

Universidad Central “Marta Abreu” de Las Villas  
Facultad de Matemática, Física y Computación  
Licenciatura en Ciencia de la Computación



# TRABAJO DE DIPLOMA

*“Técnicas estadísticas para evaluar el comportamiento del delito”*

AUTORES: Sandra Cabrera Robaina  
Carlos Moreno Casaña

TUTORES: Dr. Gladys Casas Cardoso  
Ing. Manuel Cárdenas Cabrera

CONSULTANTE: Dr. Ricardo Grau Ábalo

“Año 52 de la Revolución”

Santa Clara, 2010

Hacemos constar que el presente Trabajo de Diploma ha sido realizado en la facultad de Matemática, Física y Computación de la Universidad Central “Marta Abreu” de Las Villas como parte de la culminación de los estudios de Licenciatura en Ciencia de la Computación, autorizando a que el mismo sea utilizado por la institución para los fines que estime conveniente, tanto de forma total como parcial y que además no podrá ser presentado en eventos ni publicado sin la previa autorización de la UCLV.

---

Firma del Autor

---

Firma del Autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

---

Firma del tutor

---

Firma del jefe del  
Laboratorio



*“En el ejercicio de la profesión se pone a prueba el espíritu creador del hombre.”*

*Ernesto Guevara de la Serna*

*A mis padres y mi hermana, que son las personas que más amo y de las que más amor y cariño he recibido.*

*Sandra*

*A mis padres, en especial a mi madre que es alguien muy importante para mí.*

*Carlos*

*Quiero agradecerle a todo aquel que de una forma u otra me brindo su ayuda y sin la cual no hubiese sido posible la realización de este trabajo, en especial:*

*A Gladys Casas tutora de este trabajo, por su entera disposición y paciencia.*

*A Manuel Cabrera también tutor de este trabajo por dedicarme momentos importantes de su tiempo.*

*A mis padres por su esfuerzo, dedicación y confianza.*

*A mi hermana que siempre ha estado para lo que he necesitado.*

*A mis tíos que me han apoyado siempre.*

*A Yaniel por su ayuda y amor incondicional.*

*A Leyda por toda la ayuda brindada.*

*A Amparo por toda su colaboración en este trabajo.*

*A mis compañeros de la carrera que me han ayudado a lo largo de estos 5 años.*

*A mi compañero de tesis Carlos por acompañarme en los momentos de tensión y escuchado durante todo este tiempo de tesis.*

*Al grupo de informática del Minint que también apporto mucho para este trabajo, con los cuales estuve todo este tiempo de tesis.*

*A la universidad y a todos los profesores que tuve durante toda la carrera por formarme como profesional y prepararme para la vida.*

*A la Revolución, que nos permite llegar a ser lo que seamos capaces, dependiendo solo de nuestra voluntad y esfuerzo.*

*A todos.....muchas gracias de corazón. Sandra*

*A mis padres que me han sabido guiar y apoyarme en todo momento bajo cualquier circunstancia.*

*A mi tía Hilda por quererme como un hijo y darme todo su apoyo.*

*A mis abuelos, tíos, primos y a toda mi familia que de una forma u otra he podido contar con ellos para lo que sea necesario.*

*A mi amigo Ernesto que me ha dedicado parte de su tiempo al colaborar en la elaboración de este trabajo.*

*A mi compañera de tesis Sandra por haberme comprendido y existir una buena concordancia durante todo el tiempo que hemos compartido juntos.*

*A mis tutores Gladys Y Manuel por habernos guiados en la metodología y ejecución de este proyecto.*

*A todos aquellos que han aportado su granito de arena para que se lleve a cabo este trabajo, en especial al grupo de informática del MININT.*

*A mis profesores que contribuyeron con mi formación durante toda la carrera.*

*Gracias a todos...*

*Carlos*

## **RESUMEN**

El presente trabajo está basado en la implementación y aplicación de técnicas no paramétricas de estadística utilizando como lenguaje de programación PL/SQL (o lenguaje procedimental de bases de datos), con vistas a efectuar un análisis estadístico inferencial sobre el almacén de datos, para extraer patrones del comportamiento del delito que ayuden a la toma de decisiones.

Para realizar todo el trabajo de implementación se utilizó la herramienta *Oracle SQL Developer* como sistema gestor de bases de datos; y para graficar se utilizó la herramienta *Oracle Business Intelligence*.

## **ABSTRACT**

The present thesis is related with the implementation and application of non parametric statistical tests using PL/SQL programming language (language of data bases). The main goal is to do a statistical analysis over a data warehouse in order to extract behavior patterns of misconduct. This process is important for the decision supports problem.

In order to do the implementation, the *Oracle SQL Developer* tool was used. In order to obtain the graphics, the *Oracle Business Intelligence* tool was used.

**ÍNDICE**

RESUMEN .....	VIII
ABSTRACT .....	IX
ÍNDICE .....	1
INTRODUCCIÓN .....	1
CAPÍTULO 1. Almacenes de datos e investigaciones estadísticas .....	4
1.1. Almacenes de datos (DW) .....	5
1.1.1. Componentes de un DW .....	7
1.1.2. Funciones de un DW .....	8
1.1.2. Data Marts .....	9
1.1.4. Arquitectura de un DW .....	10
1.2. Sistema gestor de bases de datos .....	11
1.3. Inteligencia de negocios .....	13
1.3.1. Componentes de BI .....	14
1.4. Aplicación de las técnicas estadísticas a las investigaciones .....	15
1.4.1. Pruebas estadísticas no paramétricas .....	17
Consideraciones generales del capítulo .....	19
CAPÍTULO 2. Implementación de métodos estadísticos .....	21
2.1. El caso de dos muestras independientes .....	21
2.1.1. Prueba U de Mann-Whitney .....	22
2.1.2. Algoritmo de U Mann-Whitney .....	24
2.2. El caso de k muestras independientes .....	25
2.2.1. Prueba de Kruskal-Wallis .....	26
2.2.2. Algoritmo de Kruskal-Wallis .....	27
2.3. El caso de dos muestras relacionadas .....	28
2.3.1. Prueba de Wilcoxon de rangos señalados .....	29
2.3.2. Algoritmo de Wilcoxon .....	32
2.4. El caso de k muestras relacionadas .....	33
2.4.1. Prueba de Friedman .....	34
2.4.2. Algoritmo de Friedman .....	36
2.5. Lenguaje de programación PL/SQL .....	37
2.6.1. Integración de PL/SQL con Oracle .....	37
Consideraciones generales del capítulo .....	40
CAPÍTULO 3. Aplicación de las pruebas estadísticas .....	41
3.1. Manual de usuario .....	41
3.2. Resultados de las pruebas estadísticas .....	47
3.3. Validación utilizando el SPSS .....	58
Consideraciones generales del capítulo .....	71
CONCLUSIONES .....	72
RECOMENDACIONES .....	73
BIBLIOGRAFÍA .....	74

