de desarrollo de sistemas (The Systems Development Life Cycle, SDLC), es el proceso lógico utilizado por un analista de sistemas para desarrollar un sistema de información, incluyendo, los requisitos, la validación, formación, y la propiedad del usuario.

4.2.3. Fases de desarrollo del sistema

El ciclo de desarrollo técnico de un sistema se divide en las siguientes fases:

Iniciación/Planeación: Con el propósito de determinar los objetivos del proyecto. El estudio de viabilidad se utiliza para exponer el proyecto a la administración con la finalidad de obtener financiación. Se evalúan los proyectos generalmente de manera económica, operativa y técnica.

Análisis de requerimientos: Los sistemas de análisis tienen como objetivo determinar los problemas segmentando el sistema en diversos diagramas de procesos, mostrando así claramente los requisitos de los usuarios.

Diseño de requerimientos: Las diferentes actividades y finciones en el diseño de sistemas se describen a detalle incluyendo diseños de pantalla, reglas de negocio, diagrames de procesos, etc. De esta manera se obtiene una colección de módulos o subsistemas.

Desarrollo del diseño: Durante esta etapa se realiza el código de programación. Va de la mano con la siguiente etapa debido a que los módulos necesitan de estas pruebas antes de ser integradas al proyecto principal.

Pruebas: Para que un módulo de un sistema sea aceptado se realizan pruebas en diferentes niveles lógicos. Esto permite brindar una confiabilidad en las funciones que realizará dicho módulo del sistema.

Operaciones y mantenimiento: La entrega de un sistema incluye también cambios y mejoras antes de iniciar su operación. De la misma manera, mantener el sistema es un aspecto involucrado en el SDLC.

4.2.4. Tecnologías de desarrollo

En esta sección se hace un análisis del lenguaje de programación, bases de datos y entorno de desarrollo integrado (Integrated Development Environment, IDE) utilizados para el desarrollo de esta aplicación.

Lenguaje de programación

Las herramientas que permiten desarrollar aplicaciones de alto nivel debieran proporcionar portabilidad entre plataformas de hardware, optimizar el tiempo de respuesta a los algoritmos y ser adaptables a diferentes entornos de desarrollo.

Entre las características con las que cuenta el lenguaje de programación Java se encuentran las siguientes:

- Brinda independencia de la plataforma. Es decir, puede ejecutarse en cualquier plataforma Windows, Unix (Solaris, Silicon Graphics) y Power/Mac.
- Es uno de los lenguajes mas usados y con la mayor comunidad alrededor del mundo.
- Existen librerías para el tratamiento y manipulación de gráficos, en concreto, se empleó la librería Java 3D.
- Optimiza la velocidad de ejecución y uso de la memoria del procesador.
- Evita la desaceleración provocada para el acceso a recursos fuera de la memoria del procesador.
- Mayor nivel de abstracción.

- Obligatoriedad de estructura modular.
- Programación orientada a objetos.
- Amplio soporte de funciones matemáticas.
- Ejecución de programas multiprocesador.
- Es robusto, lo cual permite detectar los errores en el momento de producirse, lo que facilita la depuración.
- Software de licencia GPL (General Public License).

Java utiliza la tecnología de generación de código en tiempo de ejecución, las instrucciones que implementan a los métodos, se convierten en código nativo directamente en la máquina en la que se ejecuta, esta característica permite que las aplicaciones desarrolladas en Java sean funcionales en múltiples plataformas proporcionando escalabilidad y portabilidad, proporcionando al usuario: transparencia, mantenimiento, control de versiones, independencia del sistema operativo y ejecución local de la aplicación.

Con base en estas características se optó por continuar con este lenguaje, un lenguaje de programación orientado a objetos y de alto nivel.

Base de datos

La importancia de las bases de datos es: almacenar, resguardar y mantener la confidencialidad de la información para soportar un sistema auxiliado por una computadora, proporcionando herramientas de respaldo y recuperación de datos.

Los objetivos generales un sistema de manejador de bases de datos son:

Abstracción de la información: Existen diferentes niveles de abstracción y la finalidad es ahorrarles a los usuarios detalles en cuanto al almacenamiento físico de los datos.

Independencia: Se refiere al poder modificar un esquema físico/lógico de una base de datos sin la necesidad de que estos cambios re realicen directamente en las aplicaciones que dependen de ella.

Consistencia: Cuando no es posible eliminar la redundancia de los datos, se debe verificar que la información repetida se actualice de manera coherente y simultánea.

Seguridad: Se refiere a los permisos con los que deben contar usuarios o grupos de usuarios dependiendo de la categoría, para así garantizar que la información se encuentra segura.

Manejo de transacciones: Se proporcionan herramientas para programar modificaciones en los datos de una manera más simple.

Tiempo de respuesta: Se refiere al tiempo en que tarda en devolver la información requerida y en guardar los cambios que se realicen, el cual debe ser mínimo.

Dentro de las ventajas que ofrece un manejador de bases de datos están:

- La manipulación de grandes volúmenes de datos se realiza de manera sencilla.
- Garantizan que los cambios realizados en la base serán consistentes sin importar si existen errores, para esto se deben políticas de respaldo adecuadas.
- Si se explota de manera adecuada, se pueden disminuir los tiempos de desarrollo y aumentar la calidad del sistema desarrollado.
- La mayoría de las veces proporcionan interfaces y lenguajes de consulta que hacen más fácil la recuperación de los datos.

Actualmente existen una amplia variedad de sistemas manejadores de bases de datos, entre ellos se encuentran PostgreSQL y Miscrosoft SQL Server, los cuales se emplean en la aplicación.

SQL Server ya era utilizado en un inicio por la aplicación, se buscó el uso de un manejador que contara con una licencia GPL por lo que se definió agregar también el uso de PostgreSQL ya que es un sistema de administración de bases de datos relacional orientada a objetos de software libre, multiplataforma, escalable y de amplia capacidad para almacenar y gestionar grandes volúmenes de datos.

Entorno de desarrollo integrado

Un entorno de desarrollo integrado (Integrated Development Environment, IDE) es un programa compuesto por una serie de herramientas que utilizan los programadores para desarrollar código. Esta herramienta puede dar cabida a uno o varios lenguajes de programación simultáneamente.

Las herramientas que componen un IDE son: editor de texto, compilador, intérprete, depurador, herramientas para la automatización, sistema de ayuda para la construcción de interfaces gráficas de usuario (User Interface Graphics, GUI) y un control de versiones.

Actualmente los IDE proporcionan un marco de trabajo para los lenguajes de programación existentes en el mercado. El IDE seleccionado para dar soporte al desarrollo del presente proyecto ha sido Netbeans, dadas sus características de desarrollo de: aplicaciones multiplataforma, bibliotecas para desarrollo de gráficos 3D, cliente CVS integrado y crecimiento de plataforma por medio de plugins.

4.3. SQL para la generación de cubos

Como se revisó en el capítulo anterior, un hecho es una entidad de datos que relaciona una lista de elementos con las medidas de interés, este hecho es tomado de una lista de dimensiones. Los datos pueden ser representados como tuplas en las tablas de hechos. Una base de datos multidimensional comprende un conjunto de dimensiones, junto con una tabla de hechos base, es decir, una tabla de hechos cuyos atributos son las categorías inferiores de las dimensiones. Si la dimensión tiene muchas categorías inferiores, es posible definir una categoría nueva y única, que contiene todos los elementos de fondo, para hacer frente a los hechos básicos.

En OLAP, los usuarios necesitan analizar los hechos agregados en varios niveles de abstracción. La consulta básica que obtiene hechos en base a una granularidad dada por una lista de categorías, uno por dimensión, es conocida como una vista de cubo. Una vista del cubo se puede definir, utilizando una función de agregación, por la consulta siguiente:

```
1: SELECT < Dimension(es) >,
```

```
2: < Funcion(Medida(s)) >
3: FROM < Tabla >
4: GROUP BY < Dimension >
```

En SQL se puede definir un cubo de datos como el conjunto de todas las posibles vistas del cubo definidas sobre una lista de dimensiones, una tabla base, y las medidas agregadas. En el contexto de un cubo de datos, se puede denotar una vista de cubo simplemente como CV [G] donde G es una granularidad (lista de categorías).

Con esta idea principal, se muestra a continuación el procedimiento y cómo se emplearon las consultas SQL para la generación de cubos en cada una de las pruebas estadísticas.

4.3.1. Tablas en común

En todas las pruebas la tabla "Result" es creada al inicio como se muestra en las siguientes líneas, el número de dimensiones Dn varía de acuerdo a las que hayan sido seleccionadas, los demás campos servirán de apoyo para el cálculo de totales.

```
1: CREATE TABLE < T >_Result (
2: id int DEFAULT NEXTVAL('seq_< T >_Result') NOT NULL
3: , < D1 > int
4: , < D2 > int
5: , < D3 > int
6: , N float
7: , L_< M > float
8: , Q_< M > float
9: , PRIMARY KEY
10: (id));
```

Para llenar la tabla creada, se hace una consulta a la tabla principal con los datos estadísticos, es aquí donde se obtienen todas las "poblaciones" ya que se hace una agrupación por los valores de las dimensiones, se obtiene el total de cada agrupación, la sumatoria del valor correspondiente al campo de la medida elegida, así como la sumatoria de cuadrados, esto con el objetivo de calcular valores estadísticos más adelante.

```
1: INSERT INTO < T >_Result
2: (< D1 >, < D2 >, < D3 >, N, L_< M >, Q_< M >)
3: SELECT
4: < D1 >
5: , < D2 >
6: , < D3 >
7: , sum(1.0) as N
8: , sum(< M >)
9: , sum(< M >*< M >)
10: FROM < T >
11: GROUP BY < D1 >, < D2 >, < D3 >;
```

De la tabla "Result" se hace una selección y se crea una nueva tabla "tempMSV" donde además de obtener las agrupaciones y el total de miembros en ellas, se obtienen datos estadísticos como son la media, varianza y desviación estándar.

```
1: SELECT
2: N
3: , < D1 >
4: , < D2 >
5: , < D3 >
6: , (L_< M >/N) as MEAN_< M >
7: , ((Q_< M >/N) - (L_< M >/N)*(L_< M >/N)) as VARIANCE_< M >
8: , sqrt(((Q_< M >/N) - (L_< M >/N)*(L_< M >/N))) as STD_< M >
9: INTO tempMSV FROM (
10: SELECT
11: *
```

```
12: FROM < T >_Result
13: WHERE ( (< D1 > <>-1) ) AND ( (< D2 > <>-1) ) AND ( (< D3 > <>-1)
    )
14: ) AS t1;
```

4.3.2. Media de una sola población

Posterior a crear como primer paso las tablas en común para las pruebas, se crea una nueva tabla donde se obtendrán los valores para el estadístico en cuestión y cada registro contendrá la población analizada, las dimensiones de la población se pueden identificar en las líeas 5 a 7.

```
1: CREATE TABLE < T >_TempZ (
2: PValue_< M > float
3: , ZTest_< M > float
4: , N int
5: , < D1 > int
6: , < D2 > int
7: , < D3 > int);
```

Después de crear la tabla, se hace el insert de los valores, líneas 1 y 2 tomados a partir de un select a la tabla "tempMSV", líneas 3 a 9, los valores insertados corresponden al valor estadístico Z de la prueba, en la línea 4 se observa la fórmula para obtener el estadístico, el valor V es el valor contra el que se está haciendo la comparación el cual es indicado en la aplicación, en las líneas 6 a 8 se indican las dimensiones de la población analizada, la sentencia where, línea 10, filtra las agrupaciones donde haya 25 o más integrantes, esto debido a la suposición de una distribución normal.

```
1: INSERT INTO < T >_TempZ
2: (ZTest_< M >, N, < D1 >, < D2 >, < D3 >)
3: SELECT
4: ((mean_< M > - < V >)/(std_< M > / sqrt(N) ))
5: , N
6: , < D1 >
```

```
 7: , < D2 >
 8: , < D3 >
 9: FROM tempMSV
10: WHERE ( N>=25 );
```

Se crea una tabla "Final" mediante una sentencia select into, línea 12, que contendrá los resultados obtenidos, a partir de la tabla anterior, línea 13. Dependiendo del resultado se compara con valores extraídos de la tabla de distribución para Z de acuerdo al grado de significancia establecido, líneas 1 a 5, los registros en donde se haya insertado un valor igual a "3" serán los registros donde la hipótesis ha sido rechazada.

```
 1: SELECT
 2: CASE WHEN (abs(ZTest_< M >)<1.88) THEN '0'
 3: WHEN (abs(ZTest_< M >)<1.96) THEN '1'
 4: WHEN (abs(ZTest_< M >)<2.04) THEN '2'
 5: ELSE '3'
 6: END AS PValue_< M >
 7: ,ZTest_< M >
 8: , N
 9: , < D1 >
10: , < D2 >
11: , < D3 >
12: INTO < T >_Final
13: FROM < T >_TempZ;
```

4.3.3. Diferencia entre medias de dos poblaciones

A partir de la tabla "tempMSV" se toman dos instancias de ella, líneas 17 y 18, para generar una intersección y crear una nueva, línea 16, para esto se hace una selección con la clausula "where", como se aprecia en las líneas 19 a 26, donde se fija en una instancia el valor de una dimensión y este valor tiene que ser diferente en la otra, esto es para cada una de las dimensiones por lo que dependiendo de la selección inicial esta sentencia crece. El objetivo de esta selección es identificar las

poblaciones que varíen únicamente en una sola dimensión.

```
1: SELECT
2: t1.N AS N1
3: , t1.< D1 > AS < D1 >1
4: , t1.< D2 > AS < D2 >1
5: , t1.< D3 > AS < D3 >1
6: , t1.mean_< M > as MEAN_< M >1
7: , t1.variance_< M > as VARIANCE_< M >1
8: , t1.std_< M > as STD_< M >1
9: , t2.N AS N2
10: , t2.< D1 > AS < D1 >2
11: , t2.< D2 > AS < D2 >2
12: , t2.< D3 > AS < D3 >2
13: , t2.mean_< M > as MEAN_< M >2
14: , t2.variance_< M > as VARIANCE_< M >2
15: , t2.std_< M > as STD_< M >2
16: INTO tempCompare
17: FROM tempMSV AS t1
18: , tempMSV AS t2
19: WHERE (
20: (t1.< D1 > <>t2.< D1 > AND t1.< D1 >=0 AND t1.< D2 >=t2.< D2 > AND
    t1.< D3 >=t2.< D3 >)
21: OR (t1.< D1 > <>t2.< D1 > AND t1.< D1 >=1 AND t1.< D2 >=t2.< D2 >
    AND t1.< D3 >=t2.< D3 >)
22: OR (t1.< D1 >=t2.< D1 > AND t1.< D2 > <>t2.< D2 > AND t1.< D2 >=0
    AND t1.< D3 >=t2.< D3 >)
23: OR (t1.< D1 >=t2.< D1 > AND t1.< D2 > <>t2.< D2 > AND t1.< D2 >=1
    AND t1.< D3 >=t2.< D3 >)
24: OR (t1.< D1 >=t2.< D1 > AND t1.< D2 >=t2.< D2 > AND t1.< D3 > <>t2.<
    D3 > AND t1.< D3 >=0)
25: OR (t1.< D1 >=t2.< D1 > AND t1.< D2 >=t2.< D2 > AND t1.< D3 > <>t2.<
    D3 > AND t1.< D3 >=1)
26: );
```

Ahora se crea una nueva tabla donde se obtendrán los valores para el estadístico en cuestión y cada registro contendrá las poblaciones que se están comparando, las dimensiones de la población 1 se pueden identificar en las líneas 5 a 7 y para la población 2 en las líneas 9 a 11.

```
1: CREATE TABLE <T>_TempZ (
2: PValue_<M> float
3: , ZTest_<M> float
4: , N1 int
5: , <D1>1 int
6: , <D2>1 int
7: , <D3>1 int
8: , N2 int
9: , <D1>2 int
10: , <D2>2 int
11: , <D3>2 int);
```

Una vez creada la tabla, se hace el "insert" de los valores, líneas 1 y 2 tomados a partir de un "select" a la tabla "tempCompare", líneas 3 a 14, los valores insertados corresponden al valor estadístico Z de la prueba y las poblaciones analizadas.

```
1: INSERT INTO <T>_TempZ
2: (ZTest_<M>, N1, <D1>1, <D2>1, <D3>1, N2, <D1>2, <D2>2,
   <D3>2)
3: SELECT
4: ((mean_<M>1-mean_<M>2)/(sqrt((variance_<M>1/N1)+(variance_<
   M>2/N2))))
5: , N1
6: , <D1>1
7: , <D2>1
8: , <D3>1
9: , N2
10: , <D1>2
11: , <D2>2
12: , <D3>2
```

```
13: FROM tempCompare
14: WHERE ( N1>=25 AND N2>=25 );
```

La tabla "Final" es obtenida mediante una sentencia "select into", línea 16, a partir de la tabla anterior, línea 17, y contendrá los resultados obtenidos. Dependiendo del resultado se compara con valores extraídos de la tabla de distribución para Z de acuerdo al grado de significancia establecido, líneas 1 a 5, los registros en donde se haya insertado un valor igual a "3" serán los registros donde la hipótesis ha sido rechazada y es donde se ha encontrado una relación significativa entre las poblaciones.

```
1: SELECT
2: CASE WHEN (abs(ZTest_< M >)<1.88) THEN '0'
3: WHEN (abs(ZTest_< M >)<1.96) THEN '1'
4: WHEN (abs(ZTest_< M >)<2.04) THEN '2'
5: ELSE '3'
6: END AS PValue_< M >
7: , ZTest_< M >
8: , N1
9: , < D1 >1
10: , < D2 >1
11: , < D3 >1
12: , N2
13: , < D1 >2
14: , < D2 >2
15: , < D3 >2
16: INTO < T >_Final
17: FROM < T >_TempZ;
```

4.3.4. Proporción para una población

Para esta prueba estadística, la tabla "tempMSV" es un poco diferente, no se calculan valores como la media, desviación estándar o varianza debido a que sólo se involucran totales, es por ello que sólo se crea el campo NT, línea 6.

```
1: SELECT
2: N
3: , < D1 >
4: , < D2 >
5: , < D3 >
6: , 1 as NT
7: INTO tempMSV FROM (
8: SELECT
9: *
10: FROM < T >_Result
11: WHERE ( (< D1 > <>-1) ) AND ( (< D2 > <>-1) ) AND ( (< D3 > <>-1)
    )
12: ) AS t1;
```

El campo NT comentado anteriormente se obtiene a partir del siguiente "select", líneas 1 a 3 y es actualizado, líneas 5 a 7.

```
1: SELECT
2: sum(n) as n
3: FROM tempMSV;
4:
5: UPDATE
6: tempMSV
7: SET nt = < N >;
```

Similar a las pruebas anteriores, se crea la tabla "TempZ".

```
1: CREATE TABLE < T >_TempZ (
2: PValue float
3: , ZTest float
4: , N int
5: , < D1 > int
6: , < D2 > int
7: , < D3 > int);
```

Para esta tabla "TempZ" donde se alojan los valores del estadístico obtenido, esta vez se obtiene utilizando el parámetro P dado mediante la aplicación, línea 4, de igual manera se almacenan las dimensiones que conforman la población en cuestión, líneas 6 a 8.

```
1: INSERT INTO < T >_TempZ
2: (ZTest, N, < D1 >, < D2 >, < D3 >)
3: SELECT
4: ((N/NT) - (< P >)) / sqrt(((< P >) * (1 - (N/NT))) / (N) )
5: , N
6: , < D1 >
7: , < D2 >
8: , < D3 >
9: FROM tempMSV
10: WHERE ( N>=25 );
```

Se hacen las comparaciones debidas, líneas 1 a 5 para determinar si la hipótesis es rechazada.

```
1: SELECT
2: CASE WHEN (abs(ZTest)<1.88) THEN '0'
3: WHEN (abs(ZTest)<1.96) THEN '1'
4: WHEN (abs(ZTest)<2.04) THEN '2'
5: ELSE '3'
6: END AS PValue
7: , ZTest
8: , N1
9: , < D1 >
10: , < D2 >
11: , < D3 >
12: INTO < T >_Final
13: FROM < T >_TempZ;
```

4.3.5. Diferencia entre las proporciones de dos poblaciones

Así como la prueba anterior, sólo se toman en cuenta totales y al involucrar dos poblaciones se crean los campos $NT1$ y $NT2$, líneas 10 y 11, el filtro que se hace, líneas 14 a 21, obtiene las poblaciones donde únicamente varíen en una dimensión.

```
1: SELECT
2: t1.N AS N1
3: , t1.< D1 > AS < D1 >1
4: , t1.< D2 > AS < D2 >1
5: , t1.< D3 > AS < D3 >1
6: , t2.N AS N2
7: , t2.< D1 > AS < D1 >2
8: , t2.< D2 > AS < D2 >2
9: , t2.< D3 > AS < D3 >2
10: , 0 as NT1
11: , 0 as NT2
12: INTO tempMSV
13: FROM < T >_Result AS t1
14: ,< T >_Result AS t2 WHERE (
15: (t1.< D1 > <>t2.< D1 > AND t1.< D1 >=0 AND t1.< D2 >=t2.< D2 > AND
    t1.< D3 >=t2.< D3 >) OR
16: (t1.< D1 > <>t2.< D1 > AND t1.< D1 >=1 AND t1.< D2 >=t2.< D2 > AND
    t1.< D3 >=t2.< D3 >) OR
17: (t1.< D1 >=t2.< D1 > AND t1.< D2 > <>t2.< D2 > AND t1.< D2 >=0 AND
    t1.< D3 >=t2.< D3 >) OR
18: (t1.< D1 >=t2.< D1 > AND t1.< D2 > <>t2.< D2 > AND t1.< D2 >=1 AND
    t1.< D3 >=t2.< D3 >) OR
19: (t1.< D1 >=t2.< D1 > AND t1.< D2 >=t2.< D2 > AND t1.< D3 > <>t2.<
    D3 > AND t1.< D3 >=0) OR
20: (t1.< D1 >=t2.< D1 > AND t1.< D2 >=t2.< D2 > AND t1.< D3 > <>t2.<
    D3 > AND t1.< D3 >=1)
21: );
```