## **INTRODUCCIÓN**

### **Breves antecedentes generales del tema**

El Ministerio del Interior (Minint) desde su creación ha establecido un control estricto sobre su trabajo. Al inicio este control se realizaba a través de archivos, pero con la llegada de la era digital fueron pasando gradualmente toda su información a formato digital.

A partir de 1999 en nuestro país comenzó a recibirse la influencia de la aplicación internacional de la Inteligencia Criminal, la cual utiliza técnicas computacionales para la toma de decisiones en el enfrentamiento a las distintas manifestaciones delictivas, y cuya implantación en el Ministerio del Interior parte del perfeccionamiento de la organización, informatización y modernización de los sistemas de trabajo.

El Minint trabaja en modernizar los métodos tradicionales de análisis a partir de asumir la Inteligencia Criminal como sistema de trabajo sustentado en la obtención, el procesamiento y análisis dirigido a determinar los rasgos de los delitos, a establecer vínculos entre los hechos, a evaluar redes de delincuentes y a ubicar cruces de información, con el objetivo de lograr un enfrentamiento cualitativamente más efectivo. En todo este proceso desempeña un rol importante la utilización óptima de los datos almacenados en los Registros Operativos.

Los primeros pasos dados en esta dirección se han concentrado en el Grupo de Tratamiento y Análisis de Información Criminal en la Dirección Nacional Antidrogas (DNA 2002), el Departamento Ministerial de Enfrentamiento a los Repetidores de Viajes y el Plan de Acción Integral, el cual se desarrolla de forma experimental en Ciudad de la Habana, La Habana, Sancti Spíritus y Santiago de Cuba, y se encuentra en proceso de extensión a otros territorios.

El presente trabajo forma parte de la estrategia del Minint para la implementación paulatina de un Sistema de Inteligencia Criminal aplicando un proceso de estadística sobre el almacén de datos implementado en el gestor de bases de datos Oracle.

## **Formulación del problema**

La Dirección del Ministerio del Interior realiza una tarea de suma importancia, ya que está responsabilizada por llevar un control adecuado del comportamiento del delito. En el Minint se tramita un alto volumen de información, la cual se encuentra en un almacén de datos; al que se le denomina: DW\_SAJO. Dicho almacén está implementado en el gestor de base de datos Oracle, sobre el que se desea utilizar técnicas de estadística para la determinación de factores relacionados con los delitos y detectar patrones de comportamiento de estos, para otorgar a los especialistas de argumentos sólidos para la toma de decisiones en las líneas de enfrentamiento.

## **Objetivo General**

- Extender las aplicaciones de la herramienta Oracle establecida como estándar, para efectuar el análisis inferencial sobre el almacén de datos DW\_SAJO y desarrollar completamente el control de los delitos, con vistas a realizar un proceso de extracción de patrones que nos ayuda a la toma de decisiones y así elevar de forma cualitativa el trabajo operativo del Minint.

## **Objetivos Específicos:**

- Realizar el estudio del almacén de datos DW\_SAJO donde se encuentra todo lo relacionado con el “Comportamiento del Delito”, para comprender los distintos procesos y tareas que allí se realizan.
- Analizar y proponer qué tipo de métodos de Estadística son más útiles y eficientes en el análisis de delitos.

- Implementar métodos estadísticos utilizando como lenguaje de programación PLSQL sobre el gestor de base de datos Oracle.
- Graficar los resultados que muestran estos métodos estadísticos en un portal web utilizando la herramienta Oracle Business Intelligence.

### **Preguntas de investigación**

- ¿Cuáles técnicas de Estadística serían más útiles y eficientes en el análisis de delitos?
- ¿Cuáles son las herramientas que ofrece Oracle para implementar las técnicas de estadística?

### **Justificación**

Es conocido que los hechos delictivos tienen un impacto negativo en la sociedad, por lo que es necesaria la aplicación de nuevas tecnologías de la Inteligencia Criminal para tratar de reducirlos. Esto es posible mediante la determinación de los factores que influyen en ellos, haciendo una labor profiláctica con estos y apoyándose en el desarrollo de las técnicas estadísticas y computacionales.

### **El presente Trabajo de Diploma se encuentra estructurado de la siguiente forma:**

El Capítulo 1, llamado “Almacenes de datos e investigaciones estadísticas”, hace referencia a todos los aspectos teóricos necesarios para el desarrollo de este trabajo. Los conceptos más importantes sobre almacén de datos, las distintas herramientas a utilizar y las pruebas estadísticas no paramétricas de manera general.

En el Capítulo 2, denominado “Implementación de las pruebas estadísticas”, se expone la teoría de cada método no paramétrico en particular, así como sus algoritmos y el lenguaje de programación que se utilizó.

En el Capítulo 3, tiene como nombre “Aplicación de las pruebas estadísticas”, donde se expone un manual de usuario; también se muestran los resultados de manera gráfica de algunos análisis en los que se aplicaron estas pruebas.

## **CAPÍTULO 1. Almacenes de datos e investigaciones estadísticas**

Como se mencionó, el presente capítulo tiene como finalidad realizar un estudio preliminar de los datos que se encuentran en el *Data Warehouse (DW)*, describir en detalle el comportamiento de las variables para la formulación de métodos estadísticos. También se mencionan los métodos de estadística más comunes y útiles para el análisis de delitos.

Estos métodos permiten detectar con mayor claridad dónde se llevarán a cabo los crímenes, los sectores que son más propensos a sufrir ataques en las distintas fechas, las horas y los días de la semana en que se concentran los asaltos según los municipios, entre otros.