Lo siguiente es un procedimiento que se realiza con el fin de actualizar los valores de los totales $NT1$ y $NT2$ de cada relación de poblaciones, se sabe que en cada registro existen dos poblaciones que varían en una dimensión D, por lo que se toman estos valores $v1a$ y $v1b$, líneas 4 y 9, que son diferentes y se obtienen los totales Na y Nb. Estos valores son los que se utilizan para actualizar $NT1$ y $NT2$, línea 13, de la tabla "tempMSV", esto es solamente en poblaciones donde una dimensión tenga los valores $v1a$ y $v1b$ y las demás dimensiones sean iguales, como se puede ver en la condición "where" de las líneas 15 a 17, este procedimiento se lleva a cabo para cada registro encontrado.

```
1: SELECT
2: sum(n) as n
3: FROM tempMSV
4: WHERE ( < D > = < v1a > );
5:
6: SELECT
7: sum(n) as n
8: FROM tempMSV
9: WHERE ( < D > = < v1b > );
10:
11: UPDATE
12: tempMSV
13: SET nt1 = < Na >, nt2 = < Nb >
14: WHERE
15: < D1 >1 = < v1a > AND < D1 >2 = < v1b > AND
16: < D2 >1 = < v2 > AND < D2 >2 = < v2 > AND
17: < D3 >1 = < v3 > AND < D3 >2 = < v3 >;
```

Para continuar se crea la tabla conocida de procedimientos anteriores "TempZ".

```
1: CREATE TABLE < T >_TempZ (
2: PValue float
3: , ZTest float
4: , N1 int
```

```
5: , < D1 >1 int
6: , < D2 >1 int
7: , < D3 >1 int
8: , N2 int
9: , < D1 >2 int
10: , < D2 >2 int
11: , < D3 >2 int );
```

Para la inserción de los valores se hace el cálculo correspondiente al estadístico, línea 4.

```
1: INSERT INTO < T >_TempZ
2: (ZTest, N1, < D1 >1, < D2 >1, < D3 >1, N2, < D1 >2, < D2 >2, <
   D3 >2)
3: SELECT
4: (( n1/nt1) - (n2/nt2) ) / sqrt( ((1.0/nt1) + (1.0/nt2)) * ((n1+n2)/(nt1+nt2))
   * (1.0-((n1+n2)/(nt1+nt2))) )
5: , N1
6: , < D1 >1
7: , < D2 >1
8: , < D3 >1
9: , N2
10: , < D1 >2
11: , < D2 >2
12: , < D3 >2
13: FROM tempMSV
14: WHERE ( N1>=25 AND N2>=25 );
```

Con los valores obtenidos se hacen las comparaciones pertinentes, líneas 1 a 5, y se determina el resultado para la hipótesis.

```
1: SELECT
2: CASE WHEN (abs(ZTest)<1.88) THEN '0'
3: WHEN (abs(ZTest)<1.96) THEN '1'
4: WHEN (abs(ZTest)<2.04) THEN '2'
```

```
5: ELSE '3'
6: END AS PValue
7: , ZTest
8: , N1
9: , < D1 >1
10: , < D2 >1
11: , < D3 >1
12: , N2
13: , < D1 >2
14: , < D2 >2
15: , < D3 >2
16: INTO < T >_Final
17: FROM < T >_TempZ;
```

4.3.6. Varianza de una sola población

Después de crear las tablas comunes, se crea la tabla "TempZ" para obtener los resultados correspondientes, al involucrar en la prueba a sólo una población, dependiendo del número n de dimensiones se crean los campos necesarios, como en las líneas 5 a 7.

```
1: CREATE TABLE < T >_TempZ (
2: PValue_< M > float
3: , ZTest_< M > float
4: , N int
5: , < D1 > int
6: , < D2 > int
7: , < D3 > int);
```

En la inserción de resultados, se obtiene el estadístico involucrado utilizando el valor V que se da en un inicio para la prueba, línea 4.

```
1: INSERT INTO < T >_TempZ
2: (ZTest_< M >, N, < D1 >, < D2 >, < D3 >)
3: SELECT
4: ( ((N - 1) * (variance_< M >)) / < V >)
```

```
5: , N
6: , < D1 >
7: , < D2 >
8: , < D3 >
9: FROM tempMSV
10: WHERE ( N>=25 );
```

Creación de la tabla "Final" para la comparación de hipótesis.

```
1: SELECT
2: PValue_< M >
3: , ZTest_< M >
4: , N
5: , < D1 >
6: , < D2 >
7: , < D3 >
8: INTO < T >_Final
9: FROM < T >_TempZ;
```

Para obtener los resultados en esta prueba, se hace a través de un procedimiento, esto es debido a que los grados de libertad involucrados en cada prueba de acuerdo al número de dimensiones pueden variar. Se obtienen todos los registros de la tabla "Final" como se observa en las líneas 1 a 8, para cada registro se hace la comparación debida y al final se actualiza el campo "PValue" que registra si se acepta o rechaza la hipótesis, líneas 10 a 15.

```
1: SELECT
2: PValue_< M >
3: , ZTest_< M >
4: , N
5: , < D1 >
6: , < D2 >
7: , < D3 >
8: FROM < T >_Final
9:
```

```
10: UPDATE
11: < T >_Final
12: SET
13: PValue_< M >=< V >
14: WHERE
15: < D1 >1=< v1 > AND < D2 >1=< v2 > AND < D3 >1=< v3 >
```

4.3.7. Razón de las varianzas de dos poblaciones

Se crea la tabla para obtener resultados al realizar el cálculo del estadístico, esta vez si se emplean campos para dos poblaciones diferentes, las cuales serán comparadas.

```
1: CREATE TABLE < T >_TempZ (
2: PValue_< M > float
3: , ZTest_< M > float
4: , N1 int
5: , < D1 >1 int
6: , < D2 >1 int
7: , < D3 >1 int
8: , N2 int
9: , < D1 >2 int
10: , < D2 >2 int
11: , < D3 >2 int);
```

Con el siguiente "insert" se hace la operación necesaria, se utiliza una sentencia "case" dado que para esta prueba se requiere que la varianza de mayor valor sea la que se encuentre en la posición del numerador, líneas 3 a 9.

```
1: INSERT INTO < T >_TempZ
2: (ZTest_< M >, N1, < D1 >1, < D2 >1, < D3 >1, N2, < D1 >2, < D2 >2,
   < D3 >2)
3: SELECT
4: CASE WHEN (variance_< M >1 > variance_< M >2) THEN (variance_< M >1
   / variance_< M >2)
```

```
5: ELSE (variance_< M >2 / variance_< M >1 )
6: END AS Ztest_< M >
7: , CASE WHEN (variance_< M >1 > variance_< M >2 ) THEN N1
8: ELSE N2
9: END AS N1
10: , < D1 >1
11: , < D2 >1
12: , < D3 >1
13: , N2
14: , < D1 >2
15: , < D2 >2
16: , < D3 >2
17: FROM tempCompare
18: WHERE ( N1>=25 AND N2>=25 );
```

Creación de la tabla "Final", a partir de un "select into", para la determinación de resultados.

```
1: SELECT
2: PValue_< M >
3: , ZTest_< M >
4: , N1
5: , < D1 >1
6: , < D2 >1
7: , < D3 >1
8: , N2
9: , < D1 >2
10: , < D2 >2
11: , < D3 >2
12: INTO < T >_Final
13: FROM < T >_TempZ
```

Similar al procedimiento de la prueba para una varianza, se hace una consulta para obtener todos los registros, líneas 1 a 12, y determinar si se acepta o se rechaza la hipótesis para posteriormente actualizar el campo correspondiente líneas 14 a 21,

la razón es la misma, debido a que el grado de libertad varía de acuerdo al número de dimensiones, los valores que se requieren para la comparación también son diferentes.

```
1: SELECT
2: PValue_< M >
3: , ZTest_< M >
4: , N1
5: , < D1 >1
6: , < D2 >1
7: , < D3 >1
8: , N2
9: , < D1 >2
10: , < D2 >2
11: , < D3 >2
12: FROM < T >_Final;
13:
14: UPDATE
15: < T >_Final
16: SET
17: PValue_< M >=3
18: WHERE
19: < D1 >1=< v1a > AND < D1 >2=< v1b > AND
20: < D2 >1=< v2a > AND < D2 >2=< v2b > AND
21: < D3 >1=< v3a > AND < D3 >2=< v3b >;
```

4.3.8. Chi-cuadrada (prueba de independencia)

Se fija el valor de la dimensión con la cual se quiere realizar la prueba para determinar si guarda alguna relación, línea 6, y nuevamente se realiza el insert en la tabla "Result".

```
1: INSERT INTO < T >_Result
2: (< D1 >, < D2 >, < D3 >, N)
3: SELECT
```

```
4: < D1 >
5: , < D2 >
6: , -1
7: , sum(N) as N
8: FROM (SELECT
9: *
10: FROM < T >_Result
11: WHERE ( ((< D1 > <>-1) OR (< D1 > is null)) )
12: AND ( ((< D2 > <>-1) OR (< D2 > is null)) )
13: AND ( ((< D3 > <>-1) OR (< D3 > is null)) ) ) as t
14: , (SELECT
15: count(*) as cnt
16: FROM < T >_Result
17: ) as t1
18: GROUP BY < D1 >, < D2 >;
```

Para crear la tabla tempMSV se complementa el nombre con las dimensiones con las que se esta haciendo la comparación tempMSV$<$ $D1$ $><$ $D2$ $>$, línea 10, además se agregan los campos $T1$ y $T2$ que servirán para obtener totales, líneas 5 y 6.

```
1: SELECT
2: N
3: , < D1 >
4: , < D2 >
5: , 0 as T1
6: , 0 as T2
7: INTO tempMSV< D1 >< D2 > FROM (
8: SELECT
9: *
10: FROM < T >_Result
11: WHERE ( (< D1 > <> -1) ) AND ( (< D2 > <> -1) )
12: ) AS t1
13: WHERE ( (< D3 > = -1) );
```

Para obtener los totales se realiza un procedimiento. De la tabla result, se extraen las combinaciones existentes de los valores de las dimensiones, si existe, simplemente

se actualiza el valor de los totales T1 y T2, de lo contrario se hace un insert en la tabla y posteriormente se actualizan los totales.

Selección de las diferentes combinaciones.

```
1: SELECT distinct D1 from tempMSV< D1 >< D2 >;
2:
3: SELECT distinct D2 from tempMSV< D1 >< D2 >;
```

Se determina si existe la combinación.

```
1: SELECT ‘existás exist FROM tempMSV< D1 >< D2 >
2: WHERE < D1 > = 1 AND < D2 > = 1;
```

Si existe la comibinación en la tabla, se hacen las actualizaciones.

```
1: UPDATE tempMSV< D1 >< D2 >
2: SET T1 = ( SELECT count(1) FROM < T > WHERE < D1 > = < v1 > )
3: WHERE < D1 > = < v1 > AND < D2 > = < v2 >;
4:
5: UPDATE tempMSV< D1 >< D2 >
6: SET T2 = ( SELECT count(1) FROM < T > WHERE < D2 > = < v2 > )
7: WHERE < D1 > = < v1 > AND < D2 > = < v2 >;
```

De lo contrario se hace el insert.

```
1: INSERT INTO tempMSV< D1 >< D2 > VALUES(< v1 >,< v2 >,< V >);
```

Para la parte final, se obtiene el total de registros en la base y se aplica la operación para obtener las frecuencias esperadas, esto simula el obtener la tabla de frecuencias durante el procedimiento establecido para esta prueba.

```
1: SELECT
2: count(1) as TOTAL
3: FROM < T >
4: WHERE ( < D1 > is NOT NULL AND < D2 > is NOT NULL );
```

```
1: SELECT n ,< D1 >,< D2 >,t1,t2,
2: (cast(t1 as float)*cast(t2 as float))/< TOTAL > as fe
3: into tempCompare< D1 >< D2 >
4: FROM tempMSV< D1 >< D2 >;
```

Para crear la tabla "TempZ" se hacen los cálculos correspondientes para la prueba y obtener el estadístico.

```
1: INSERT INTO < T >_TempZ< D1 >< D2 >
2: (ZTest, < D1 >1, < D2 >1, < D1 >2, < D2 >2) SELECT sum((n-fe)*(n-fe)/fe)
   , < D1 >1 , < D2 >1 , < D1 >2 , < D2 >2 FROM tempCompare< D1 >< D2 >;
```

Por último, para la tabla "Final" se hacen las comparaciones necesarias para determinar si se acepta o se rechaza la prueba.

```
1: SELECT
2: CASE WHEN (abs(ZTest)<4.709) THEN '0'
3: WHEN (abs(ZTest)<5.024) THEN '1'
4: WHEN (abs(ZTest)<5.412) THEN '2'
5: ELSE '3ÉND AS PValue, ZTest
6: < D1 >1
7: , < D2 >1
8: , < D1 >2
9: , < D2 >2
10: INTO < T >_Final< D1 >< D2 >
11: FROM < T >_TempZ< D1 >< D2 >;
```

Capítulo 5

Experimentación

5.1. Plataforma

Los experimentos fueron realizados con la aplicación a la que se ha hecho mención durante la introducción, a continuación se presentan las características del equipo de cómputo en donde dicha aplicación fue instalada para su uso.

5.1.1. Hardware

Se utilizó una notebook con las siguientes características:

Procesador: Intel Celeron 925 2.30 Ghz.

Disco duro: 500 Gb.

Memoria: DDR3 de 3 Gb.

5.1.2. Software

En cuanto a los programas necesarios para llevar a cabo los experimentos, se contó con lo siguiente:

- Sistema operativo Windows 7 Professional, versión de 32-bit.
- Sistema manejador de bases de datos:

  - PostgreSQL 9.0.
  - Microsft SQL Server 2005.

- Java en su versión 6.0_29, además de la librería J3D para la visualización de los gráficos en 3D y los respectivos conectores JDBC hacia los manejadores de las bases de datos (PostgreSQL y SQL Server).

- Aplicación desarrollada

5.2. Base de datos

Para los experimentos se utilizó una base de datos compuesta por 2 tablas.

5.2.1. Tabla 1 (zmed655)

Esta tabla contiene datos personales acerca del historial médico de algunos pacientes, el historial médico cuenta con características de los pacientes tales como el sexo, si fuma o no, si padece de diabetes, etc., estos datos son numéricos y depende de la codificación establecida lo que representa cada uno, es decir, un 1 en el campo de sexo indica que es de sexo masculino y un 0 femenino, otros datos están relacionados con mediciones acerca de las arterias del corazón, indican en porcentaje el estrechamiento que existe en determinada arteria, por ejemplo el campo lm indica la medida para la arteria principal izquierda y puede tener un valor de 0 - 100 donde 100 es el máximo nivel de estrechamiento.

Los datos utilizados en los artículos [27] y [28] fueron un conjunto de 256 registros. En el Apéndice A se muestra el ejercicio donde se verificó y se llegó a la conclusión de que los datos no presentan una distribución normal, hecha esta comprobación se ingresaron más datos ficticios y se utilizaron solamente con el fin de mostrar durante el experimento las características de la aplicación, por lo que las conclusiones a las que se llegan no tienen ningún fundamento para asegurar que son verdaderas. En total son 313 registros los que contiene la tabla.

Su estructura se muestra en la Figura 5.1

5.2.2. Tabla 2 (tbbm)

Es similar a la tabla anterior, ya que contiene la historia clínica de algunos pacientes, en este caso, son pacientes que padecen de tuberculosis y a diferencia de la anterior, los datos son reales pero solamente contiene variables categóricas por lo que no se realizó la misma prueba de distribución que con los anteriores. El propósito de tener esta tabla es el de realizar un análisis con la aplicación aquí descrita, mostrando únicamente la funcionalidad al elegir la prueba de hipótesis chi-cuadrada, para comparar los resultados que se obtengan contra los que se consiguieron al momento de realizar una investigación nacional de salud, ya que dicha tabla contiene los mismos datos que fueron recopilados en el trabajo mencionado.

En total son 142 campos y 1262 registros, en la Figura 5.2 se muestran sólo los campos que se verán involucrados durante el segundo experimento:

zmed655	
rid	Integer
oldyn	Integer
sex	Integer
hta	Integer
diab	Integer
hyplpd	Integer
fhcad	Integer
smoke	Integer
claudi	Integer
pangio	integer

pstroke	Integer
pcarsur	Integer
highchol	Integer
lm	Integer
lad	Integer
lcx	Integer
rca	Integer
al	Double
as_	Double
sa	Double

si	Double
is_	Double
il	Double
li	Double
la	Double
ap	Double
lm2	Integer
diabb	Integer
oldynb	Integer
sexb	integer

Figura 5.1: Tabla 1

tbbm	
folio	character varying(7)
sexo	numeric(10)
escolrec	numeric(10)
peso	numeric(73)
vih	numeric(10)
derechoh	numeric(10)
fiebre	numeric(10)
pdia3	numeric(10)
expua2	numeric(10)
alcohol	numeric(10)
fuma	numeric(10)

Figura 5.2: Tabla 2

5.3. Experimentos

5.3.1. Experimento 1 (datos manipulados)

Para mostrar las pantallas de la aplicación, las bondades así como las funciones que se presentan en este, se realizó un experimento con una base de datos simulando datos de pacientes de un hospital con su historia clínica.

Al iniciar, permite elegir entre dos manejadores de bases de datos diferentes, PostgreSQL y SQLserver, originalmente la aplicación sólo utilizaba SQLserver, esta funcionalidad se agregó dado que en ocasiones no se cuenta con una licencia de Microsoft SQLserver, y a diferencia de este Postgres es un manejador para el cual no se requiere adquirir una licencia.

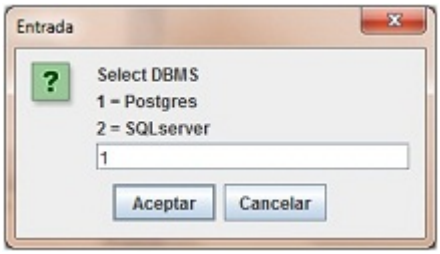

Figura 5.3: Selección del manejador de bases de datos

El siguiente paso es indicar la tabla donde se encuentran todos los registros con los que se trabajará, esta funcionalidad también es nueva dado que anteriormente la aplicación estaba casada con una sola tabla.

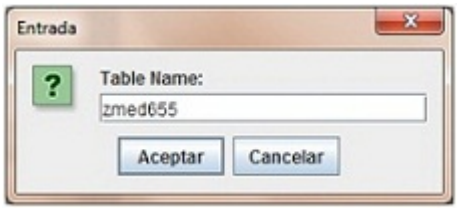

Figura 5.4: Selección de la tabla con la que se trabaja

A continuación se extraen de la tabla indicada todos los campos en ella, dado que el usuario conoce dichos campos, se debe hacer una distinción de ellos para determinar si representan una dimensión o una medida, no es necesario clasificar todos los campos, únicamente los necesarios en el experimento.

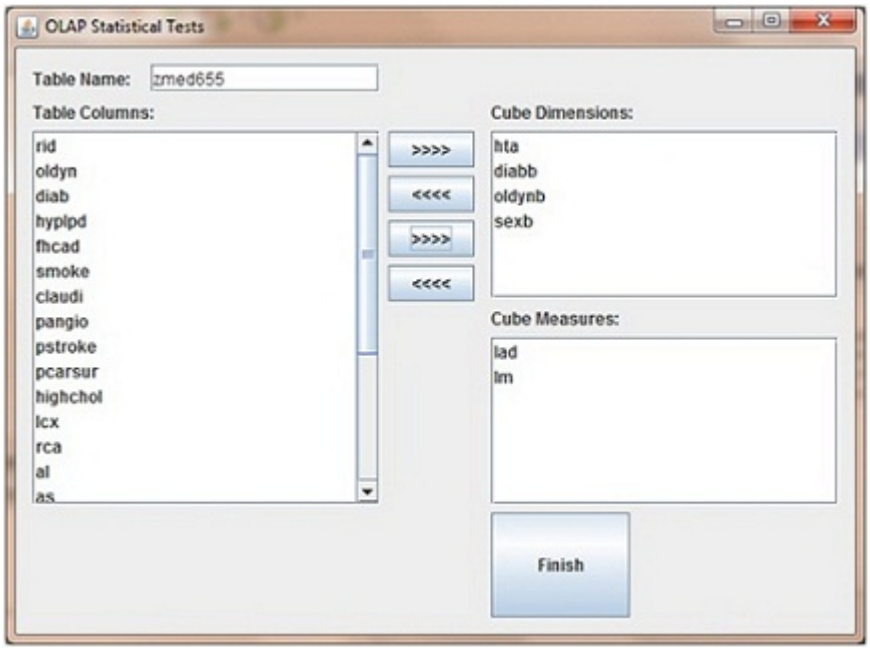

Figura 5.5: Clasificación de las columnas en dimensiones y medidas

Antes de comenzar el análisis, es necesario indicar que dimensiones y medidas de las clasificadas anteriormente se utilizarán, así como la prueba estadística que se aplicará.

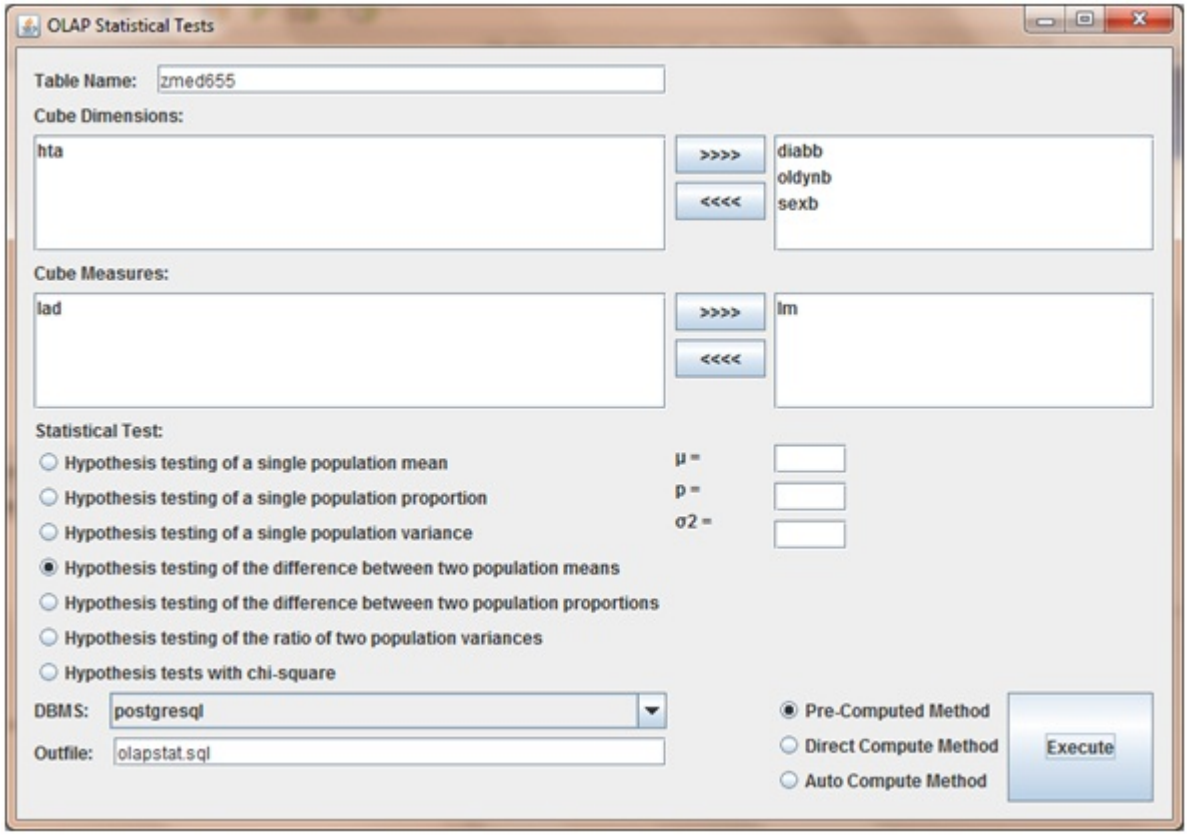

Figura 5.6: Selección de dimensiones, medidas y prueba estadística

Se obtiene una nueva pantalla con los resultados, en la parte superior indica la prueba estadística que se empleó, en la sección más grande se muestra el cubo se crea dependiendo de las dimensiones seleccionadas, en este caso fueron 3, dependiendo de los distintos valores encontrados en los registros será el tamaño del cubo hacia cada dimensión, en este caso se encontraron 4 valores distintos para cada una, en la parte lateral se muestran las gráficas de distribución de los datos, estas varían de acuerdo al cuboide que pueda ser seleccionado. Finalmente la parte que indica los resultados encontrados son los cuboides que tienen un color distinto a la mayoría.

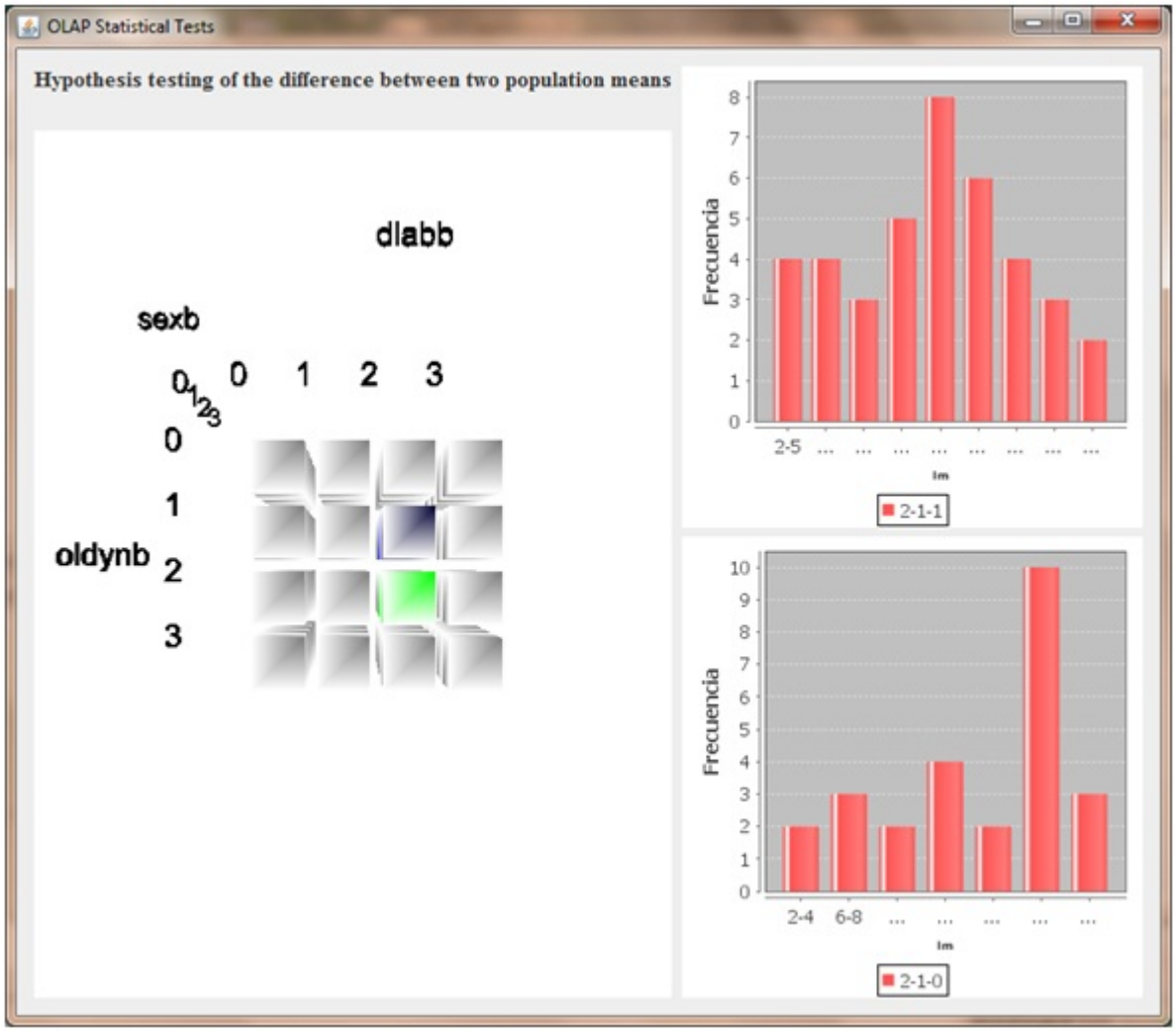

Figura 5.7: Pantalla de resultados

Cada cuboide representa una población distinta, las poblaciones que se observan del mismo color indican que la hipótesis nula que dice que "El par de poblaciones contrastado no mantiene ninguna relación" fue rechazada, por lo que se acepta la hipótesis alternativa y visible ahora qué poblaciones guardan una relación, del mismo modo se aprecia en que dimensión es donde dichas poblaciones tienen un valor distinto que las hace diferentes, y en consecuencia se intuye que esta característica es la que influye para que exista la relación encontrada.

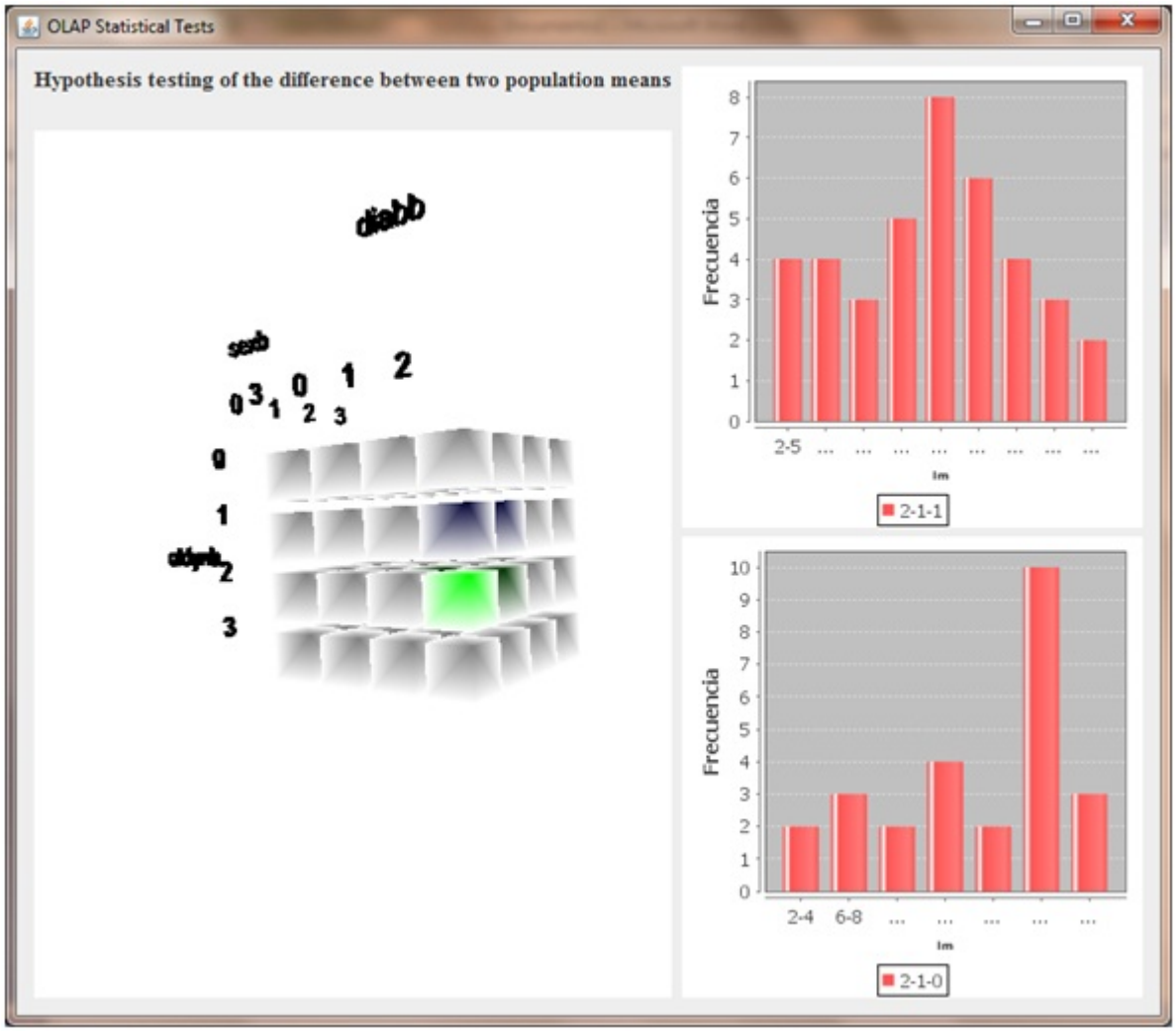

Figura 5.8: Cuboides que mantienen una relación en color

Desde la imagen anterior se observa que las poblaciones se movieron de lugar, esta es otra funcionalidad con la que se cuenta. Debido a que el cubo se puede extender dependiendo de los valores encontrados, se puede dar el caso que algunas poblaciones queden un tanto escondidas, por esta razón los cuboides se pueden recorrer hacia la izquierda, derecha, arriba o abajo para llevarlos hacia una cara visible y apreciar las relaciones existentes. De igual forma, el cubo puede girarse y se cuenta así con una vista de 360^o.

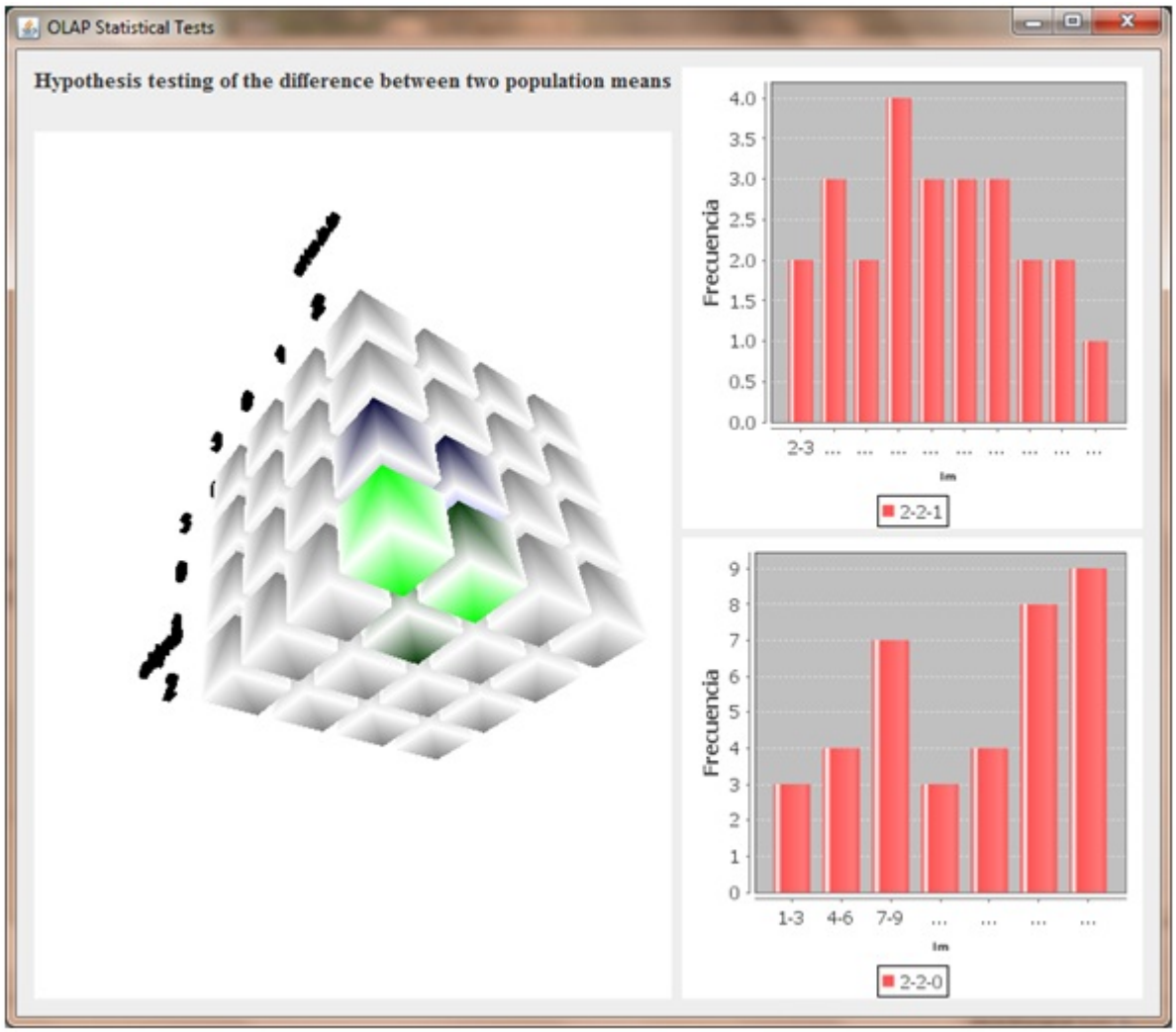

Figura 5.9: Movimiento de los cuboides hacia una cara visible

Un problema tratado en este trabajo fue el manejo de muchas dimensiones, para mostrar lo que sucedería al elegir más de 3, se muestra el siguiente ejemplo tomando 4 dimensiones.

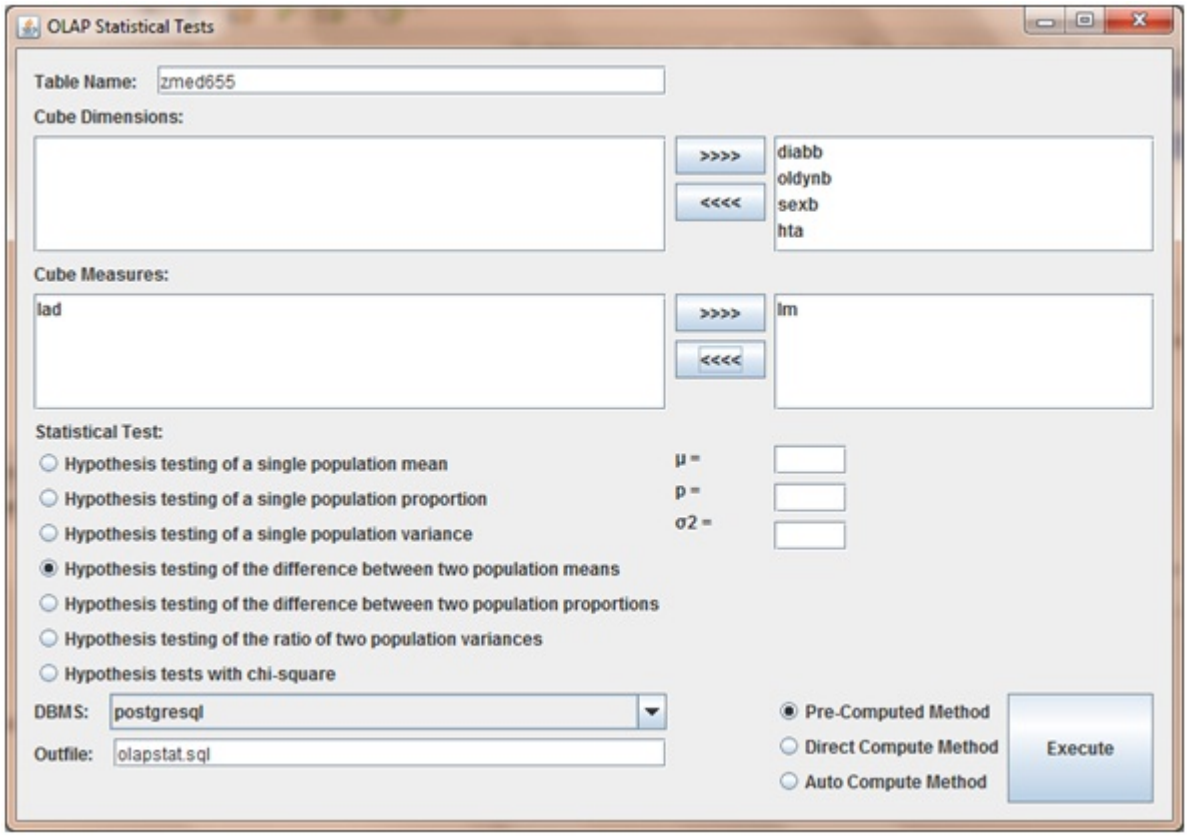

Figura 5.10: Selección de más de 3 dimensiones

Como el caso anterior se observa un cubo, aunque esta vez para una de las dimensiones sólo se encontraron dos valores diferentes. Falta una dimensión de las que fueron seleccionadas, en esta pantalla aún no es visible, sin embargo, se puede seleccionar algún cuboide de los mostrados e ir hacia "adentro" para poder apreciar las demás dimensiones que pudiera contener.

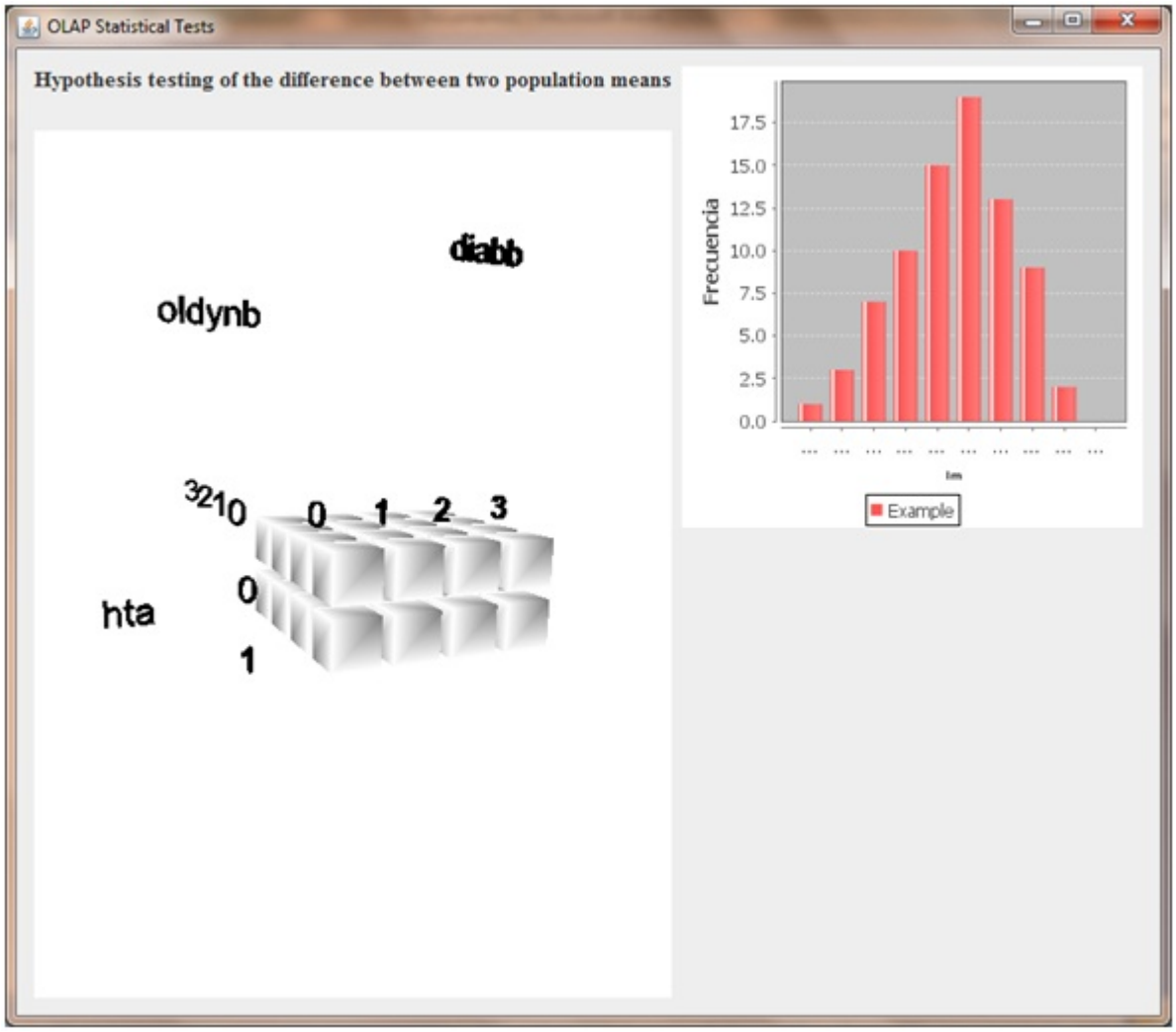

Figura 5.11: Muestra de las primeras 3 dimensiones

Al seleccionar la población donde diabb = 2, hta = 1 y oldynb = 2 se muestran las dimensiones restantes, en este caso sexb. En la parte superior, esta vez se puede apreciar además del título, la población en la cual nos adentramos.