El conocimiento de lo anterior permite realizar una asignación más eficiente y ordenada de los policías para disuadir los asaltos, es decir, tomar las medidas necesarias, como por ejemplo: enviar policías en multitudes hacia los lugares en donde los modelos indican que se perpetrarán la mayor cantidad de delitos, y de esta forma se logra prevenir que estos se materialicen para así reducir la tasa de criminalidad.

La idea que subyace detrás de la utilización de métodos estadísticos, es que los crímenes tienden a presentar patrones de comportamiento definidos a lo largo del tiempo. Estos patrones pueden ser estudiados y extraídos a través de técnicas estadísticas y matemáticas. En particular, como se desconocen las causas que motivan a los delincuentes a delinquir o, al menos, no se tiene un registro de ellas en las bases de datos, sólo es posible estudiar el comportamiento de los delitos considerando cómo estos se distribuyen a través del tiempo.

## 1.1. Almacenes de datos (DW)

Los sistemas de *Data Warehouse* han surgido como respuesta a la problemática de extraer información sintética a partir de datos atómicos almacenados en bases de datos de producción. Uno de los objetivos principales de este tipo de sistemas es servir como base de información para la toma de decisiones.

Los beneficios obtenidos por la utilización de este tipo de sistemas se basan en el acceso interactivo e inmediato a información estratégica de un área de negocios. Este acercamiento de la información al usuario final permite una toma de decisiones rápida y basada en datos objetivos obtenidos a partir de las bases de datos (eventualmente heterogéneas) de la empresa. Estos beneficios aumentan cuanto más importantes son las decisiones a tomar y cuanto más crítico es el factor tiempo.

Bill Inmon fue uno de los primeros autores en escribir sobre el tema de los almacenes de datos, define un *data warehouse* como:

“Un almacén de datos es una colección de datos integrada, orientada a temas, variante en el tiempo y no volátil, utilizada como apoyo para los procesos de toma de decisión”. (Inmon, 2005)

Estos términos significan:

- ❖ Integrada: Contiene una base de datos centralizada y consolidada que integra datos derivados de toda la organización. Los datos se almacenan en un formato consistente y existe un único esquema de representación.
  
- ❖ Orientada a temas: Los datos se organizan y se resumen por temas, por ejemplo, ventas, finanzas y transportación, para cada uno de los cuales el DW contiene sujetos, tales como productos, compradores y regiones. Por tanto, un DW se enfoca a las actividades del negocio, lo cual contrasta con los sistemas operacionales que se orientan a los procesos.

- ❖ Variante en el tiempo: Los datos se asocian con un punto en el tiempo o con un periodo. La toma de decisiones se apoya en diferentes modelos, estadísticos o de otro tipo, que necesitan información histórica. Esta característica básica de los datos en un DW difiere del comportamiento en el ambiente operacional donde los datos reflejan exactamente el momento actual.
- ❖ No volátil: Los datos no se modifican una vez introducidos (solo-lectura). Ello permite la optimización del acceso a los datos, puesto que el sistema no tiene que efectuar frecuentemente los chequeos de integridad requeridos por las operaciones de modificación. Además, se garantiza la disponibilidad de datos históricos.

Inmon defiende una metodología descendente (*top-down*) a la hora de diseñar un almacén de datos, ya que de esta forma se considerarán mejor todos los datos corporativos. En esta metodología los Data Marts se crearán después de haber terminado el *data warehouse* completo de la organización. (Inmon, 2005)

Ralph Kimball es otro conocido autor en el tema de los *data warehouse*, define un almacén de datos como: "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis". También fue Kimball quien determinó que un *data warehouse* no era más que: "la unión de todos los Data Marts de una entidad". Defiende por tanto una metodología ascendente (*bottom-up*) a la hora de diseñar un almacén de datos.

Las definiciones anteriores se centran en los datos en sí mismos. Sin embargo, los medios para obtener y analizar esos datos, para extraerlos, transformarlos y cargarlos, así como las diferentes formas para realizar la gestión de datos, son componentes esenciales de un almacén de datos. Muchas referencias a un almacén de datos utilizan esta definición más amplia. Por lo tanto, en esta definición se incluyen herramientas para la inteligencia

empresarial, herramientas para extraer, transformar y cargar datos en el almacén de datos, y herramientas para gestionar y recuperar los metadatos.

### 1.1.1. Componentes de un DW

Estos componentes son parte de una funcionalidad específica cada una por separado, y juntas constituyen los diversos elementos físicos y funcionales que conforman el Almacén de Datos.

- ❖ Orígenes de los datos: Están formados principalmente por los distintos sistemas de procesamiento de transacciones de datos. Los datos se recopilan y se consolidan para formar conjuntos de datos coherentes que reflejen con precisión las operaciones y el historial de la organización.
  
- ❖ Bases de datos relacionales: Se utilizan en el DW para clasificar, limpiar y transformar los datos recibidos desde los orígenes de datos, así como para contener y administrar cantidades masivas de datos en la base de datos del DW. Deben proporcionar capacidad para transferir y actualizar los datos rápidamente; realizar indizaciones eficaces y flexibles, y permitir sofisticadas capacidades de consulta, efectivas para organizar y recuperar datos.
  
- ❖ Área de preparación de datos: El área de preparación de datos, a veces denominada área de ensayos de datos, consiste en una base de datos relacional hacia la que se extraen los datos desde sus orígenes y en la que se transforman en formatos comunes; se comprueba su coherencia y su integridad; y se preparan para su carga definitiva en la base de datos del DW.
  
- ❖ Servicios de presentación: Están contruidos por el conjunto de todas las herramientas disponibles para facilitar el proceso de análisis y de evaluación, y varían desde simples informes hasta sofisticados algoritmos de minería de datos.

### 1.1.2. Funciones de un DW

El almacén de datos contiene datos que son necesarios o útiles para una organización, es decir, que se utiliza como un repositorio de datos para posteriormente transformarlos en información útil para el usuario. Debe entregar la información correcta a la gente indicada en el momento óptimo y en el formato adecuado, da respuesta a las necesidades de usuarios expertos, y utiliza Sistemas de Soporte a Decisiones (**DSS**), Sistemas de Información Ejecutiva (**EIS**) o herramientas para hacer consultas o informes. Los usuarios finales pueden hacer fácilmente consultas sobre sus almacenes de datos sin tocar o afectar la operación del sistema. (Inmon, 2005)

En el funcionamiento de un almacén de los datos son muy importantes las siguientes ideas:

- Integración de los datos provenientes de bases de datos distribuidas por las diferentes unidades de la organización y que con frecuencia tendrán diferentes estructuras (fuentes heterogéneas). Se debe facilitar una descripción global y un análisis comprensivo de toda la organización en el almacén de datos.
- 
- Separación de los datos usados en operaciones diarias en el almacén de datos para los propósitos de divulgación, de ayuda en la toma de decisiones, para el análisis y para operaciones de control. Ambos tipos no deben coincidir en la misma base de datos, ya que obedecen a objetivos muy distintos y podrían entorpecerse entre sí.

Periódicamente, se importan datos al almacén de datos de los distintos sistemas de planeamiento de recursos de la entidad (ERP) y de otros sistemas de software relacionados con el negocio para la transformación posterior. Es práctica común normalizar los datos antes de combinarlos en el almacén de datos mediante herramientas de extracción, transformación y carga (ETL). Estas herramientas leen los datos primarios

(a menudo bases de datos OLTP de un negocio), realizan el proceso de transformación al almacén de datos (filtración, adaptación, cambios de formato, etc.) y escriben en el almacén.

### 1.1.2. Data Marts

Los *Data marts* son subconjuntos de datos de un *data warehouse* para áreas específicas.

Entre las características de un *data mart* destacan:

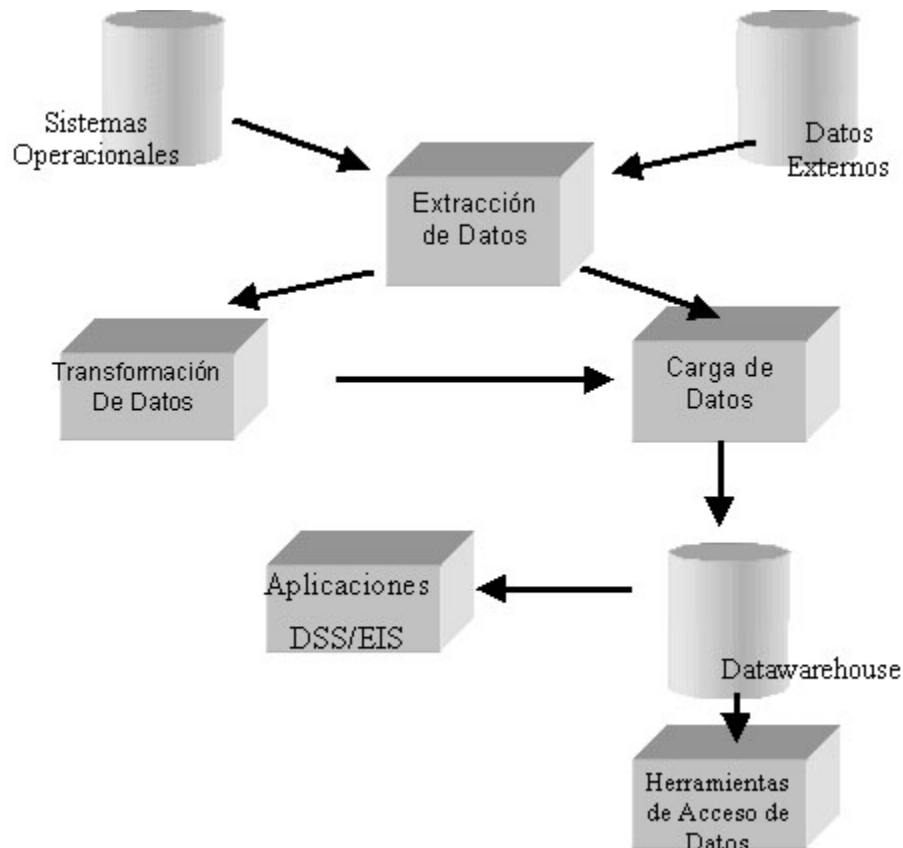
- Usuarios limitados.
- Área específica.
- Tiene un propósito específico.
- Tiene una función de apoyo.

Un ***Data mart*** es una versión especial de almacén de datos. Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de múltiples formas para que diversos grupos de usuarios los exploren de la forma más conveniente según sus necesidades. (Inmon, 2005)

El *Data mart* es un sistema orientado a la consulta, en el que se producen procesos *batch* de carga de datos (altas) con una frecuencia baja y conocida. Es consultado mediante herramientas OLAP (On Line Analytical Processing-Procesamiento Analítico en Línea) que ofrecen una visión multidimensional de la información. Sobre estas bases de datos se pueden construir EIS (Executive Information Systems, Sistemas de Información para Directivos) y DSS (Decision Support Systems, Sistemas de Ayuda a la toma de Decisiones). Por otra parte, se conoce como *Data Mining* al proceso no trivial de análisis de grandes cantidades de datos con el objetivo de extraer información útil, por ejemplo, para realizar clasificaciones o predicciones.

En síntesis, se puede decir que los *data marts* son pequeños data warehouse centrados en un tema o un área de negocio específico dentro de una organización.

#### 1.1.4. Arquitectura de un DW



**Figura 1.1** Arquitectura de un DW

- Datos operacionales: un origen de datos para el componente de almacenamiento físico DW.(Inmon, 2005)
- Extracción de Datos: selección sistemática de datos operacionales usados para poblar el componente de almacenamiento físico DW.
- Transformación de datos: procesos para resumir y realizar otros cambios en los datos operacionales para reunir los objetivos de orientación a temas e integración principalmente.

- Carga de Datos: inserción sistemática de datos en el componente de almacenamiento físico DW.
- Data warehouse: almacenamiento físico de datos de la arquitectura DW.
- Herramientas de Acceso al componente de almacenamiento físico DW: herramientas que proveen acceso a los datos.

## 1.2. Sistema gestor de bases de datos

Oracle es un sistema gestor de base de datos relacional extremadamente potente y flexible. Esta potencia y flexibilidad, sin embargo, implica también una cierta complejidad. Para poder diseñar aplicaciones útiles basadas en Oracle es necesario entender cómo manipula Oracle los datos almacenados en el sistema. Está orientado al acceso remoto y a internet. (David Planeaux, 2007)

Puede ser implementado en diferentes plataformas: UNIX, las distintas versiones Windows, Linux, etc. También se asienta en diferentes arquitecturas en cuanto a procesadores: Intel, Alpha, Sparc, RISC y algunas otras. Aunque la plataforma para la que fue pensado inicialmente Oracle fue UNIX, este ha logrado adaptarse a otras versiones como Windows, y ha desplazado a los gestores propios de este sistema operativo.