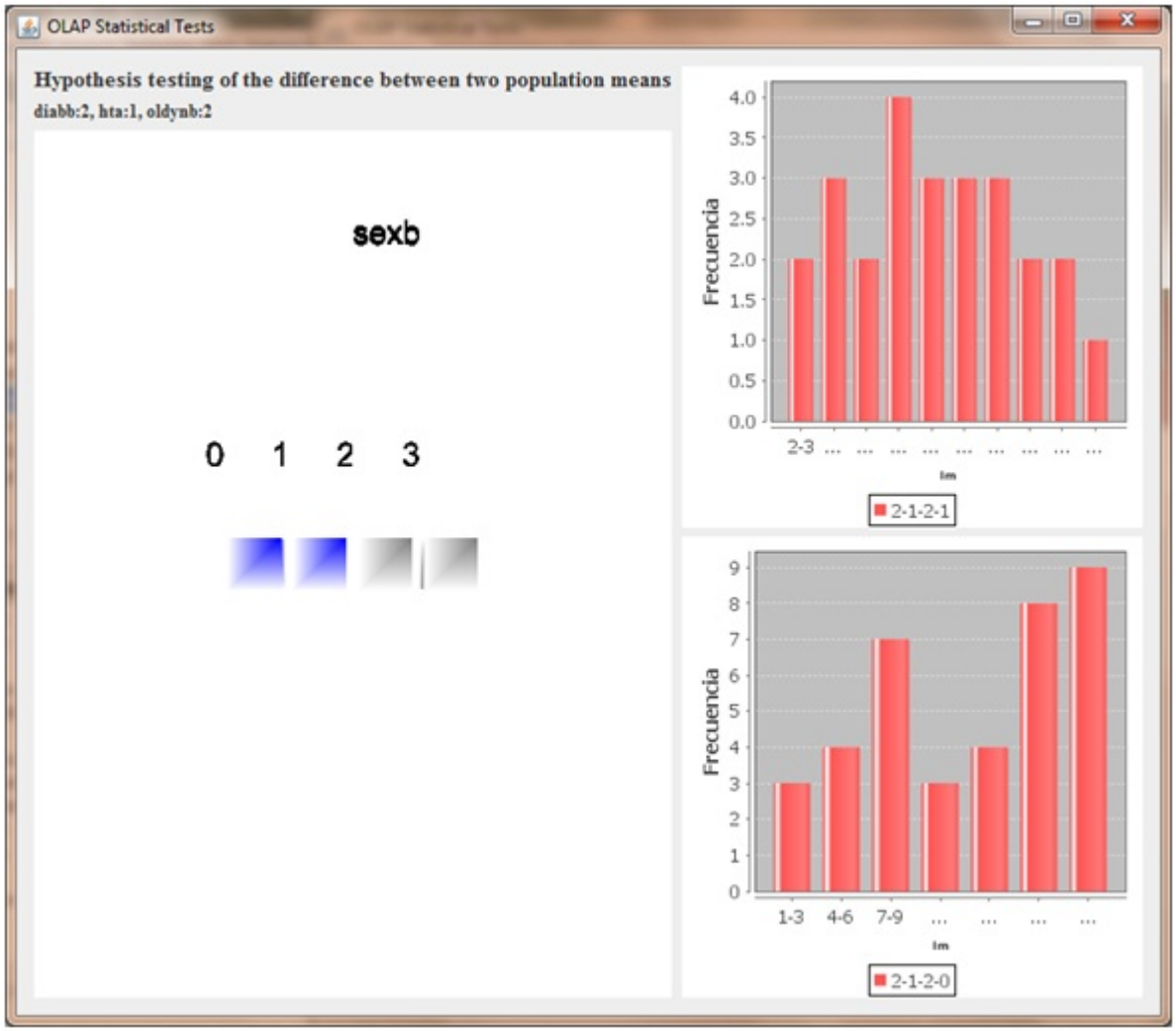

Figura 5.12: Selección de una población en específico para observar las demás dimensiones

Adicionalmente como se aprecia en la siguiente imagen, las gráficas también pueden manipularse para observar completamente los valores en los ejes.

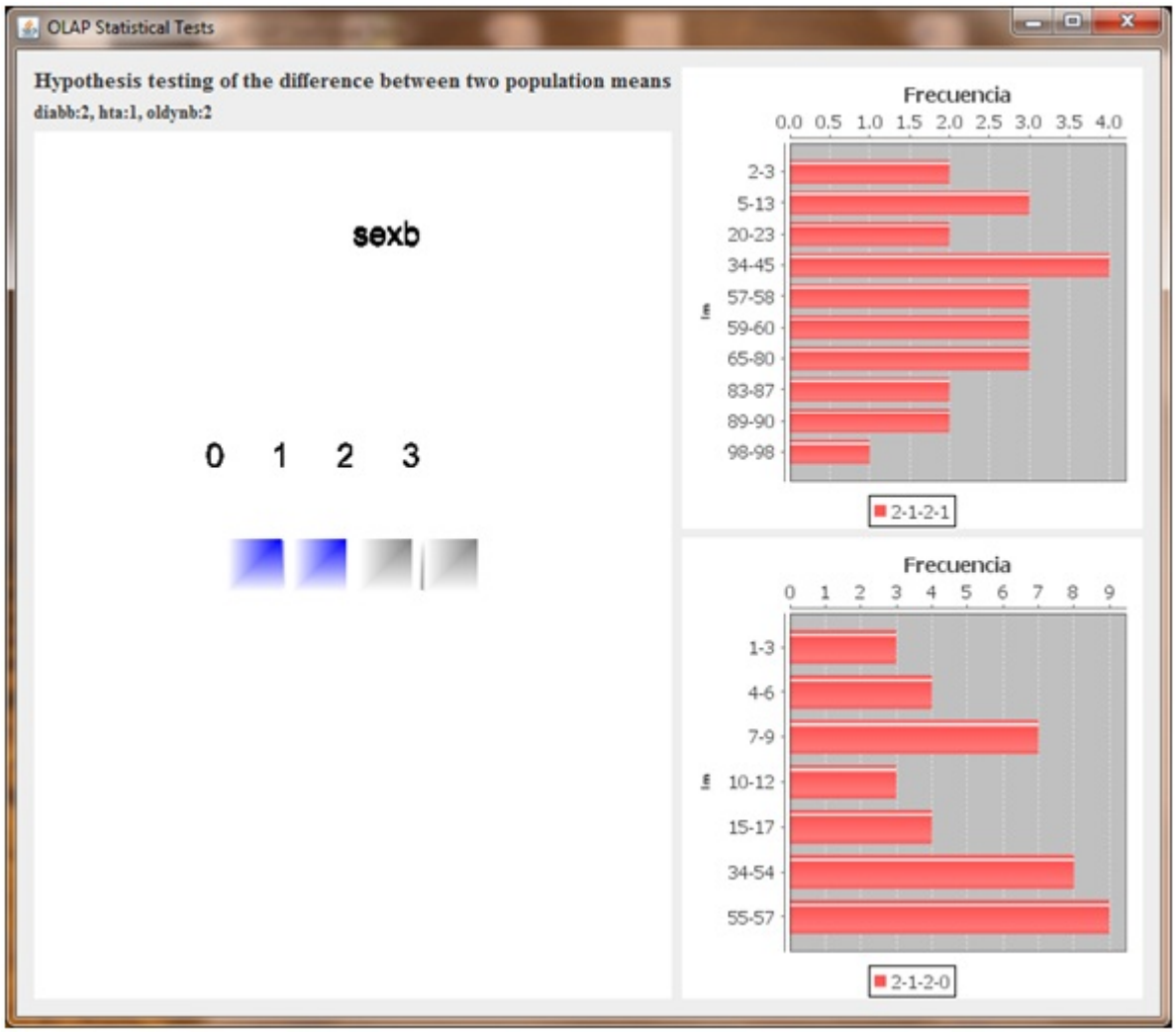

Figura 5.13: Intercambio de ejes en las gráficas

5.3.2. Experimento 2 (datos reales)

Para experimentar y comparar los resultados obtenidos a través de la aplicación, además de mostrar la visualización de estos mismos, se realizó una comparación con resultados obtenidos a través de un experimento realizado y publicado por un instituto nacional de salud pública.

Resumen

El artículo muestra los resultados obtenidos de un estudio para determinar las consecuencias clínicas de la tuberculosis pulmonar en pacientes con diabetes y el riesgo de presentar la enfermedad de quienes están expuestos a la bacteria que provoca la enfermedad.

La base de datos con la que se cuenta para la experimentación son datos recopilados a cerca de casos confirmados en México entre 1995 y 2012 así como información de sus contactos entre 2001 y 2004. En total existen 1262 participantes con tuberculosis en la base.

A los resultados a los que se llegó es que el riesgo de padecer tuberculosis en personas con diabetes es el triple que en personas sin diabetes. Los pacientes con diabetes y tuberculosis tienen manifestaciones clínicas más graves y una mayor probabilidad de fracaso del tratamiento, así como de recurrencia y recaída. Muchos estudios han explorado la relación entre diabetes y tuberculosis, de la misma manera se ha demostrado que el riesgo de presentar tuberculosis en personas con diabetes es el triple que en personas sin diabetes.

El objetivo principal del artículo es describir las manifestaciones clínicas y resultados de tratamientos en pacientes con diabetes y con tuberculosis confirmada, comparados con pacientes sin este tipo de diagnósticos. Del mismo modo se intenta determinar el riesgo de que esta enfermedad progrese rápidamente entre pacientes con diabetes que viven con pacientes con tuberculosis.

Todos los datos fueron analizados utilizando SATA 10.0 (StataCorp LP, Texas, USA)

STATA

Stata es un paquete de software estadístico. Permite, entre otras funcionalidades, la gestión de datos, el análisis estadístico, el trazado de gráficos y las simulaciones. El tipo de archivos que utiliza son .do, denominados do files.

Pone cientos de herramientas estadísticas; técnicas avanzadas como modelos de supervivencia, regresiones con datos en panel dinámico, modelos multinivel mezclados, modelos con selección de muestra, ARCH y estimación con muestras complejas; técnicas estándar como modelos lineales generalizados, regresiones con respuesta binaria, ANOVA/MANOVA, ARIMA, análisis cluster, tabulaciones básicas y resumen estadístico

Es un paquete estadístico diseñado para el análisis descriptivo de datos y la implementación de diferentes técnicas de estimación.

Resultados

La comparación entre pacientes con y sin diabetes tuvo los siguientes resultados: Los pacientes con más probabilidades de tener diabetes fueron mujeres, personas mayores y personas con un nivel socioeconómico más alto, con mayor acceso a seguridad social y más años de educación formal. Personas que utilizan drogas ilegalmente, alcohol y tabaco, con falta de vivienda e infección de VIH fueron las más frecuentes entre personas sin diabetes. No se observaron diferencias entre los pacientes con y sin diabetes, respecto a las drogas.

Algunos de los resultados obtenidos en este artículo se resumen en la siguiente tabla:

Los valores están representados por la proporción N/Total (%)

La base de datos con la que trabaja el programa como se mencionó anteriormente contiene 1262 registros de pacientes, todos ellos son pacientes con tuberculosis, y cada uno es un registro con información adicional sobre su historia clínica o sobre algunas características particulares, algunas de ellas son sobre las cuales se realizó el análisis estadístico. En la tabla se muestran las características de sexo, edad, seguridad social, alcohol, fumar, falta de vivienda, VIH y fiebre para cada una de estas características se obtienen los resultados representativos en la tabla:

- Pacientes con TB: Son los pacientes que cumplen con la característica específica, por ejemplo, ser mujer, para este caso, sabemos que 532 pacientes de los

Característica	Pacientes con TB	Pacientes con TB y con DB	Pacientes con TB y sin DB	Valor estadístico p
Mujeres	532/1262 (42.16)	180/374 (48.13)	352/888 (39.64)	0.005
Más de 6 años de educación formal	846/1260 (64.71)	271/373 (72.65)	575/887 (64.83)	0.007
Acceso a seguridad social	451/1262 (35.74)	184/374 (49.20)	267/888 (30.07)	0.001
Uso de Alcohol	546/1260 (43.33)	144/373 (38.61)	402/887 (45.32)	0.028
Fumar	291/1259 (23.11)	68/373 (18.23)	223/886 (25.17)	0.008
Falta de vivienda	46/1257 (3.66)	4/374 (1.07)	42/883 (4.76)	0.001
VIH	24/1228 (2.00)	2/362 (0.55)	22/886 (2.54)	0.022
Fiebre	905/1259 (71.88)	276/374 (73.80)	529/885 (71.07)	0.017

TB – Tuberculosis, DB – Diabetes

Figura 5.14: Tabla de Resultados

1262 son mujeres, si lo queremos saber en porcentaje, este es el 42.16 %.

- Pacientes con TB y con DB: Esta columna nos indica del total de pacientes con TB, cuántos de ellos también tienen presencia de DB. Siguiendo con el ejemplo, de los 1262 pacientes 374 tienen TB y además presentan DB, de esos 374, 180 son mujeres, en cuanto a la proporción representa el 48.13 %.

- Pacientes con TB y sin DB: Similar a la columna anterior, nos indica del total de pacientes con TB cuántos de ellos no tienen presencia de DB. Observando los datos, de 1262 pacientes 888 tienen TB pero no DB y 352 de ellos son mujeres, es decir un 39.64 %.

El programa STATA trabaja con la prueba estadística de χ^2 , lo que se pretende conocer es cuales factores o características pueden influir en que las personas con TB también desarrollen DB. Esto se puede determinar a partir del valor estadístico P que devuelve el análisis en la tabla. En el caso de la característica que se refiere al ser mujer, se muestra un valor de 0.005, esto nos dice que existe una probabilidad del 99.5 % de que el ser mujer sea un factor para que las personas con TB desarrollen DB. Los investigadores son los que se encargarán de determinar si este valor es significativo.

En realidad la base de datos contiene como característica el sexo. Existe una tabla donde un campo que dice si el registro pertenece a un hombre o a una mujer, por ejemplo, si el valor es 1 significa que es hombre o si es un 0 será mujer. La prueba estadística de χ^2 evalúa dos características en una población y lo que se determina es si esas 2 características mantienen una relación significativa. En este caso, se plantea la hipótesis de que la presencia de DB y el sexo son características que son independientes, se realiza el procedimiento necesario para obtener el valor estadístico que nos dirá si se acepta o se rechaza la hipótesis. En caso de ser rechazada la hipótesis se determina que la presencia de DB y el sexo son características dependientes. En este caso, el investigador tuvo que realizar algún procedimiento extra para poder determinar cuál de los dos valores para el sexo es el que influye, si 0 o 1, es decir, si el ser hombre o el ser mujer.

Resultados de la aplicación propuesta

Se trabajó con la misma base de datos que en el experimento del artículo, los datos fueron cargados en postgresql. En la primera parte del programa se eligieron los campos de la base de datos a utilizar, haciendo una correspondencia con la tabla anterior se tiene:

Característica	Campo en base
Mujeres	sexo
Más de 6 años de educación formal	escolrec
Acceso a seguridad social	derechoh
Uso de Alcohol	alcohol
Fumar	fuma
Falta de vivienda	expua2
Fiebre	fiebre

Figura 5.15: Correspondencia en los campos de la base de datos

La característica que representa la presencia de diabetes, en la base de datos, es el campo pdia3.

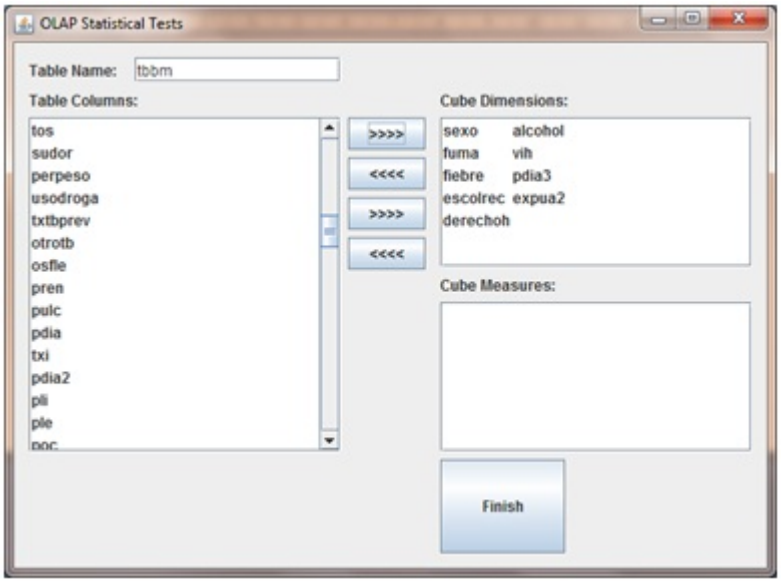

Figura 5.16: Selección de las dimensiones

En la siguiente pantalla lo principal es la selección de la prueba estadística a emplear, en este caso será la prueba para χ^2. En la sección para las dimensiones del cubo se deben elegir las características para las cuales se quiere saber si son dependientes, en este caso se seleccionará sexo y pdia3, es decir, el género y la presencia de diabetes.

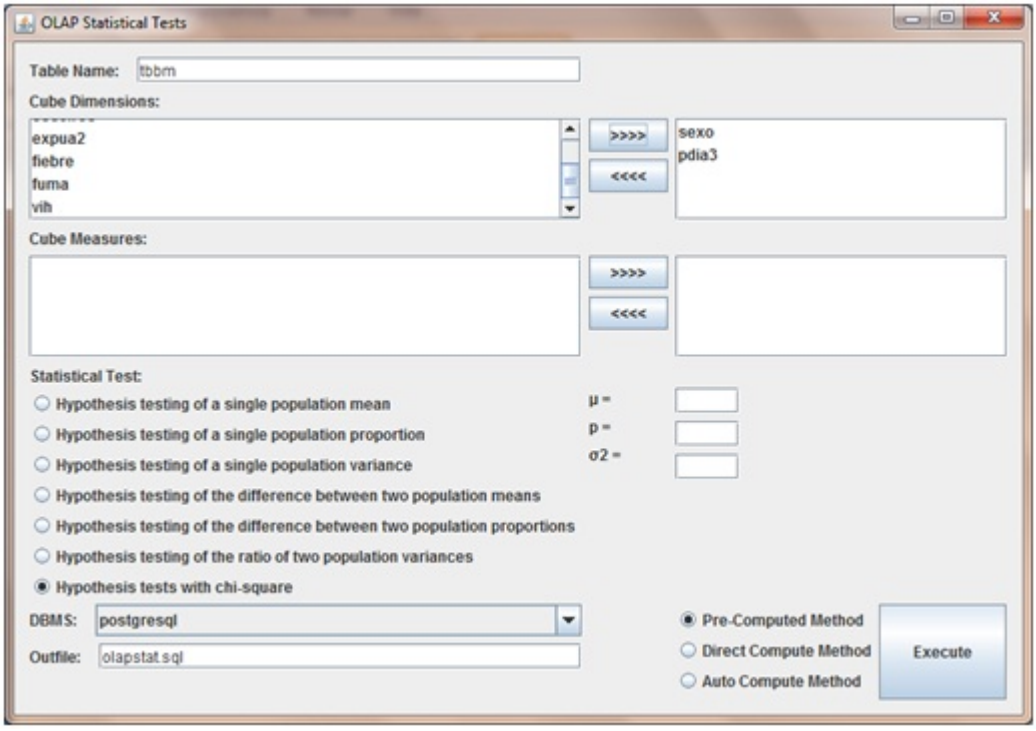

Figura 5.17: Selección de la prueba estadística y de las características a evaluar

En este caso la hipótesis será que "Las características sexo y presencia de diabetes son características dependientes".

El programa realiza la prueba estadística correspondiente, a diferencia del análisis realizado en el artículo el cual determina un valor P en cada análisis, en esta ejecución se realiza la operación para obtener el valor de la variable estadística y al final se compara con un valor fijo para determinar si se acepta o se rechaza la hipótesis, el nivel de confianza establecido es de 99 %, por lo que el valor que se toma para P es de 0.01.

Revisando el ejemplo para el sexo:

		Diabetes		TOTAL
Sexo		0 – Sin	1 – Con	
	0 – Mujer	352 (374.3391)	180 (157.6608)	532
	1 – Hombre	536 (513.6608)	194 (216.3391)	730
TOTAL		888	374	1262

Figura 5.18: Tabla de contingencia con frecuencias esperadas y frecuencias observadas

El estadístico resultante es:

$$\chi^2 = 7{,}7765$$

Observando los valores de la tabla para la distribución χ^2 se tiene:

P	0.005	0.01
	7.879	6.635

Figura 5.19: Valores extraídos de tablas para el valor P

En el artículo determina que el valor más cercano para el valor de χ^2 obtenido es 7.870 por lo que el valor para $P = 0{,}005$, es decir que la hipótesis se rechazará cuando se tenga un valor $P \leq 0{,}005$.

Para el caso de la aplicación propuesta se toma el valor de $P = 0{,}01$ por lo que en nuestro caso la hipótesis se rechazará cuando se tenga un valor $P \leq 0{,}01$, es decir, cuando $\chi^2 \leq 6{,}635$.

En este ejemplo se cumple que $\chi^2 > 6{,}635$, debido a que el valor obtenido fue $7{,}7765 > 6{,}635$ por lo que la hipótesis nula se rechaza y se acepta la alternativa, esto es, se concluye que "Las características de sexo y presencia de diabetes son dependientes".

A manera de resumen tomando las mismas características presentadas en la tabla de resultados esto es lo que debe de generar la aplicación para cada caso:

Característica	Valor estadístico p obtenido	Valor estadístico p comparado	Resultado esperado
Mujeres	0.005	0.01	Rechaza
Más de 6 años de educación formal	0.007	0.01	Rechaza
Acceso a seguridad social	0.001	0.01	Rechaza
Uso de Alcohol	0.028	0.01	Acepta
Fumar	0.008	0.01	Rechaza
Falta de vivienda	0.001	0.01	Rechaza
Fiebre	0.017	0.01	Acepta

Figura 5.20: Comparación de resultados

Para las dimensiones de sexo y presencia de diabetes la hipótesis nula se rechaza y se acepta la alternativa, visualmente se puede apreciar que los cubos se iluminan de color verde.

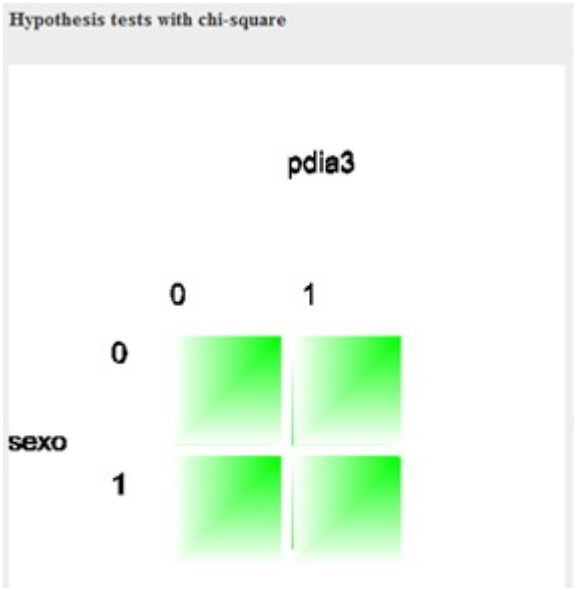

Figura 5.21: sexo vs presencia de diabetes

Para las dimensiones de alcohol y presencia de diabetes la hipótesis nula se acepta, visualmente se puede apreciar ya que los cubos se quedan en color gris.

Figura 5.22: alcohol vs presencia de diabetes

Los demás resultados se muestran a continuación:

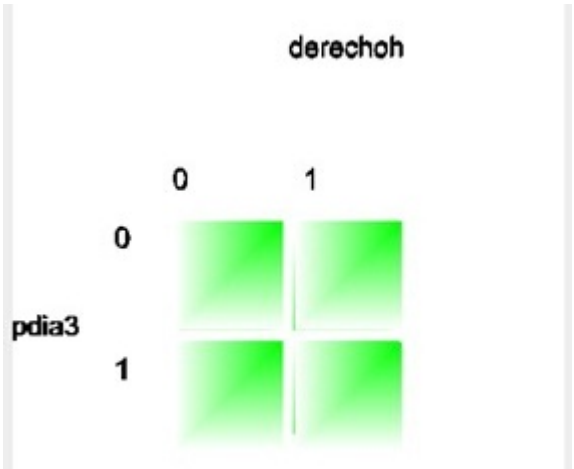

Figura 5.23: acceso a seguridad social vs presencia de diabetes

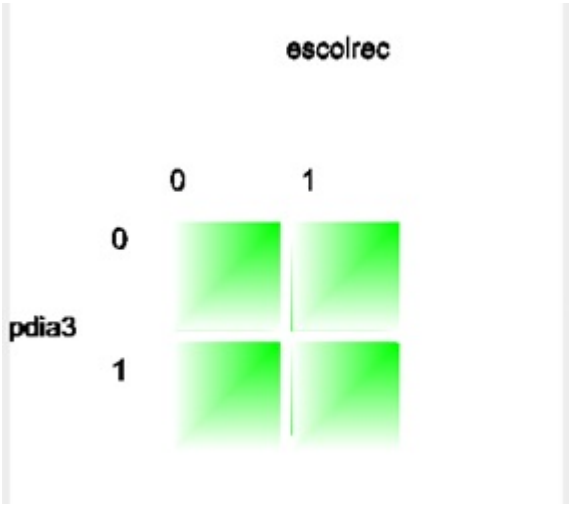

Figura 5.24: más de seis años de educación formal vs presencia de diabetes

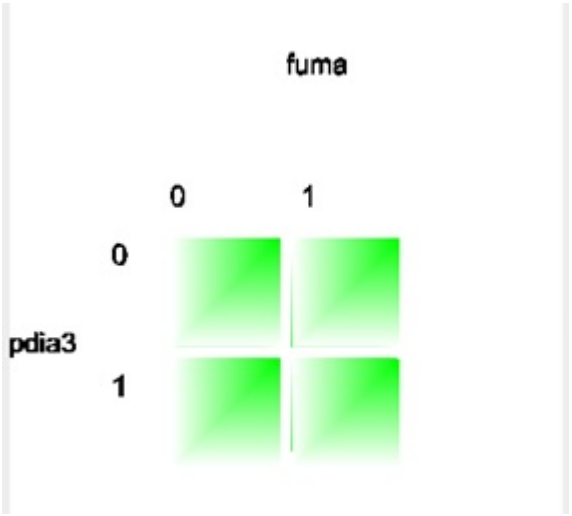

Figura 5.25: fumar vs presencia de diabetes

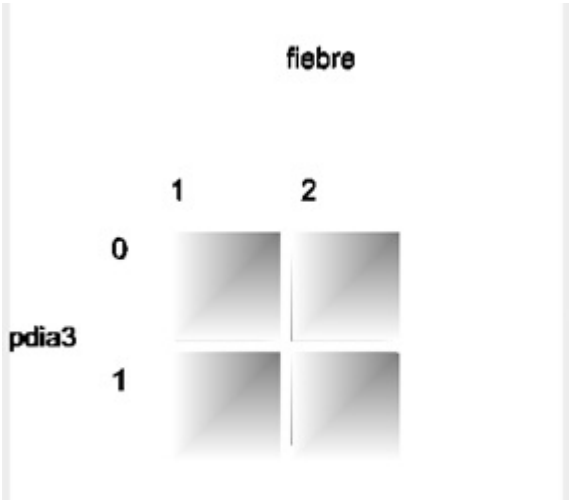

Figura 5.26: fiebre vs presencia de diabetes

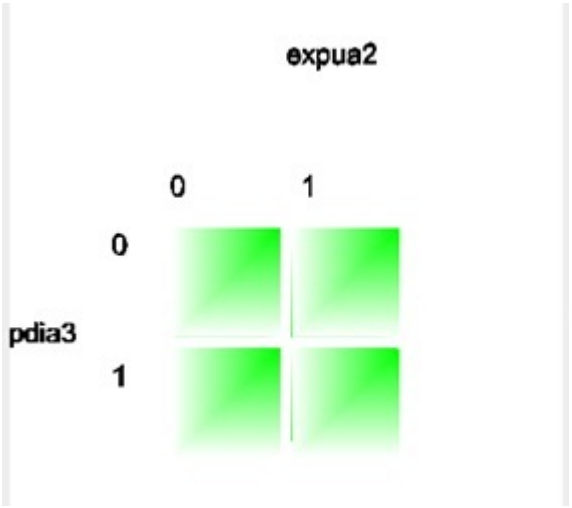

Figura 5.27: fiebre vs presencia de diabetes

Como se mencionó, la prueba χ^2 solamente dice si las dimensiones son independientes o dependientes, adicionalmente se puede emplear la prueba de la diferencia entre las proporciones de dos poblaciones para determinar cuál es en específico el valor que mantiene esa dependencia.

Siguiendo con las dimensiones de sexo y presencia de diabetes se sabe que: Existen 532 pacientes mujeres de las cuales 180 tienen diabetes, además 730 hombres, 194 de ellos con diabetes. Obteniendo las proporciones, la proporción de mujeres con diabetes $MD = 180/532 = 0{,}338$ y la proporción de hombres de hombres con diabetes $HD = 194/730 = 0{,}265$.

Se quiere saber si se puede concluir que la proporción de pacientes con presencia de diabetes es mayor entre mujeres que entre hombres. Las hipótesis quedarían de la siguiente manera:

- H_0 = "La proporción de mujeres con presencia de diabetes es menor o igual que la proporción de hombres con presencia de diabetes"
- H_a = "La proporción de mujeres con presencia de diabetes es mayor que la proporción de hombres con presencia de diabetes"

Estadística de prueba:

$$z = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}}$$

$$\sigma_{p_1-p_2} = \sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}, \qquad \bar{p} = \frac{x_1 + x_2}{n_1 + n_2}, \qquad \bar{q} = 1 - \bar{p}$$

El nivel de confianza que se tomó para esta prueba es de un 95 % por lo que el valor para $p = 0{,}05$, en tablas se puede obtener el valor para $z_p = 1{,}645$

Haciendo la comparación:

$$z \leq z_p$$

Lo que se observa es que al realizar la prueba estadística se rechaza H_0 y en consecuencia se acepta H_a.

De esta manera se puede concluir lo que se buscaba, la proporción de mujeres en la población de pacientes con presencia de diabetes es mayor que la proporción de hombres. Ligado con la conclusión anterior de que existe una relación entre el sexo y la presencia de diabetes, se puede determinar, coincidiendo con el artículo que: "El ser mujer es un factor para desarrollar diabetes en pacientes con tuberculosis".

En la aplicación se pueden revisar visualmente estos resultados:

Se seleccionan las dos dimensiones involucradas, en este caso, sexo y diabetes, además de seleccionar la prueba de hipótesis para la diferencia entre proporciones.

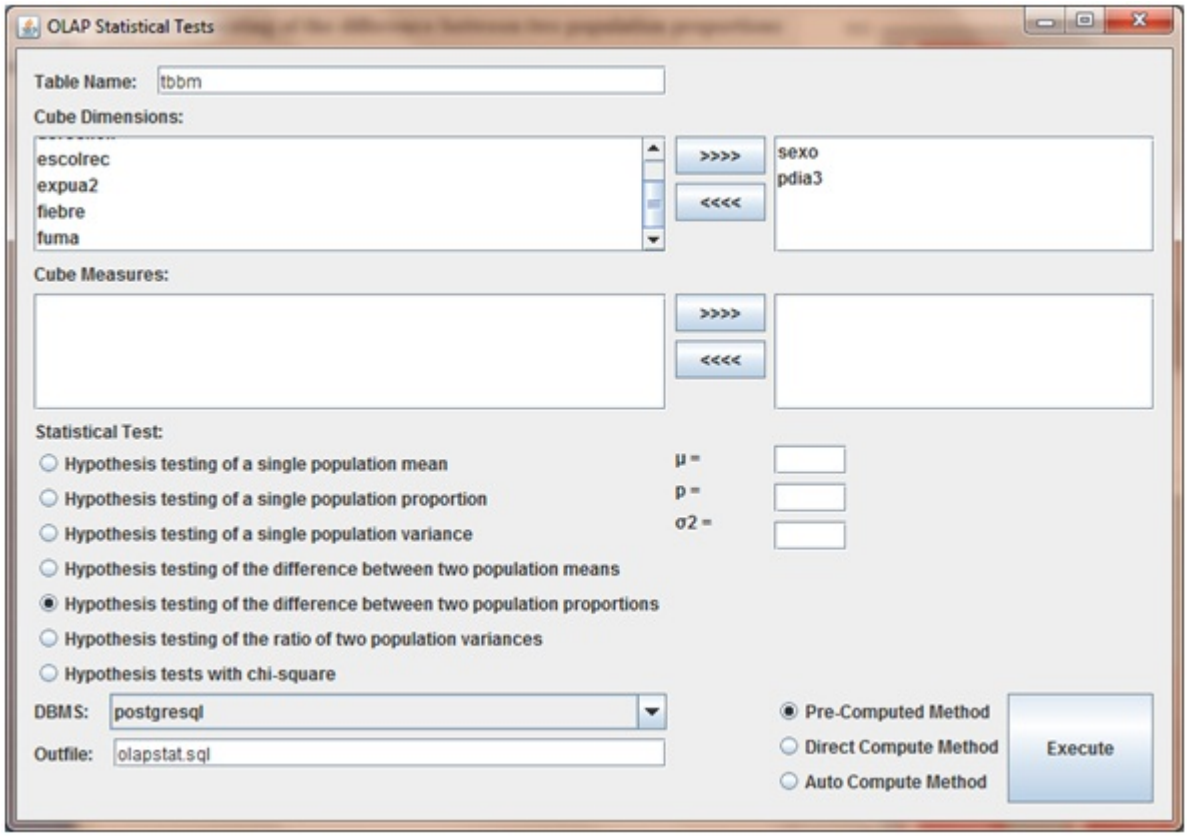

Figura 5.28: Selección de elementos para realizar la prueba

En los resultados se aprecian las poblaciones en las que la hipótesis fue rechazada, cada par se encuentra identificado con un color diferente. Las poblaciones de interés son donde existe la presencia de diabetes (pdia3 = 1), de inicio pareciera que no hubo una hipótesis confirmada entre las poblaciones donde hay presencia de diabetes y el ser mujer (sexo = 0) o ser hombre (sexo = 1), sin embargo, al girar el modelo en 3D, se puede apreciar que el cubo con coordenadas (1,0) mantiene 2 relaciones, una con (0,0) y otra con (1,1) por lo que se aprecian dos colores diferentes, el color verde claro en los dos cubos (1,0) y (1,1) muestra que la hipótesis nula entre ambos fue rechazada por lo que se acepta la alternativa y la proporción mayor que se observa entre las poblaciones es confirmada.

Las proporciones se pueden obtener a partir de las gráficas en la pantalla, la barra de "Total" indica el número total de pacientes, hombres o mujeres, según sea el caso, y la barra "N" dice el número de pacientes, hombres o mujeres, respectivamente y que además tienen presencia de diabetes.

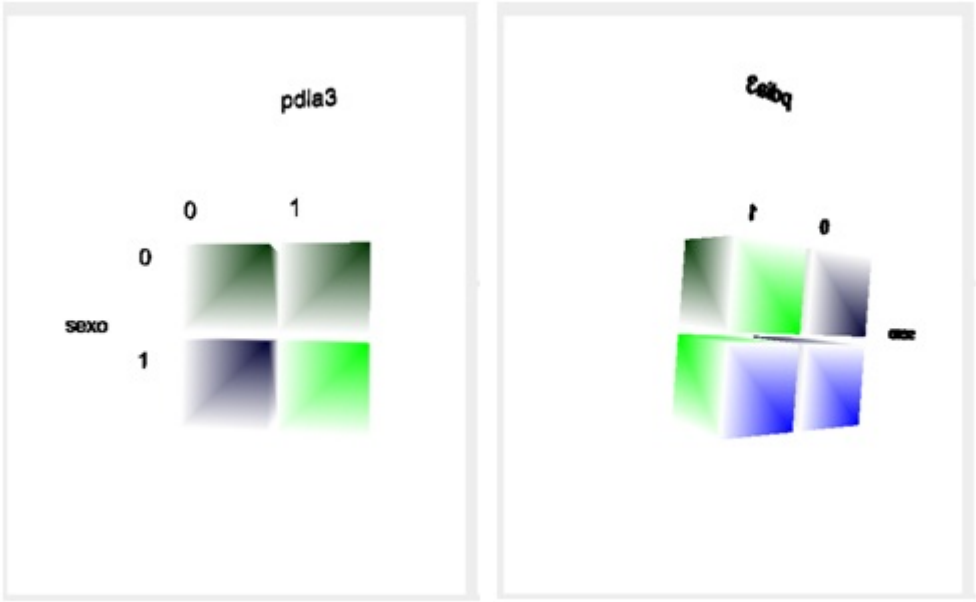

Figura 5.29: Resultados de la prueba de proporciones

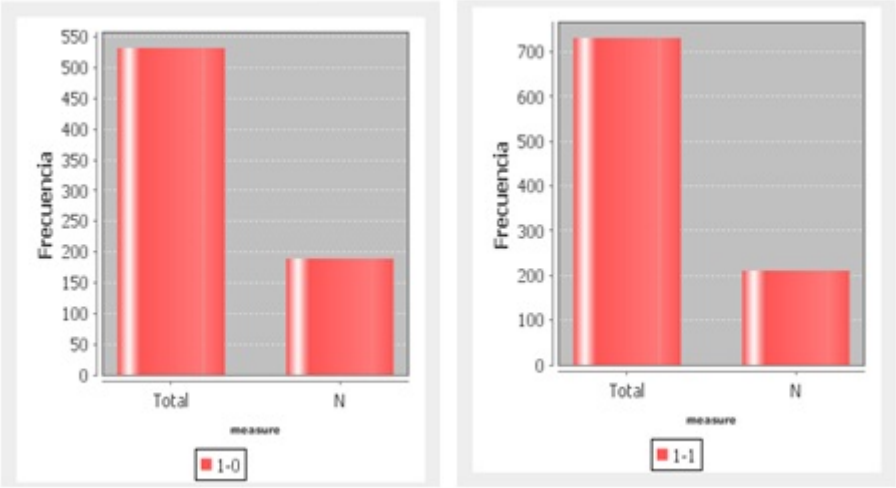

Figura 5.30: Gráficas que muestran las proporciones de las poblaciones (1-0) y (1-1)

Capítulo 6

Conclusiones y trabajo futuro

Se hizo uso de una herramienta enfocada a un problema en específico a través del análisis OLAP, para crear una nueva herramienta en este campo de análisis donde todo el proceso tiene un sustento estadístico. Se implementaron nuevas pruebas estadísticas para contar con una variedad de estas para ser empleadas en distintos campos de investigación. Para la fácil visualización y comprensión de los resultados se crearon nuevos métodos y formas de mostrar los datos, así como funcionalidades para explorarlos, en el mismo contexto se muestra una nueva forma de visualizar los datos y adentrarse en todas las dimensiones cuando se cuenta con muchas de ellas.

Dichas funcionalidades y mejoras se utilizaron y comprobaron a través de experimentos, uno de ellos con datos un tanto manipulados pero con el objetivo de mostrar todas las bondades con las que cuenta la herramienta. Para la comprobación de resultados se tomó un caso real analizado en un artículo de investigación, donde tomando la misma base de datos recopilada en dicho estudio, se empleo la herramienta para llegar a las mismas conclusiones.

En lo que se refiere a trabajos futuros, el campo de la estadística es amplio y se puede analizar la manera de utilizar aún más pruebas de hipótesis para el análisis de los datos. En este trabajo se hizo uso de la programación para la visualización en 3D, en este punto no existe algún límite para ideas que pudieran surgir encaminadas a mejorar la manipulación o representación del cubo y sus componentes. Así mismo, se sabe que en el mercado existen algunas herramientas que trabajan con visualización de datos las cuales pueden realizar cosas distintas a las presentadas en este trabajo,

cabe la posibilidad de que estas herramientas existentes pudieran ser complementadas con el análisis propuesto, por lo que sería conveniente proporcionar la posibilidad tanto de exportar como de importar datos con el sistema trabajado.

Bibliografía

[1] . Visualization and knowledge discovery for high dimensional data. En *Proceedings of the Second International Workshop on User Interfaces to Data Intensive Systems (UIDIS'01)*, UIDIS '01, págs. 1–. IEEE Computer Society, Washington, DC, USA, 2001. ISBN 0-7695-0834-0. URL http://dl.acm.org/citation.cfm?id=572716.884774.

[2] Y. Bédard, T. Merrett, y J. Han. Fundaments of spatial data warehousing for geographic knowledge discovery. *Geographic datamining and knowledge discovery*, Research Monographs in GIS(11):53–73, 2000.

[3] Surajit Chaudhuri y Umeshwar Dayal. An overview of data warehousing and olap technology. *SIGMOD Rec.*, 26(1):65–74, 1997. ISSN 0163-5808. doi:10.1145/248603.248616. URL http://doi.acm.org/10.1145/248603.248616.

[4] Yeow Wei Choong, Dominique Laurent, y Patrick Marcel. Computing appropriate representations for multidimensional data. En *Proceedings of the 4th ACM international workshop on Data warehousing and OLAP*, DOLAP '01, págs. 16–23. ACM, New York, NY, USA, 2001. ISBN 1-58113-437-1. doi:10.1145/512236.512239. URL http://doi.acm.org/10.1145/512236.512239.

[5] Howard B. Christensen y the Statistics Instructional Development Team. *Statistics step by step*. Houghton Mifflin, Boston, 2000. ISBN 58-648-411-4.

[6] E. F. Codd, S. B. Codd, y C. T. Salley. Providing OLAP (On-Line analytical processing) to User-Analysis: An IT mandate. 1993.

[7] Alfredo Cuzzocrea, Domenico Saccà, y Paolo Serafino. A hierarchy-driven compression technique for advanced olap visualization of multidimensional

data cubes. En *Proceedings of the 8th international conference on Data Warehousing and Knowledge Discovery*, DaWaK'06, págs. 106–119. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 3-540-37736-0, 978-3-540-37736-8. doi: 10.1007/11823728_11. URL http://dx.doi.org/10.1007/11823728_11.