Como base de datos, Oracle ha acaparado casi todo el mercado de su sector, ya que ofrece una solución integral para resolver cualquier cuestión empresarial y se ha convertido en el software que casi todas las empresas están empleando para el asesoramiento a su propia toma de decisiones. Es actualmente uno de los paquetes de software más ampliamente extendidos en todas las compañías que tienen que gestionar una cantidad importante de información.

Oracle contiene una avanzada seguridad; los clientes pueden encriptar de manera transparente todos los datos de sus aplicaciones o solo algunas columnas específicas de la base de datos si así lo prefieren. Nos ofrece un rendimiento mucho mayor que cualquier otra plataforma de base de datos. Al permitirnos asignar nuestras propias zonas de memoria a sus datos y cualidades, podemos tener en todo momento controlados tanto el crecimiento como el rendimiento de los distintos esquemas que componen nuestra base; aunque, por otro lado, esto suponga un problema, ya que debemos estar pendientes en todo momento de su configuración para no sufrir fallos debido a algún problema de almacenamiento.(Lumpkin, 2007)

Resumiendo, Oracle alcanza hoy en día un buen nivel de madurez y de profesionalidad gracias especialmente a:

- Su transportabilidad funciona sobre decenas de plataformas.
- La potencia de sus instrumentos de desarrollo de aplicaciones.
- La riqueza de su diccionario de datos.
- Los mecanismos encargados de la seguridad y la confidencialidad.
- Una experiencia probada sobre el terreno y una buena presencia Oracle a nivel de formación, consejo y soporte técnico.

Producto de las características que hacen sobresalir este SGBD en su sector del mercado mundial y al hecho de estar desarrollando una aplicación que puede tornarse complicada, compleja y traer como consecuencia una base de datos igualmente compleja, se ha decidido utilizar Oracle como sistema gestor de base de datos para el desarrollo del Sistema Informático.

### 1.3. Inteligencia de negocios

Algo peor que no tener información disponible resulta tener mucha información y no saber qué hacer con ella. La Inteligencia de Negocios o *Business Intelligence* (Spiegel) es la solución a ese problema, pues por medio de dicha información puede generar escenarios, pronósticos y reportes que apoyen a la toma de decisiones, lo que se traduce en una ventaja competitiva. La clave para BI es la información y uno de sus mayores beneficios, la posibilidad de utilizarla en la toma de decisiones. En la actualidad hay una gran variedad de software de BI con aplicaciones similares que pueden ser utilizados en las diferentes áreas de la empresa, tales como: ventas, marketing, finanzas, etc. Son muchas las empresas que se han beneficiado por la implementación de un sistema de BI; además, se pronostica que con el tiempo se convertirá en una necesidad de toda empresa. (Vallejos, 2006)

En este nuevo mundo, la información reina, afirma Geoffrey A. Moore, director de *Chasm Group*. Vivimos en una época en que la información es la clave para obtener una ventaja competitiva en el mundo de los negocios. Para mantenerse competitiva una empresa, los gerentes y los tomadores de decisiones requieren de un acceso rápido y fácil a información útil y valiosa de la empresa. Una forma de solucionar este problema es por medio del uso de Business Intelligence o Inteligencia de Negocios.

#### **¿Qué es Inteligencia de Negocios?**

La Inteligencia de Negocios se puede definir como el proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos. Dentro de la categoría de bienes se incluyen las bases de datos de clientes, información de la cadena de suministro, ventas personales y cualquier actividad de marketing o fuente de información relevante para la empresa. (Vallejos, 2006)

BI apoya a los tomadores de decisiones con la información correcta, en el momento y

lugar correcto, lo que les permite tomar mejores decisiones de negocios. La información adecuada en el lugar y momento adecuado incrementa efectividad de cualquier empresa.

La tecnología de BI no es nueva, ha estado presente de varias formas por lo menos en los últimos 20 años, comenzando por generadores de reportes y sistemas de información ejecutiva en los años 80, afirma Candice Goodwin. Entiéndase como sinónimos de tecnología de BI los términos aplicaciones, soluciones o software de inteligencia de negocios.

Hay una gran variedad de aplicaciones o software que brindan a la empresa la habilidad de analizar de una forma rápida por qué pasan las cosas, y enfocarse a patrones y amenazas.

Oracle BI Enterprise Edition (OBI EE) es una suite completa e integrada de herramientas de análisis, diseñadas para proporcionar una gran visibilidad del negocio y conocimiento útil al espectro más amplio de usuarios. Con esta suite se permite que los usuarios, a través de una *interface web* y de forma autónoma, tengan acceso a la información más actualizada y relevante para su toma de decisiones y para crear acciones adecuadas como consecuencia del conocimiento obtenido.

### **1.3.1. Componentes de BI**

Todas las soluciones de BI tienen funciones parecidas, pero deben reunir al menos los siguientes componentes:

- ❖ Multidimensionalidad: la información multidimensional se puede encontrar en hojas de cálculo, bases de datos, etc. Una herramienta de BI debe ser capaz de reunir información dispersa en toda la empresa e incluso en diferentes fuentes, para así proporcionar a los departamentos la accesibilidad, el poder y flexibilidad

que necesitan para analizar la información. Por ejemplo, un pronóstico de ventas de un nuevo producto en varias regiones no está completo, si no se toma en cuenta también el comportamiento histórico de las ventas de cada región y la forma en que la introducción de nuevos productos se ha desarrollado en cada región en cuestión.(Vallejos, 2006)

- ❖ Data Mining: Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende, por lo general, de la habilidad para ver las nuevas tendencias o los cambios en ellas. Las aplicaciones de *data mining* pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no son muy evidentes.
  
- ❖ Agentes: Los agentes son programas que piensan. Ellos pueden realizar tareas a un nivel muy básico sin necesidad de intervención humana. Por ejemplo, un agente pueden realizar tareas un poco complejas, como elaborar documentos, establecer diagramas de flujo, etc.
  