[8] Stephen G. Eick. Visualizing multi-dimensional data. *SIGGRAPH Comput. Graph.*, 34(1):61–67, 2000. ISSN 0097-8930. doi:10.1145/604446.604454. URL http://doi.acm.org/10.1145/604446.604454.

[9] Michael Gebhardt, Matthias Jarke, y Stephan Jacobs. A toolkit for negotiation support interfaces to multi-dimensional data. *SIGMOD Rec.*, 26(2):348–356, 1997. ISSN 0163-5808. doi:10.1145/253262.253344. URL http://doi.acm.org/10.1145/253262.253344.

[10] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, y Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Min. Knowl. Discov.*, 1(1):29–53, 1997. ISSN 1384-5810. doi:10.1023/A:1009726021843. URL http://dx.doi.org/10.1023/A:1009726021843.

[11] Jiawei Han, Micheline Kamber, y Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd ed$^{\underline{\text{ón}}}$., 2011. ISBN 0123814790, 9780123814791.

[12] P. Hanrahan, C. Stolte, y J. Mackinlay. Visual analysis for everyone: Understanding data exploration and visualization. 2007.

[13] Jaqueline Hurtado de Becerra. *Metodología de la investigación holística*. Caracas, Venezuela, 2000.

[14] W. H. Inmon. *Building the Data Warehouse,3rd Edition*. John Wiley & Sons, Inc., New York, NY, USA, 3rd ed$^{\underline{\text{ón}}}$., 2002. ISBN 0471081302.

[15] W. H. Inmon y Chuck Kelley. *Rdb/VMS, developing the data warehouse*. Q E D Pub Co, 1994.

[16] Su Mi Ji, Beom Seok Lee, Kyoung Il Kang, Sung Gook Kim, Cheolwhan Lee, Oh-young Song, Joon Yeon Choeh, Ran Baik, y Sung Wook Baik. A study on the generation of olap data cube based on 3d visualization interaction. En *Proceedings of the 2011 International Conference on Computational Science and Its Applications*, ICCSA '11, págs. 231–234. IEEE Computer Society, Washington, DC, USA, 2011. ISBN 978-0-7695-4404-5. doi:10.1109/ICCSA.2011.40. URL http://dx.doi.org/10.1109/ICCSA.2011.40.

[17] J. R. Jiménez. *Nociones básicas de probabilidad*. Universidad de los andes, Venezuela, 2006.

[18] Daniel A. Keim y Hans-Peter Krigel. Visdb: Database exploration using multidimensional visualization. *IEEE Comput. Graph. Appl.*, 14(5):40–49, 1994. ISSN 0272-1716. doi:10.1109/38.310723. URL http://dx.doi.org/10.1109/38.310723.

[19] Gregory E. Kersten, Zbigniew Mikolajuk, y Anthony Gar on Yeh. *Decision support systems for sustainable development : a resource book of methods and applications*. Boston : Kluwer Academic, 2000. ISBN 0306475421.

[20] Hing-Yan Lee y Hwee-Leng Ong. A new visualisation technique for knowledge discovery in olap.

[21] Andreas Maniatis, Panos Vassiliadis, Spiros Skiadopoulos, y Yannis Vassiliou. Cpm: A cube presentation model for olap. En *DaWaK 2003, Prague, Czech Republic, September 3 5 (2003*. 2003.

[22] Andreas S. Maniatis, Panos Vassiliadis, Spiros Skiadopoulos, y Yannis Vassiliou. Advanced visualization for olap. En *Proceedings of the 6th ACM international workshop on Data warehousing and OLAP*, DOLAP '03, págs. 9–16. ACM, New York, NY, USA, 2003. ISBN 1-58113-727-3. doi:10.1145/956060.956063. URL http://doi.acm.org/10.1145/956060.956063.

[23] Svetlana Mansmann y Marc H. Scholl. Exploring olap aggregates with hierarchical visualization techniques. En *Proceedings of the 2007 ACM symposium on Applied computing*, SAC '07, págs. 1067–1073. ACM, New York,

NY, USA, 2007. ISBN 1-59593-480-4. doi:10.1145/1244002.1244235. URL http://doi.acm.org/10.1145/1244002.1244235.

[24] Ciro Martínez Bencardino. *Estadística y muestreo*. Ecoe Ediciones, Bogotá, 2005. ISBN 58-648-411-4.

[25] Victor Morales Sánchez. *Planeamiento y analisis de investigaciones*. Eldorado, Caracas, 2000.

[26] Carlos Ordoñez y Zhibo Chen. Evaluating statistical tests on olap cubes to compare degree of disease. *Trans. Info. Tech. Biomed.*, 13(5):756–765, 2009. ISSN 1089-7771. doi:10.1109/TITB.2008.926989. URL http://dx.doi.org/10.1109/TITB.2008.926989.

[27] Carlos Ordoñez y Zhibo Chen. Exploration and visualization of olap cubes with statistical tests. En *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, VAKD '09, págs. 46–55. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-670-0. doi:10.1145/1562849.1562855. URL http://doi.acm.org/10.1145/1562849.1562855.

[28] Carlos Ordoñez, Zhibo Chen, y Javier García-García. Interactive exploration and visualization of olap cubes. En *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, DOLAP '11, págs. 83–88. ACM, New York, NY, USA, 2011. ISBN 978-1-4503-0963-9. doi:10.1145/2064676.2064691. URL http://doi.acm.org/10.1145/2064676.2064691.

[29] Ayman Ammoura Osmar, Osmar Zaane, y Y Goebel. Towards a novel olap interface for distributed data warehouses. En *In DaWaK*, págs. 174–185. 2001.

[30] Ayala alejandro Peña. Tecnologías de la información: Su alineamiento al negocio de las organizaciones. 2006.

[31] Peter Rob y Carlos Coronel. *Database Systems, Design, Implementation and Management*. Course Technology, 5th ed$^{\underline{\text{ón}}}$., 2002.

[32] Sergio Ezequiel Rozic. Bases de datos y su aplicación con sql. 2004.

[33] P. Russom. Trends in data visualization software for business users. 2000.

[34] Sunita Sarawagi, Rakesh Agrawal, y Nimrod Megiddo. Discovery-driven exploration of olap data cubes. En *Proceedings of the 6th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '98, págs. 168–182. Springer-Verlag, London, UK, UK, 1998. ISBN 3-540-64264-1. URL http://dl.acm.org/citation.cfm?id=645338.650401.

[35] Shashi Shekhar, Chang-Tien Lu, Pusheng Zhang, y Rulin Liu. Data mining for selective visualization of large spatial datasets. En *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '02, págs. 41–. IEEE Computer Society, Washington, DC, USA, 2002. ISBN 0-7695-1849-4. URL http://dl.acm.org/citation.cfm?id=850952.853772.

[36] Mark Sifer. A visual interface technique for exploring olap data with coordinated dimension hierarchies. En *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, págs. 532–535. ACM, New York, NY, USA, 2003. ISBN 1-58113-723-0. doi:10.1145/956863.956966. URL http://doi.acm.org/10.1145/956863.956966.

[37] Chris Stolte, Diane Tang, y Pat Hanrahan. Multiscale visualization using data cubes. *IEEE Transactions on Visualization and Computer Graphics*, 9(2):176–187, 2003. ISSN 1077-2626. doi:10.1109/TVCG.2003.1196005. URL http://dx.doi.org/10.1109/TVCG.2003.1196005.

[38] Kesaraporn Techapichetvanich y Amitava Datta. Interactive visualization for olap. En *Proceedings of the 2005 international conference on Computational Science and Its Applications - Volume Part III*, ICCSA'05, págs. 206–214. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 3-540-25862-0, 978-3-540-25862-9. doi:10.1007/11424857_23. URL http://dx.doi.org/10.1007/11424857_23.

[39] David P. Tegarden. Business information visualization. *Commun. AIS*, 1(1es), 1999. URL http://dl.acm.org/citation.cfm?id=374109.374113.

[40] Angi Voss, Vera Hernandez, Hans Voss, y Simon Scheider. Interactive visual exploration of multidimensional data: Requirements for commongis with olap. En *Proceedings of the Database and Expert Systems Applications, 15th International Workshop*, DEXA '04, págs. 883–887. IEEE Computer Society, Washington, DC, USA, 2004. ISBN 0-7695-2195-9. doi:10.1109/DEXA.2004.90. URL http://dx.doi.org/10.1109/DEXA.2004.90.

[41] Wang y Yi. The application of data cubes in business data visualization. 2012.

Apéndice A

Comprobación de distribución

A.1. Resúmen

La prueba de bondad de ajuste chi-cuadrada es de ayuda para verificar que la población de la cual proviene una muestra tiene una distribución específica(en este caso se quiere comprobar una distribución normal).

A.2. Objetivo

Sea x:

x: Variable aleatoria poblacional.
$f(x)$: Ditribución de probabilidad.

Se desea probar la hipótesis:

H_0: $f(x) = f_0(x)$.

Contrastando con la hipótesis alterna:

H_a: $f(x) \neq f_0(x)$.

A.3. Desarrollo

Se tiene una muestra aleatoria de tamaño n tomada de una población con una distribución normal supuesta $f_0(x)$ la cual es de nuestro interés verificar.

Los datos fueron tomados de nuestra base de datos trabajada y son un total de 256 datos, se tomaron los valores para los campos lm, lad, rcx y rca lo cuales representan medidas.

Se hace la suposición de que las observaciones de la muestra están agrupadas en k clases, siendo o_i la cantidad de observaciones en cada clase $i = 1,2, \ldots, k$.

Las frecuencias teóricas o frecuencias esperadas e_i se obtienen a partir de la distribución normal. Una distribución puede dividirse en intervalos o clases de muchas maneras, lo más cómodo es dividir en intervalos que tengan un número idéntico de elementos para facilitar las operaciones. De esta manera los datos se dividieron en 10 clases: -1.29, -0.85, -0.53, -0.26, 0, 2.6, 0.53, 0.85 y 1.29.

En cada intervalo se tiene el 10 % de las frecuencias esperadas, en este caso al tener $n = 256$, en cada intervalo se tendrían 25.6 elementos en la frecuencia esperada.

Las clases obtenidas fueron las mostradas en la Figura A.1.

Clases		lm		lad		lcx		rca	
1	Menor a	-12.6663493		-0.57872837		-15.8229463		-14.9347432	
2	Entre	-12.6663493	-7.13758964	-0.57872837	14.4665356	-15.8229463	0.49808288	-14.9347432	1.81480097
3	Entre	-7.13758964	-3.11667354	14.4665356	25.4085457	0.49808288	12.3679223	1.81480097	13.9962877
4	Entre	-3.11667354	0.27597442	25.4085457	34.6408668	12.3679223	22.3830993	13.9962877	24.2744171
5	Entre	0.27597442	3.54296875	34.6408668	43.53125	22.3830993	32.0273438	24.2744171	34.171875
6	Entre	3.54296875	6.80996308	43.53125	52.4216332	32.0273438	41.6715883	34.171875	44.0693329
7	Entre	6.80996308	10.202611	52.4216332	61.6539543	41.6715883	51.6867652	44.0693329	54.3474623
8	Entre	10.202611	14.2235271	61.6539543	72.5959644	51.6867652	63.5566046	54.3474623	66.528949
9	Entre	14.2235271	19.7522868	72.5959644	87.6412284	63.5566046	79.8776338	66.528949	83.2784932
10	Mayor a	19.7522868		87.6412284		79.8776338		83.2784932	

Figura A.1: Clases

Las frecuencias observadas o_i para cada clase se presentan en la Figura A.2.

El cálculo de chi-cuadrada se hace utilizando la fórmula

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Clases	lm	lad	lcx	rca
1	0	0	0	0
2	0	232	231	231
3	0	9	1	1
4	231	5	8	8
5	0	5	6	6
6	0	0	4	5
7	1	2	1	0
8	0	0	2	2
9	0	2	1	2
10	24	1	2	1

Figura A.2: Frecuencias observadas

Los cálculos antes de la sumatoria se muestran en la Figura A.3.

lm	lad	lcx	rca	$f(x) = (x - 25.6)^2 / 25.6$			
0	0	0	0	25.6	25.6	25.6	25.6
0	232	231	231	25.6	1664.1	1648.01406	1648.01406
0	9	1	1	25.6	10.7640625	23.6390625	23.6390625
231	5	8	8	1648.01406	16.5765625	12.1	12.1
0	5	6	6	25.6	16.5765625	15.00625	15.00625
0	0	4	5	25.6	25.6	18.225	16.5765625
1	2	1	0	23.6390625	21.75625	23.6390625	25.6
0	0	2	2	25.6	25.6	21.75625	21.75625
0	2	1	2	25.6	21.75625	23.6390625	21.75625
24	1	2	1	0.1	23.6390625	21.75625	23.6390625

Figura A.3: Cálculo

Al obtener las sumatorias para cada registro se tiene:

Para lm $\chi^2 = 1850.953125$

Para lad $\chi^2 = 1851.96875$

Para lcx $\chi^2 = 1833.375$

Para rca $\chi^2 = 1833.6875$

A.4. Conclusión

Se observa que los valores obtenidos para χ^2 en los cuatro casos son demasiado grandes, por esta razón se concluye que los datos no siguen una distribución normal.

Apéndice B

Diagramas del sistema

B.1. Diagrama general de casos de uso

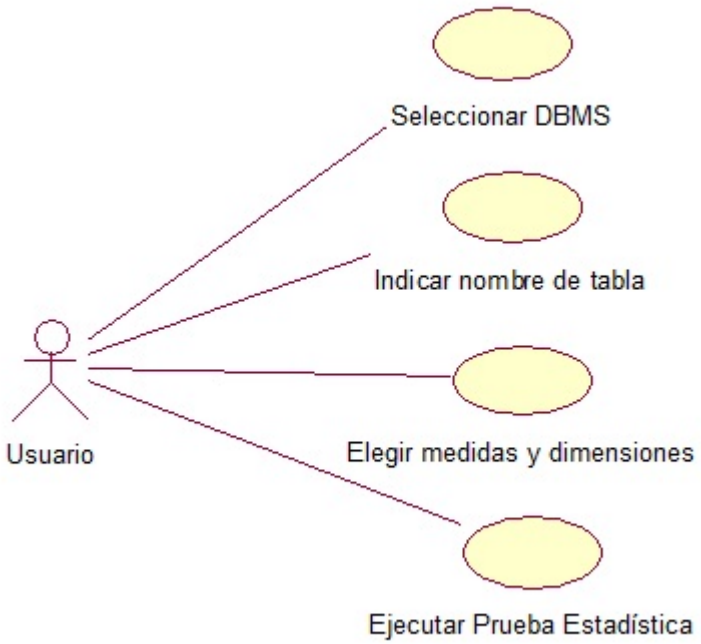

Figura B.1:

B.1.1. Caso de uso: Seleccionar DBMS

Actor: Usuario.

Descripción: Es iniciado por el usuario, sirve para elegir el manejador de bases de datos que se utilizará para la prueba estadística.

Precondiciones: El actor cuenta con un equipo de cómputo así como con la aplicación y los manejadores de bases de datos instalados.

Flujo:

ACTOR		SISTEMA		
Paso	Acción	Paso	Acción	Excp.
1	Da doble click sobre el icono de la aplicación.	2	Muestra el mensaje pidiendo que se seleccione un DBMS de dos posibles.	
3	Teclea la opción deseada.			
4	Indica Aceptar cuando haya terminado.	5	Guarda en memoria a través de una variable el DBMS seleccionado.	E1

Excepciones:

id	NOMBRE	ACCIÓN
E1	Opción que no existe en la lista.	El programa termina.

Postcondiciones: Se muestra la pantalla para la selección de la tabla.

B.1.2. Caso de uso: Indicar nombre de tabla

Actor: Usuario.

Descripción: El actor visualiza la pantalla para teclear el nombre de la tabla que se utilizará para la prueba.

Precondiciones: El actor seleccionó anteriormente el DBMS que se utilizará.

Flujo:

ACTOR		SISTEMA		
Paso	Acción	Paso	Acción	Excp.
1	Teclea el nombre de la tabla existente en la base de datos.			
2	Indica Aceptar cuando haya ingresado el nombre.	3	Guarda en memoria a través de una variable el nombre de la tabla.	
		4	Consulta en la base de datos los campos de la tabla indicada.	E1

Excepciones:

id	NOMBRE	ACCIÓN
E1	La tabla indicada no existe en la base de datos.	La pantalla siguiente no contiene datos para mostrar.

Postcondiciones: Se muestra la pantalla para seleccionar medidas y dimensiones.

B.1.3. Caso de uso: Elegir medidas y dimensiones.

Actor: Usuario.

Descripción: El actor visualiza la pantalla con todos los campos de la tabla para poder seleccionar y clasificar medidas y dimensiones que se utilizarán en la prueba.

Precondiciones: El sistema obtuvo los campos existentes en la tabla indicada.

Flujo:

ACTOR		SISTEMA		
Paso	Acción	Paso	Acción	Excp.
1	Selecciona de los campos las dimensiones para el cubo.			
2	Selecciona de los campos las medidas para el cubo.			
3	Indica terminar cuando ha seleccionado los campos necesarios.	4	Guarda en memoria las dimensiones y medidas seleccionadas.	E1

Excepciones:

id	NOMBRE	ACCIÓN
E1	No se seleccionaron ni medidas ni dimensiones.	La pantalla siguiente no contiene datos para mostrar.

Postcondiciones: Se muestra la pantalla donde hay que seleccionar las opciones para la prueba estiadística.

B.1.4. Caso de uso: Ejecutar prueba estadística.

Actor: Usuario.

Descripción: El actor visualiza la pantalla con los parámetros disponibles, selecciona los necesarios así como la prueba estadística de su interés, y finalmente la ejecuta.

Precondiciones: Se clasificaron los campos encontrados en medidas y dimensiones.

Flujo:

ACTOR		SISTEMA		
Paso	Acción	Paso	Acción	Excp.
1	Elige las dimensiones para conformar en cubo.			
2	Elige las medidas con las que se realizarán los cálculos, si son necesarias.			
3	Selecciona una prueba estadística de la lista.			
4	Indica Ejecutar, para comenzar a realizar la prueba.	5	De acuerdo a los parámetros realiza el procedimiento correspondiente a la prueba seleccionada.	E1, E2, E3

Excepciones:

id	NOMBRE	ACCIÓN
E1	El actor seleccionó medidas en alguna prueba donde no se requieren.	
E2	El actor no seleccionó ninguna dimensión.	
E3	El actor no ingresó un parámetro si la prueba estadística así lo requería.	

Postcondiciones: Se presenta la pantalla con el cubo generado y los resultados visuales de la prueba.