- ❖ Data Warehouse: Es la respuesta de la tecnología de información a la descentralización en la toma de decisiones. Coloca información de todas las áreas funcionales de la organización en manos de quien toma las decisiones. También proporciona herramientas para búsqueda y análisis.

#### **1.4. Aplicación de las técnicas estadísticas a las investigaciones**

La aplicación de las técnicas estadísticas a las investigaciones para la planificación ideal de las experiencias y el procesamiento e interpretación de sus resultados, se ha convertido

en una necesidad imperiosa de las ciencias, particularmente, para obtener la mayor información del sistema estudiado con el mínimo de experiencias necesarias en el menor tiempo posible.

Las investigaciones del tipo social han cobrado una gran importancia en los últimos tiempos, debido al desarrollo del carácter social de la producción, al desarrollo de la ciencia y a la ampliación de los campos de aplicación: sociología, psicología, medicina, etc. Para la realización de estas investigaciones, los métodos de análisis estadísticos y, en particular, las pruebas no paramétricas constituyen una herramienta matemática de suma utilidad.

En las ciencias de la conducta se efectúan investigaciones para determinar en qué medidas son aceptables las hipótesis que derivamos de las teorías de las conductas. Después de seleccionar cierta hipótesis que parece importante en una teoría determinada, se recogen los datos empíricos que dan información directa acerca de la aceptabilidad de esa hipótesis. La decisión acerca del significado puede conducir a la confirmación, revisión o el rechazo de la hipótesis y, con ella, la teoría que la originó. (Siegel, 1988a)

El primer aspecto a considerar para poder aplicar la estadística a un trabajo experimental es definir, desde el comienzo, frente a qué tipo de sistema o población nos encontramos. Desde el punto de vista de las técnicas estadísticas a utilizar se tienen dos grandes ramas: la estadística paramétrica y la estadística no paramétrica.

Existen varios criterios para la elección de una prueba estadística que sirva para tomar decisiones acerca de una hipótesis en investigación. Estos criterios pueden resumirse en:

- ❖ La potencia de la prueba.
- ❖ La aplicabilidad del modelo estadístico en que se basan los datos de investigación.
- ❖ Potencia-eficiencia.

- ❖ El nivel de medidas logrado en la investigación.

Una prueba estadística paramétrica es más eficaz cuando se reúnen todas las suposiciones de su modelo estadístico y al medir las variables analizadas por lo menos con una escala de intervalos. Sin embargo, aun cuando se satisfagan todas las suposiciones de la prueba paramétrica acerca de la población y los requerimientos de fuerza y medición, el concepto de potencia–eficiencia señala que al aumentar apropiadamente el tamaño de la muestra, se puede usar una prueba no paramétrica en lugar de una paramétrica, sin perder potencia para rechazar la hipótesis inicial.(Arsula, 1984)

En el presente trabajo se utilizaron las pruebas estadísticas no paramétricas con la siguiente clasificación:

- ❖ El caso de dos muestras independientes.
- ❖ El caso de dos muestras relacionadas.
- ❖ El caso de k muestras independientes.
- ❖ El caso de k muestras relacionadas.

### **1.4.1. Pruebas estadísticas no paramétricas**

Una prueba estadística no paramétrica es aquella cuyo modelo no especifica las condiciones de los parámetros de la población de la que se sacó la muestra. Existen algunas suposiciones que se asocian con la mayoría de las pruebas estadísticas no paramétricas; pero son pocas y mucho más débiles que las asociadas con las pruebas paramétricas. Además, las no paramétricas no requieren mediciones tan fuertes; la mayoría de las pruebas no paramétricas se aplican a datos de una escala ordinal y algunas a los de una escala nominal. (Siegel, 1988a)

#### **Ventajas de las pruebas no paramétricas**

- ❖ Las declaraciones de probabilidad obtenidas de la mayoría de las pruebas estadísticas no paramétricas son probabilidades exactas (excepto en el caso de

muestras grandes, que proporcionan excelentes aproximaciones), independientemente de la forma de la distribución de la población de la que se tomó la muestra. El cálculo de la probabilidad no depende de la forma de la población; aunque algunas pruebas no paramétricas supongan identidad de forma de dos o más distribuciones de población. (Siegel, 1988a)

- ❖ Si los tamaños de la muestra son tan pequeños como  $N=6$ , no hay alternativa de elección de una prueba estadística no paramétrica, a menos que se conozca exactamente la naturaleza de la distribución de la población.
- ❖ Hay pruebas estadísticas no paramétrica adecuadas para observaciones hechas en poblaciones diferentes. Ninguna prueba paramétrica puede manejar tales datos sin exigirnos suposiciones aparentemente irreales.
- ❖ Las pruebas estadísticas no paramétrica son útiles tanto para datos inherentes a los rangos como datos cuyos puntajes aparentemente numéricos tienen fuerza de rangos. Esto es, el investigador solamente puede decir de sus sujetos que uno comparte en mayor o menor grado cierta característica de otro, sin especificar la cantidad. Por ejemplo, al estudiar una variable como la ansiedad, se puede afirmar que el sujeto A sufre más ansiedad que el sujeto B, sin saber en modo alguno cuanta ansiedad tiene A. Datos que son inherentes a clasificaciones en rangos o solamente a categorías como positivo o negativo (más o menos, mejor o peor), pueden ser tratados por métodos no paramétricos; pero no pueden ser manejados por métodos paramétricos, a menos que se hagan suposiciones precarias y quizás irreales acerca de las distribuciones de base.
- ❖ Los métodos no paramétricos son útiles para datos clasificatorios, medidos en una escala nominal. Ninguna técnica paramétrica se aplica a tales datos.

- ❖ Las pruebas estadísticas no paramétricas son típicamente mucho más fáciles de aplicar y más sencillas que las pruebas paramétricas.

### **Desventajas de las pruebas no paramétricas**

- ❖ Si todos los supuestos del modelo estadístico paramétrico se satisfacen realmente por los datos y si la medida tiene la fuerza requerida, las pruebas estadísticas no paramétricas no aprovechan cabalmente información contenida en los datos. El grado de desperdicio se expresa por la potencia-eficiencia de la prueba no paramétrica. (Si una prueba estadística no paramétrica tiene una potencia-eficiencia de 90 %, esto significa que al ser satisfecha todas las condiciones de la prueba paramétrica, el método paramétrico apropiado será tan efectivo con una muestra menor en un 10 % que la empleada en el análisis no paramétrico.)