B.2. Diagrama de paquetes

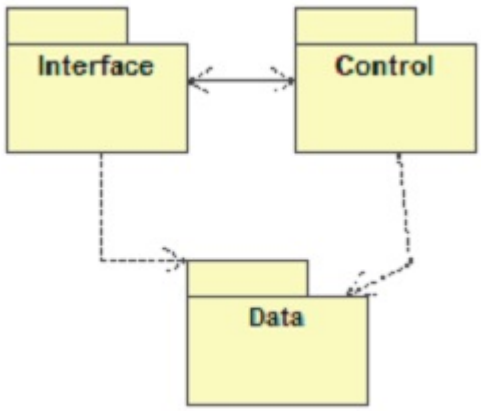

Figura B.2: Diagrama de paquetes

B.2.1. Clases de dominio del problema

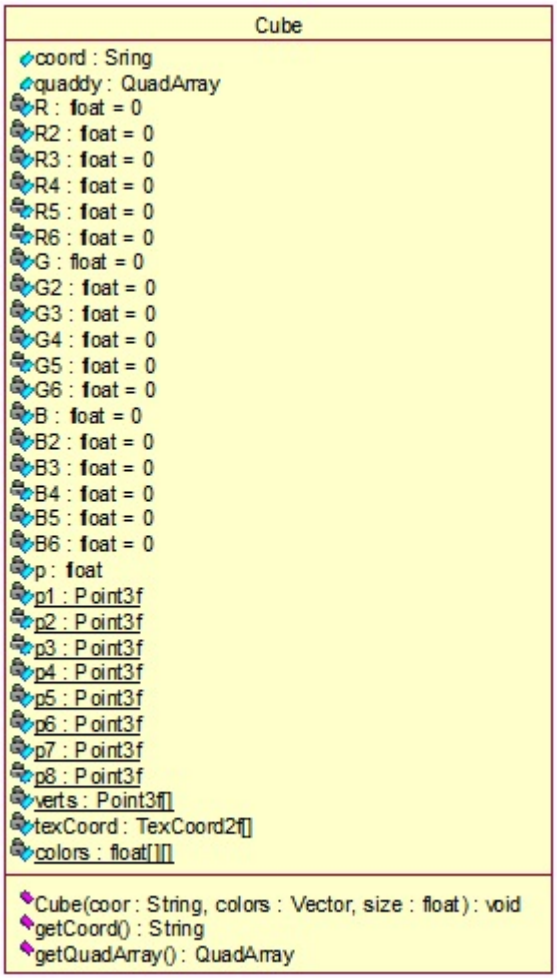

Figura B.3: Clase para la construcción de cuboides

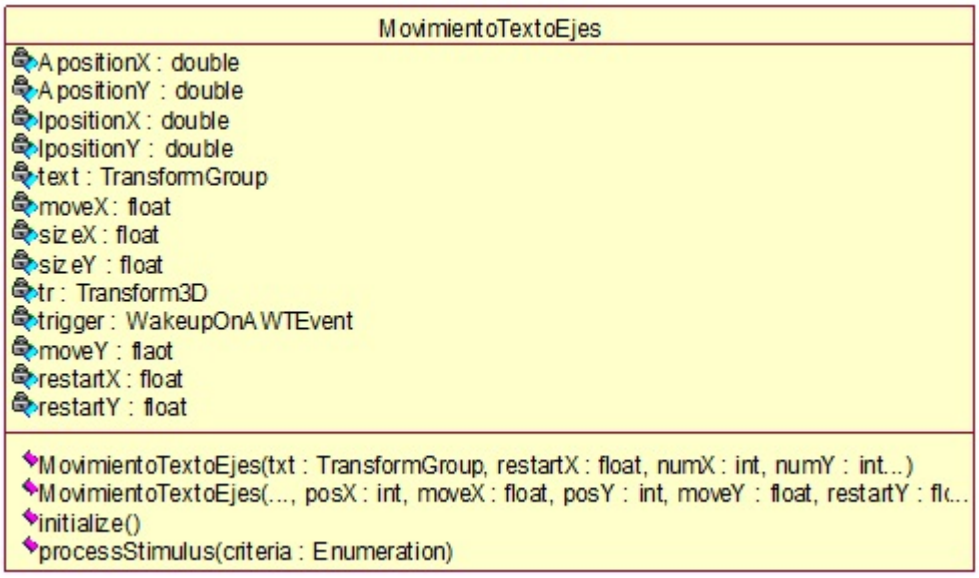

Figura B.4: Clase que maneja el movimiento del texto en 3D

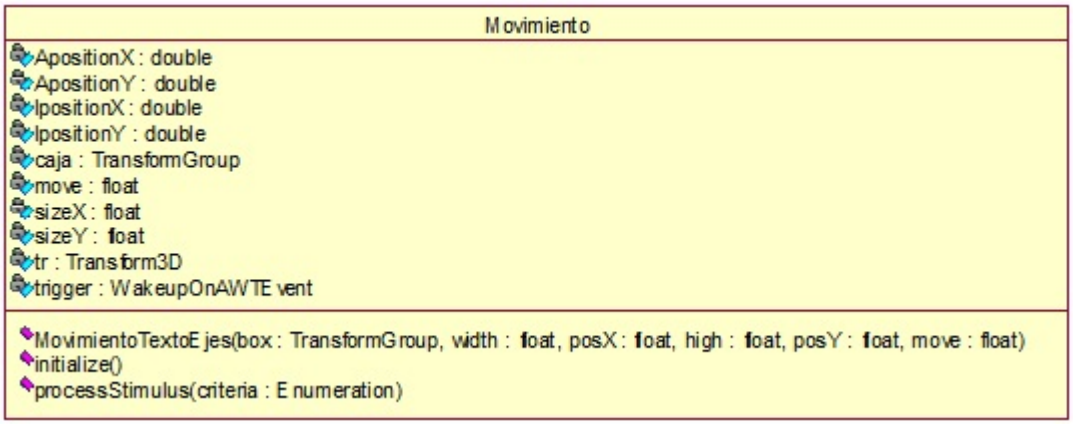

Figura B.5: Clase que maneja el movimiento de los cuboides

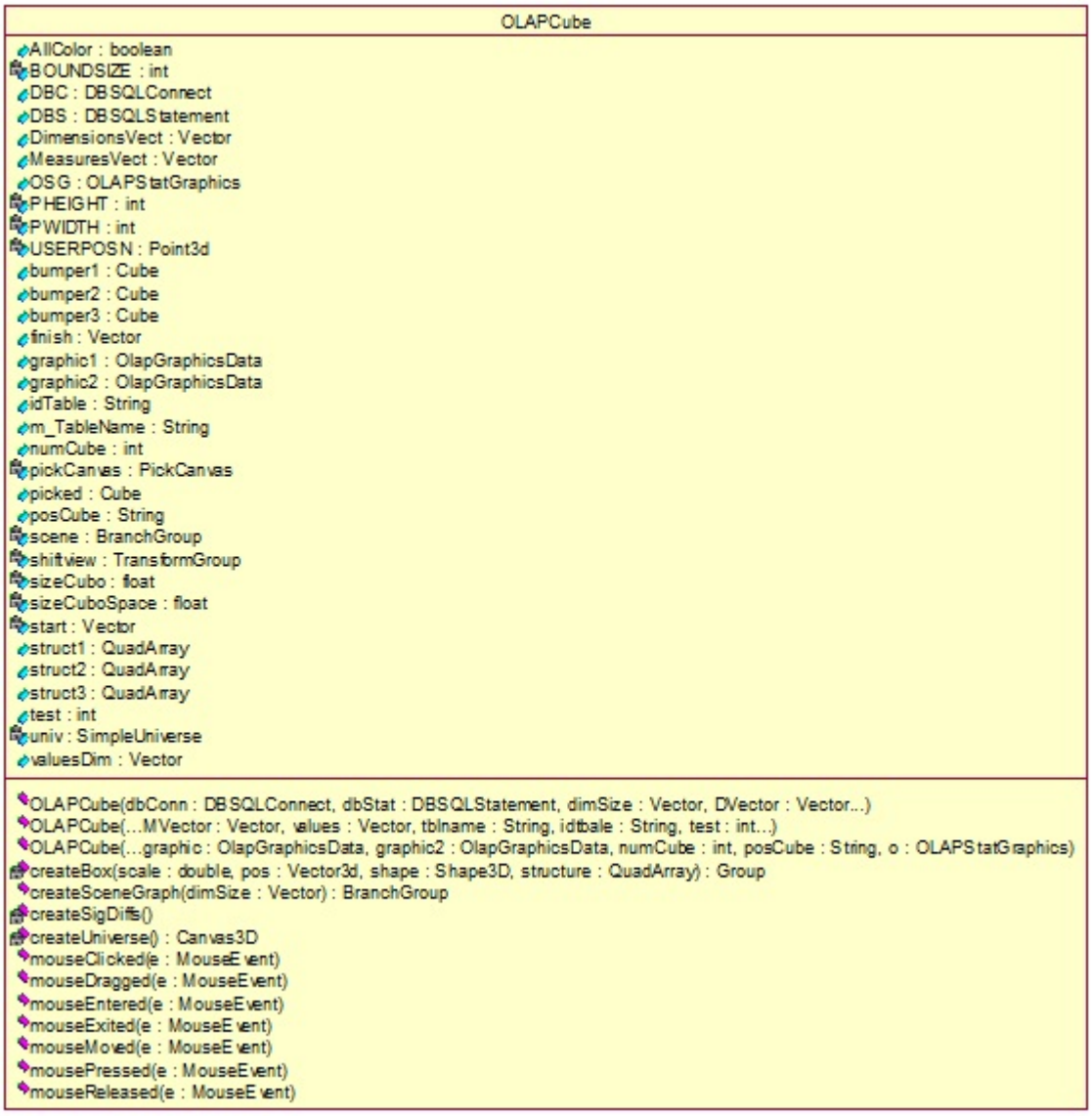

Figura B.6: Clase que maneja la escena para el cubo OLAP

Figura B.7: Clase que maneja la bitácora

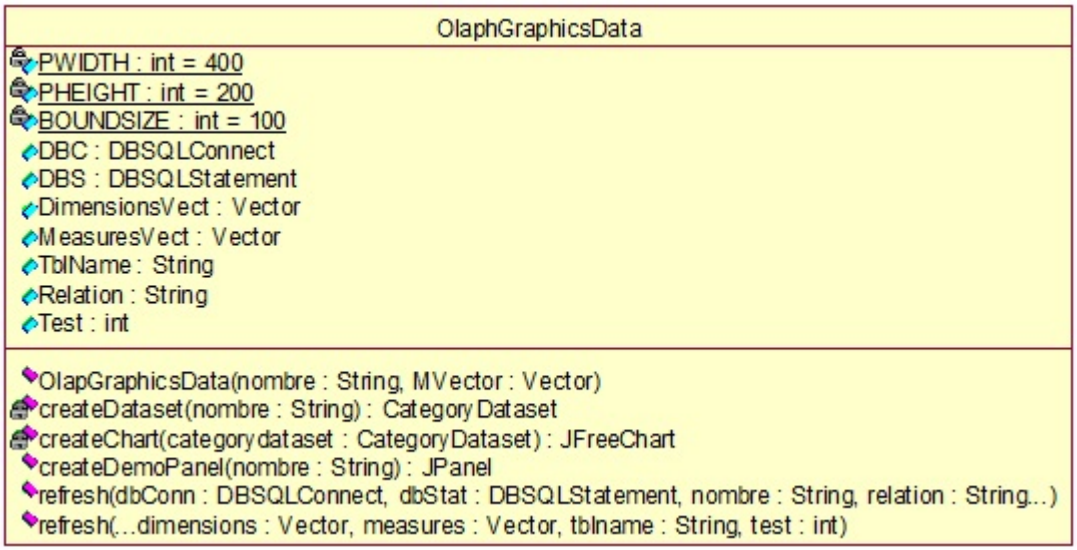

Figura B.8: Clase para manejar la creación de gráficas

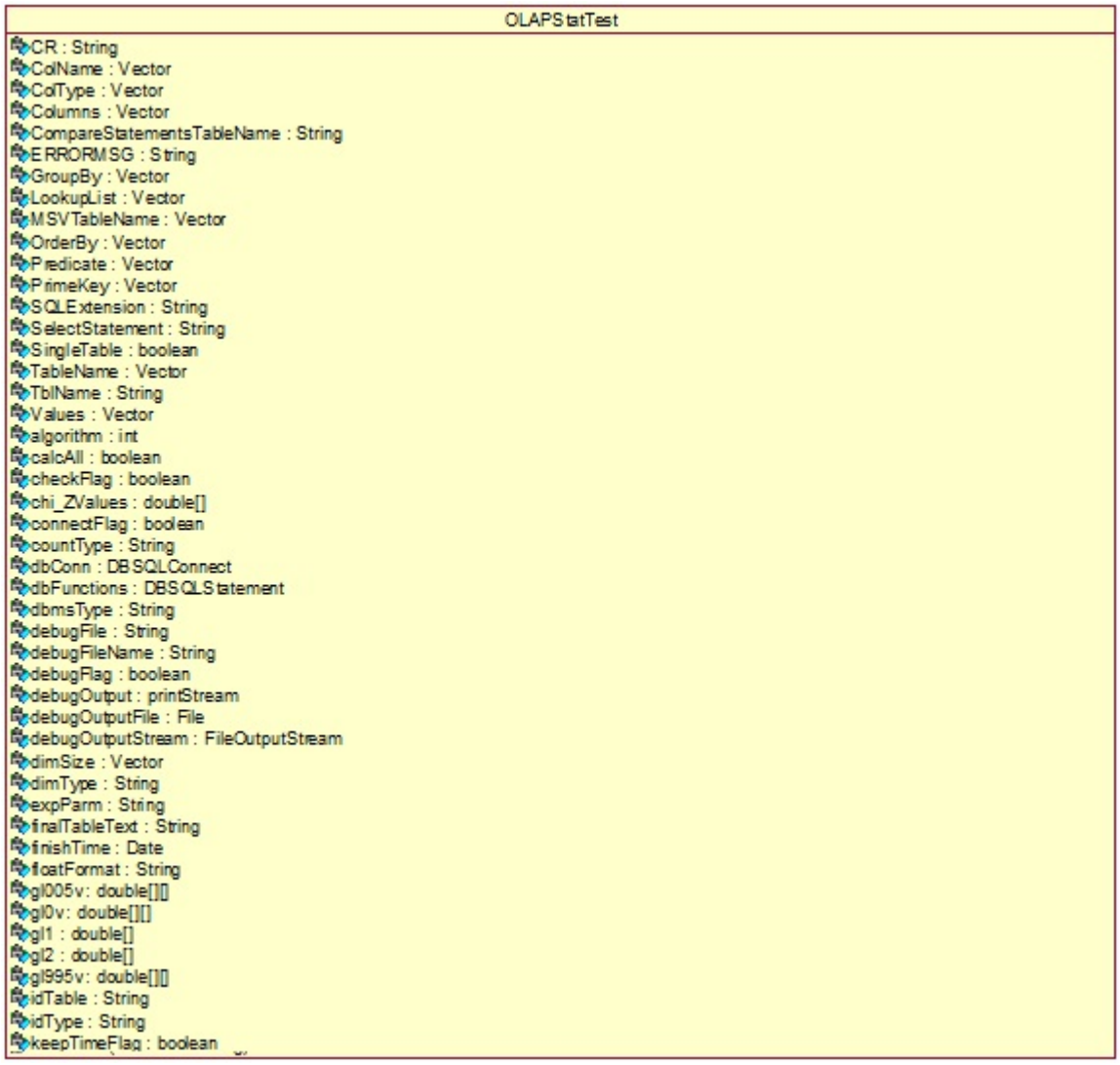

Figura B.9: Atributos clase principal

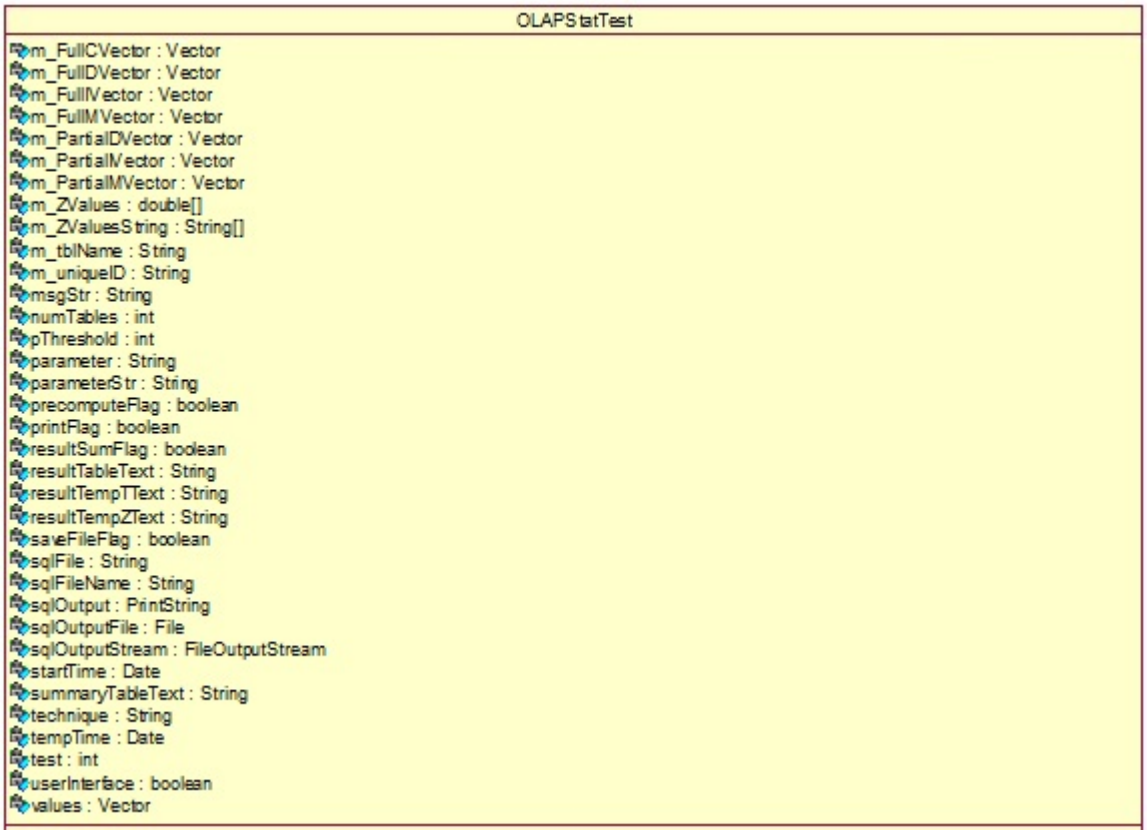

Figura B.10: Atributos clase principal

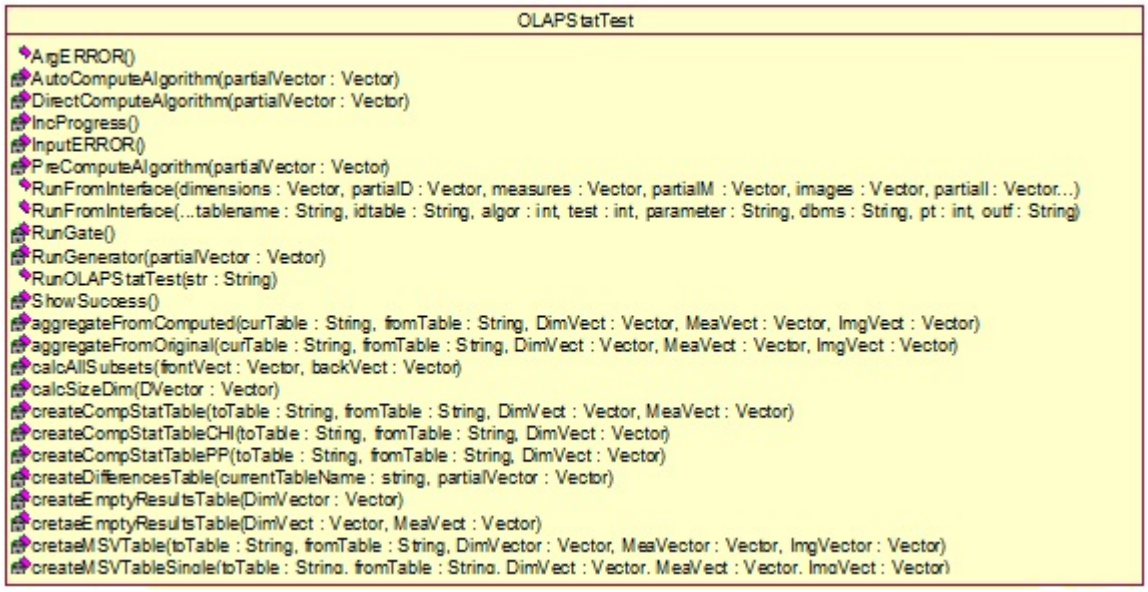

Figura B.11: Operaciones clase principal

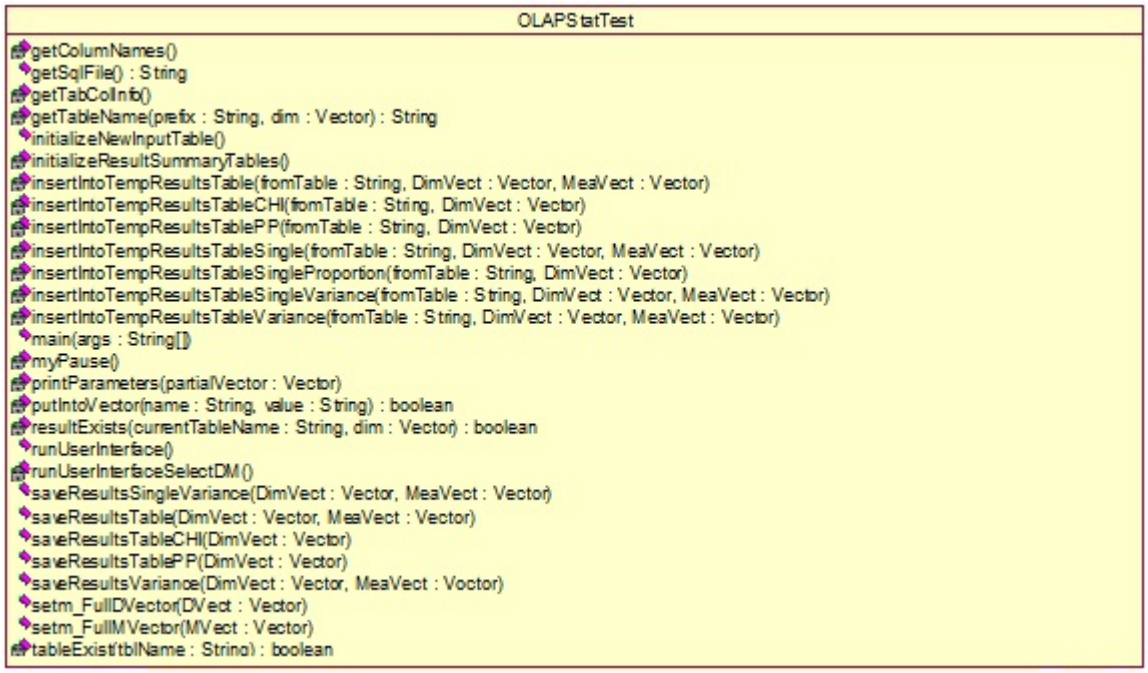

Figura B.12: Operaciones clase principal

B.2.2. Clases de Interfaz

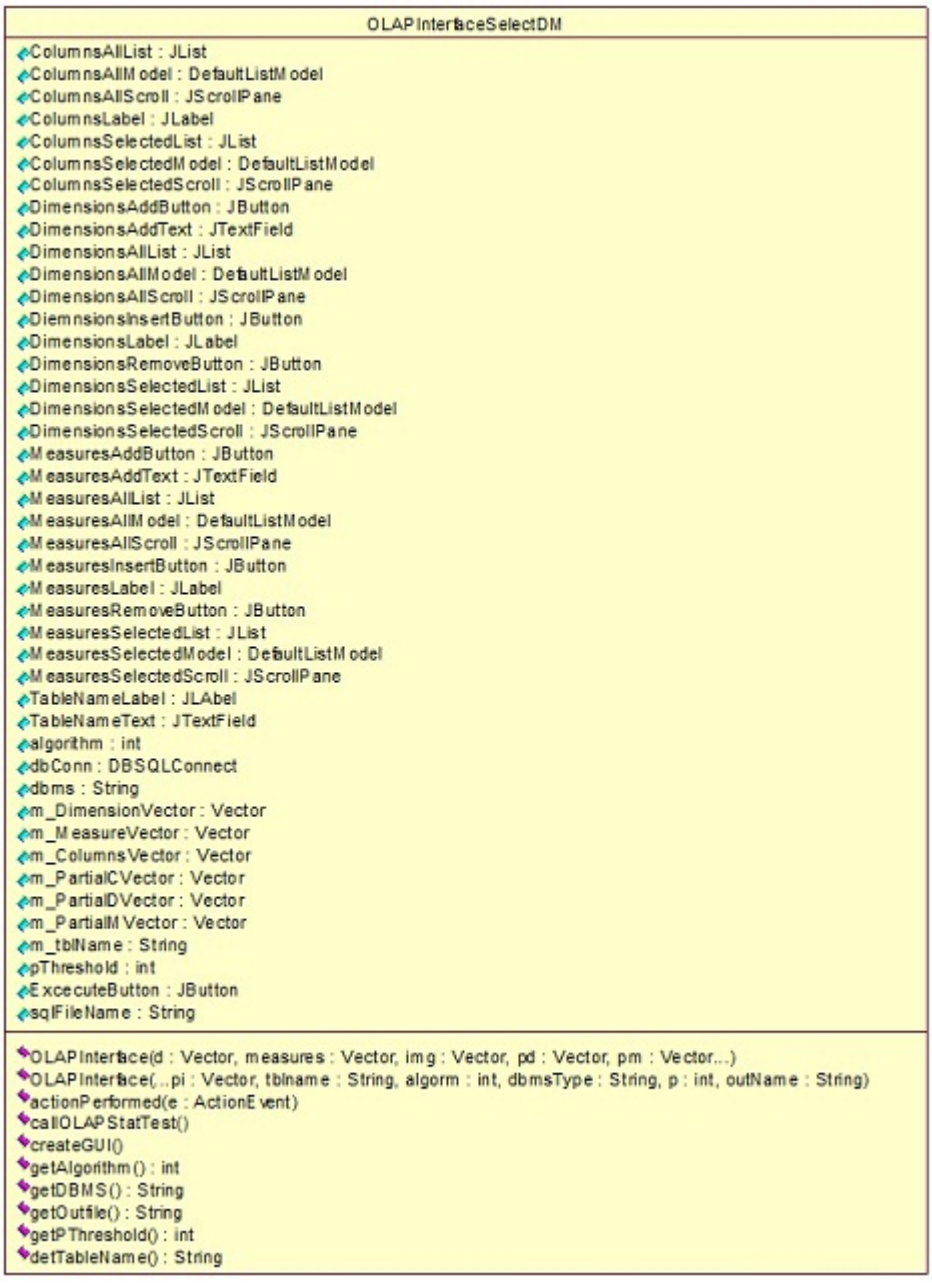