Por lo general, las pruebas paramétricas son más poderosas que las pruebas no paramétricas y deben usarse siempre que sea posible. Es importante observar que aunque las pruebas no paramétricas no hacen suposiciones sobre la distribución de la población que se muestrea, muchas veces se apoyan en distribuciones muestrales como la normal o la Chi cuadrado. (Siegel, 1988a)

### **Consideraciones generales del capítulo**

- El desarrollo de las computadoras ha repercutido de forma extraordinaria en la sociedad, al revolucionar los métodos tradicionales de trabajo y elevar la posibilidad de usar herramientas poderosas que faciliten el trabajo. La evaluación del comportamiento del delito no constituye una excepción.
- Se definió el concepto de almacén de datos, sus componentes y funcionalidades, para poder realizar estudios rápidos y efectivos sobre grandes volúmenes de datos.

- Para la realización de este trabajo se definieron los métodos de análisis estadísticos y en particular las pruebas no paramétricas, que son fundamentales para evaluar el comportamiento del delito.
- Se definieron varias herramientas para la realización definitiva de este trabajo, todas pertenecientes a Oracle establecida como estándar en el Minint; ellas son Oracle SQL Developer, como sistema gestor de bases de datos, y Oracle Business Intelligence, el cual permite a usuarios a través de una *interface web* y de forma autónoma tengan acceso a la información más actualizada y relevante para la toma de decisiones.

## **CAPÍTULO 2. Implementación de métodos estadísticos**

En el siguiente capítulo se exponen los fundamentos matemáticos de los métodos no paramétricos seleccionados, así como su algoritmo de cálculo. Se muestran detalles acerca del lenguaje de programación que se utilizó para implementarlos.

### **2.1. El caso de dos muestras independientes**

Cuando se tienen problemas en los que se quieren comparar dos muestras independientes, las pruebas estadísticas no paramétricas más importantes son:

- a) Prueba de Fisher.
- b) Prueba Chi-Cuadrado.
- c) Prueba de la Mediana.
- d) Prueba U de Mann Whitney.
- e) Prueba de Kolmogorov-Smirnov.
- f) Prueba de las Rachas de Wald-Wolfwits.

Estas pruebas no paramétricas se utilizan para determinar si existen diferencias significativas entre dos muestras independientes. En otras palabras, ellas calculan la probabilidad de que las muestras procedan de la misma población.

la prueba estadística paramétrica que se utiliza en estos casos es la prueba T, que supone que los puntajes (que se suman al calcular las medias) son observaciones independientes de poblaciones distribuidas normalmente con varianzas iguales y requiere que las observaciones se midan por lo menos en una escala de intervalos.

La prueba U de Mann-Whitney es la alternativa de elección de la prueba T (Siegel, 1988a, SIEGEL, 1988b). Ella supone que los datos a comparar tengan al menos una medida ordinal.

### 2.1.1. Prueba U de Mann-Whitney

En estadística la prueba U de Mann-Whitney, también llamada de Mann-Whitney-Wilcoxon, prueba de suma de rangos Wilcoxon o prueba de Wilcoxon-Mann-Whitney, es una prueba no paramétrica aplicada a dos muestras independientes cuyos datos han sido medidos al menos en una escala de nivel ordinal.

Fue propuesto inicialmente en 1945 por Wilcoxon para el caso de tamaños muestrales iguales. Festinger en 1946 desarrolló un procedimiento equivalente al de Wilcoxon. Pero fueron Mann y Whitney en 1947 los primeros en extender el procedimiento al caso de tamaños muestrales desiguales y los primeros también en proporcionar tablas para poder utilizar el procedimiento con muestras pequeñas. Precisamente las aportaciones de Mann y Whitney resultaron las que más contribuyeron a la divulgación del procedimiento; de ahí que, generalmente, sea conocido como prueba de Mann-Whitney. El test es virtualmente idéntico a la realización de una prueba paramétrica ordinaria T de dos muestras en los datos después de haber ordenado las muestras combinadas. (Spiegel, 2009)

La prueba calcula el llamado estadístico  $U$ , cuya distribución para muestras con más de 20 observaciones se aproxima bastante bien a la distribución normal.

El test se basa en el ranqueo de los datos de la muestra total (compuesta de dos grupos) y la observación de si estos valores ranqueados de un grupo y del otro se intercalan adecuadamente como una medida de que las distribuciones no difieren.

#### Propósito:

Para probar si dos muestras independientes han sido tomadas de la misma población.

#### Método:

a) Teoría

El estadístico  $U$  de Mann-Whitney se define como el número de veces que  $Y$  precede a  $X$  en el arreglo combinado ordenado de las dos muestras independientes.

$$X_1, X_2, \dots, X_n \quad \text{y} \quad Y_1, Y_2, \dots, Y_m$$

Es una sola secuencia de  $N = n + m$  variables.

Se define:

$$D_{ij} = \begin{cases} 1 & \text{si } Y_j \leq X_i \\ 0 & \text{si } Y_j > X_i \end{cases}$$

para  $i = 1, 2, \dots, n$  y  $j = 1, 2, \dots, m$

Se tiene que:

$$U_1 = nm - \frac{n(n+1)}{2} \qquad U_2 = nm - \frac{m(m+1)}{2}$$

O por otra expresión parecida a esta (donde se tienen calculadas  $\sum R_1$  y  $\sum R_2$ , que es la suma de los rangos para la muestra 1 y 2):

$$U_1 = nm + \frac{n(n+1)}{2} - \sum R_1 \qquad U_2 = nm + \frac{m(m+1)}{2} - \sum R_2$$

Donde  $n$  y  $m$  son el tamaño respectivo de cada muestra.

La aproximación a la normal,  $z$ , cuando tenemos muestras lo suficientemente grandes viene dada por la expresión:

$$z = \frac{(U - m_u)}{\sigma_u}$$

Donde  $m_u$  y  $\sigma_u$  son la media y la desviación estándar de  $U$  si la hipótesis nula es cierta, y vienen dadas por las siguientes fórmulas:

$$m_x = nm/2$$

$$\sigma_x = \sqrt{\frac{nm(n+m+1)}{12}}$$

Los cálculos tienen que tener en cuenta la presencia de observaciones idénticas a la hora de ordenarlas. No obstante, si su número es pequeño, se puede ignorar esa circunstancia.