Figura B.13: Clase para presentar la interfaz donde se clasifican los campos

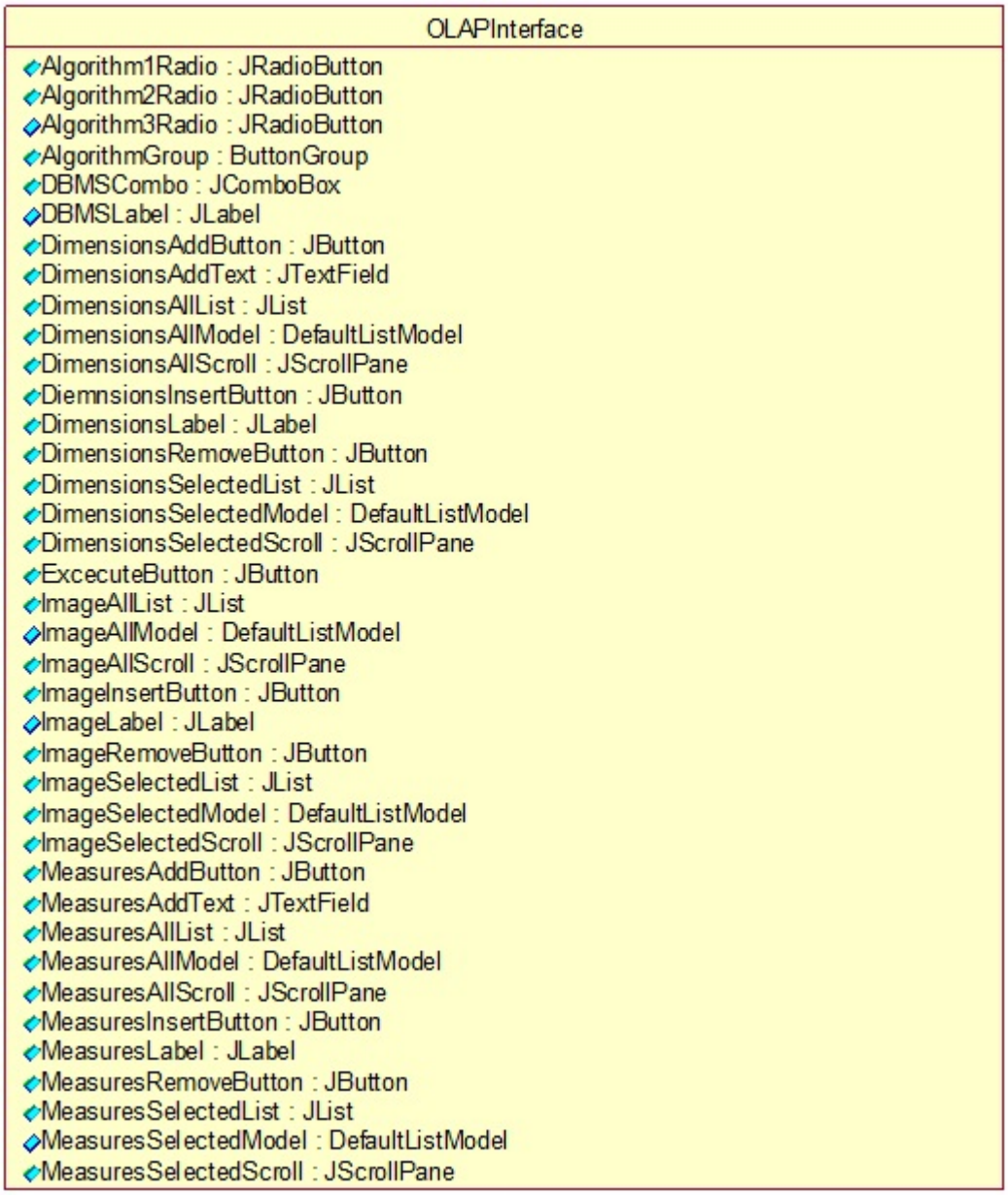

Figura B.14: Clase para presentar la interfaz donde se seleccionan los parámetros de la prueba estadística (Atributos)

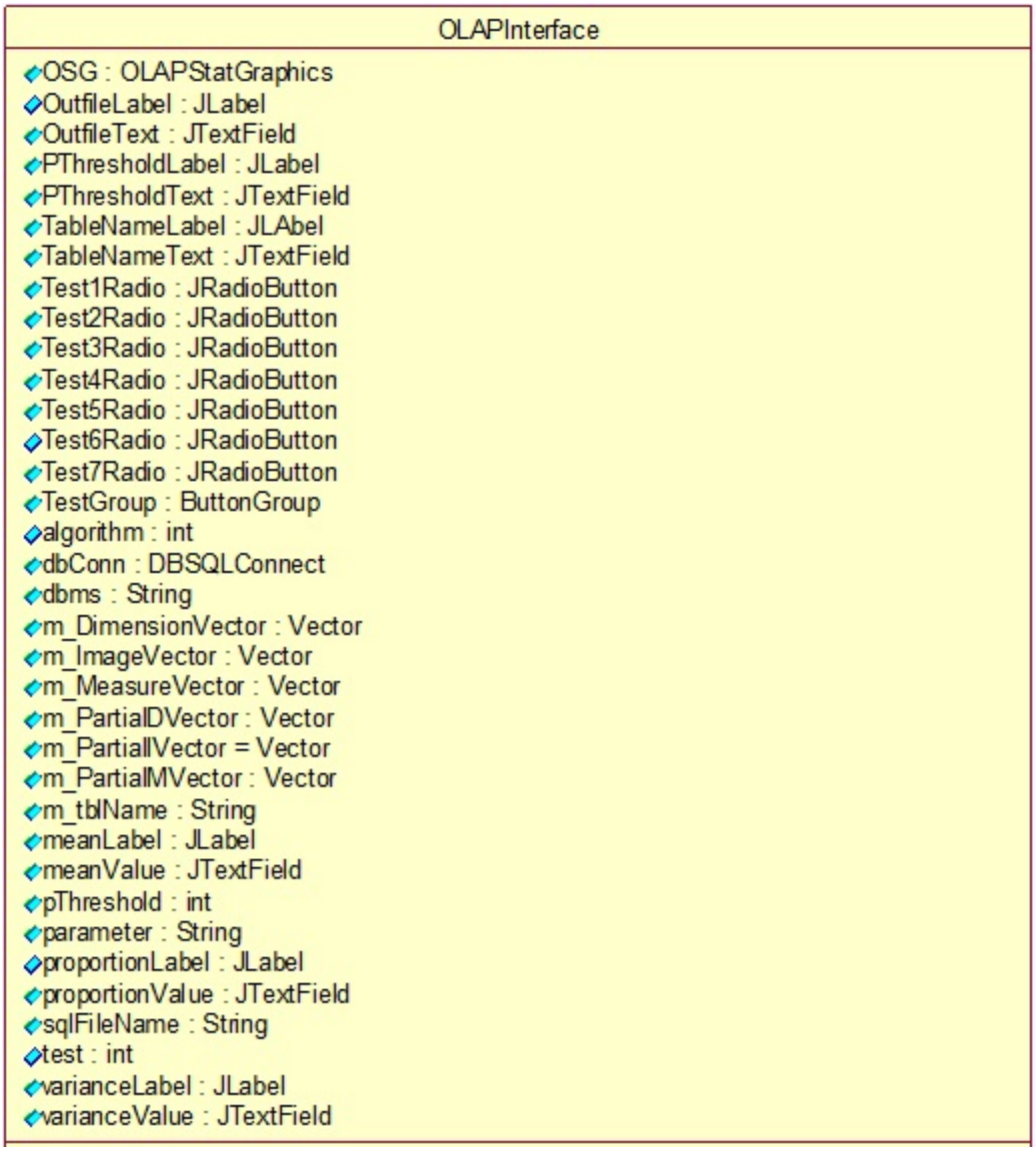

Figura B.15: Clase para presentar la interfaz donde se seleccionan los parámetros de la prueba estadística (Atributos)

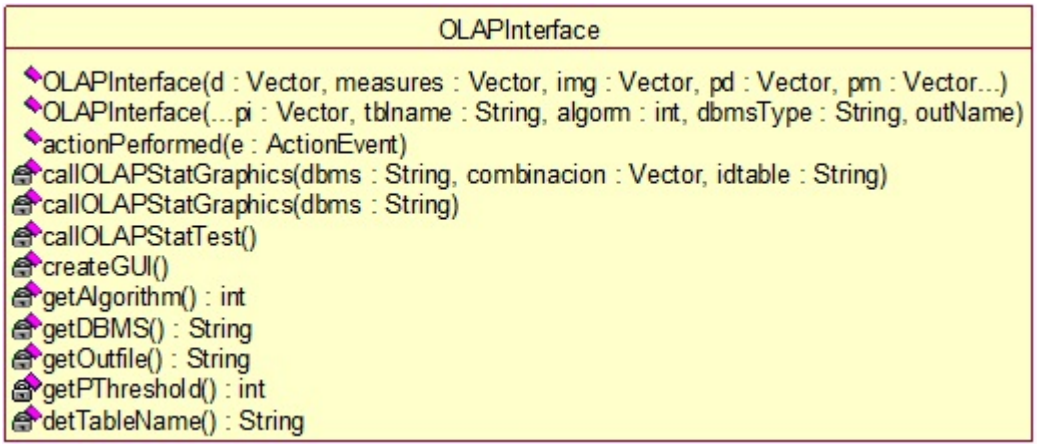

Figura B.16: Clase para presentar la interfaz donde se seleccionan los parámetros de la prueba estadística (Operaciones)

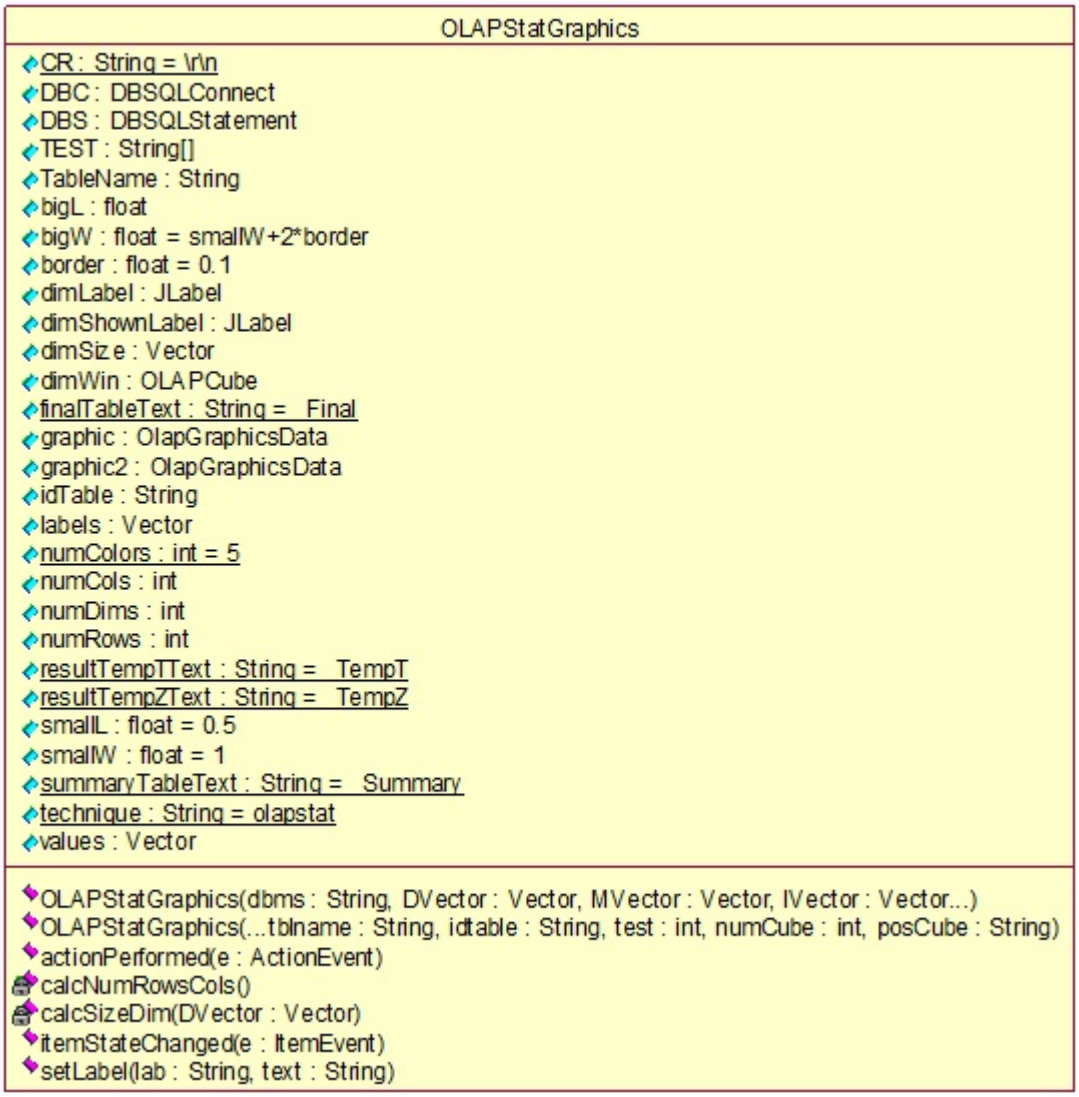

Figura B.17: Clase para presentar la interfaz donde se presentan los resultados visuales de la prueba

B.2.3. Clases de manejo de datos

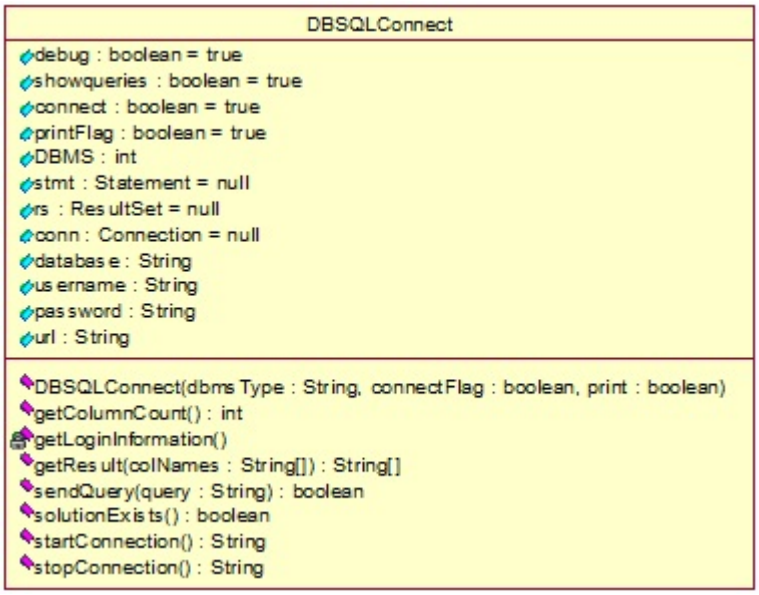

Figura B.18: Clase para el manejo de la conexión con la base de datos

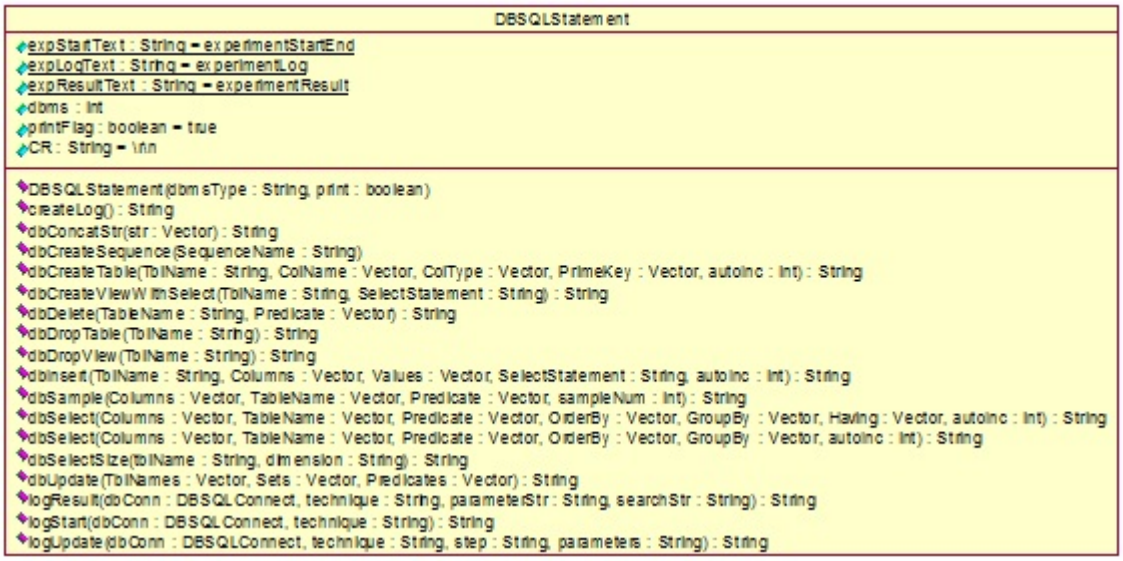

Figura B.19: Clase para la construcción de consultas SQL

B.3. Diagrama de clases

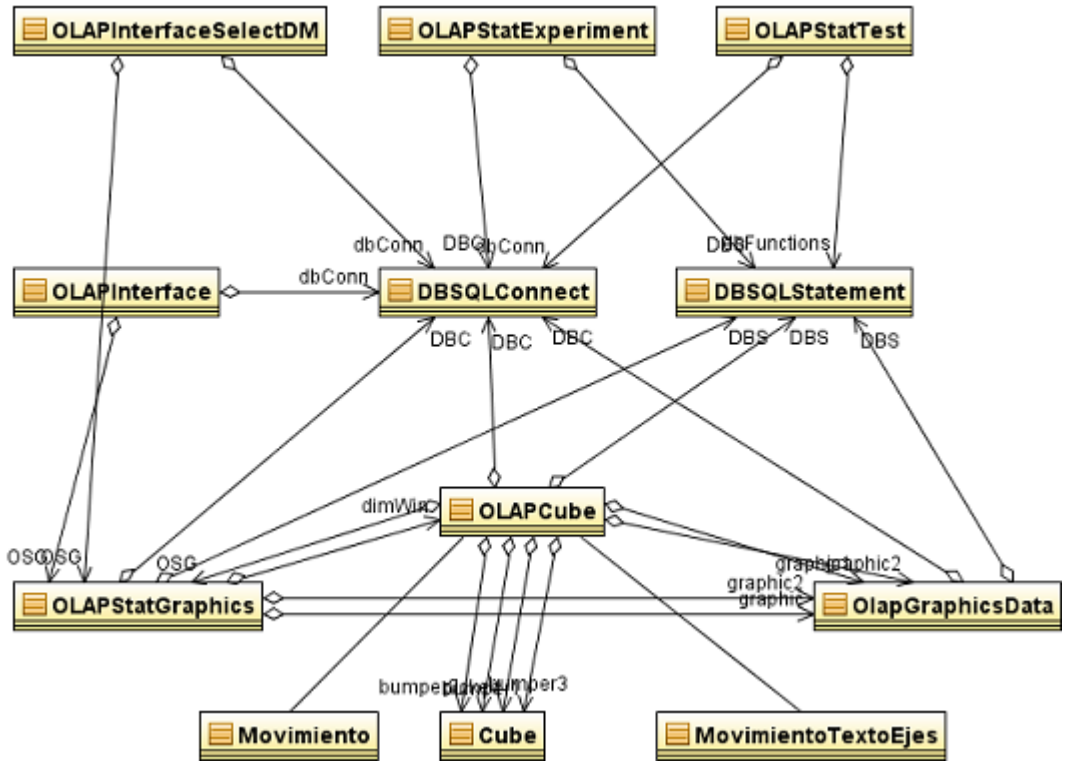

Figura B.20: Diagrama de clases