Esta prueba es una de las más adecuadas para la comparación de dos muestras de cualquier tamaño y es más exacta que la prueba de la mediana, además, alcanza hasta 90 % de la potencia o agudeza de la prueba t de Student. (Mann, 2007)

### 2.1.2. Algoritmo de U Mann-Whitney

Estos son los pasos de la prueba de U Mann-Whitney:

1. Determinar el tamaño n, m de las muestras  $X_i$  ( $i = 1, 2, \dots, n$ )

$$Y_j$$
 ( $j = 1, 2, \dots, m$ )

Si n y m son menores que 20, se consideran muestras pequeñas, pero si son mayores que 20, se consideran muestras grandes.

2. Arreglar los datos en rangos del menor al mayor valor. En caso de que existan ligas o empates de rangos iguales, se deberán detectar para un ajuste posterior.
3. Calcular los valores de  $U_1$  y  $U_2$ , de modo que se elija el más pequeño para comparar con los críticos de U Mann-Whitney de la tabla de probabilidades asociadas con valores pequeños.
4. En caso de muestras grandes, calcular el valor de Z, pues en estas condiciones se distribuye normalmente.

$$Z = \frac{U - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

5. Decidir si se acepta o se rechaza la hipótesis.

## 2.2. El caso de k muestras independientes

Para el caso de k muestras independientes, las pruebas estadísticas no paramétricas que se reportan en la literatura son: (Arsula, 1984)

- a) La prueba Chi-Cuadrado.
- b) La prueba de la Mediana.
- c) Prueba de Kruskal-Wallis.

Estas pruebas se utilizan para probar la significación de diferencias entre tres o más grupos o muestras independientes.

La prueba de Kruskal Wallis constituye una generalización natural de la de Mann-Whitney para dos muestras independientes,(Siegel, 1988a). Esa es la causa por la que se seleccionó para ser implementada. Esta prueba requiere que se tenga una medición ordinal de la variable y se utiliza para probar si k muestras independientes pueden haber sido recogidas de la misma población continua.

La segunda y la tercera pruebas pueden aplicarse a los mismos datos en estudios. La prueba de Kruskal-Wallis es más eficiente, pues usa más la información de las observaciones, convierte los puntajes en rangos, mientras que la prueba de la mediana los convierte simplemente en signos más o menos.

Para el caso de k muestras independientes la técnica paramétrica que se usa es la prueba F, que supone que las observaciones sean tomadas de manera independiente de poblaciones distribuidas

normalmente, todas las cuales tienen la misma variante, su requisito es por lo menos una medida de intervalo de la variable estudiada. (Arsula, 1984)

### 2.2.1. Prueba de Kruskal-Wallis

En estadística, la prueba de Kruskal-Wallis (de William Kruskal y W.Allen Wallis) es un método no paramétrico para probar si un grupo de datos proviene de la misma población. Intuitivamente, es idéntico al ANOVA (*Analysis of Variance*) con los datos reemplazados por categorías. Es una extensión de la prueba de la U de Mann-Whitney para 3 o más grupos.

Ya que es una prueba no paramétrica, la prueba de Kruskal-Wallis no asume normalidad en los datos, en oposición al tradicional ANOVA. Si supone bajo la hipótesis nula que los datos vienen de la misma distribución. Una forma común en que se viola este supuesto es con datos heterocedásticos. (Spiegel, 2009)

#### Propósito:

Probar si K muestras independientes provienen de la misma población.

#### Método:

##### a) Teoría:

Se tienen k muestras independientes de tamaño  $N_1, N_2, \dots, N_k$ . Las k muestras se unen en una sola de tamaño N y sus valores se sustituyen por rangos. Luego se calcula la suma de los rangos asociados a cada población.

La prueba de Kruskal-Wallis determina la desigualdad de las poblaciones. Bajo la hipótesis de que las k muestras proceden de la misma población y para valores de  $N_i$  no muy pequeños, se cumple que:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1)$$

Donde:

$n_j$ : Es el número de observaciones en el grupo  $j$

$R_j$ : Es la suma de los rangos en el grupo  $j$

$N$ : Es el número total de observaciones entre todos los grupos

Se puede realizar una corrección para los valores repetidos dividiendo  $H$  por

$$\frac{\sum t}{N^3 - N}$$

Donde  $\sum t$  indica sumar en todos los grupos de empate de tamaño  $t$ , el valor  $t^3 - 3$

Se considera a la prueba de Kruskal-Wallis la más potente para comparar más de 2 variables continuas independientes.

La prueba de Kruskal-Wallis es más eficiente que la extensión de la prueba de la mediana, pues utiliza más la información de las observaciones al convertir los puntajes en rangos en lugar de simplemente dicotomizarlos por encima y por debajo de la mediana. (Wallis, 2007)

### 2.2.2. Algoritmo de Kruskal-Wallis

Los pasos para usar el análisis de varianza de una clasificación por rangos de Kruskal-Wallis son los siguientes:

1. Se ordenan todas las observaciones de los  $k$  grupos en una sola serie, y se asignan rangos de 1 a  $N$ .
2. Se determina el valor de  $R$  (la suma de los rangos) para cada uno de los  $k$  grupos de rangos.
3. Si una gran proporción de las observaciones están ligadas, se calcula el valor  $H$  con la fórmula:

$$H = \frac{\frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1)}{1 - \frac{\sum T}{N^2 - N}}$$

De otra manera, se usa la fórmula:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1)$$

4. El método para determinar la significación del valor observado de H depende del tamaño de k y del tamaño de los grupos.
  - a) Si  $k=3$  y si  $n_1, n_2, n_3$  son iguales o menores a 5, la tabla O puede usarse para determinar la probabilidad asociada conforme a  $H_{\alpha}$  de una H tan grande como la observada.
  - b) En los otros casos, la significación de un valor tan grande como el valor observado de H puede determinarse por medio de la tabla C, con  $d_f = k - 1$ .
5. Si la probabilidad asociada con el valor observado de H es igual o menor que el nivel de significación,  $\alpha$ , previamente fijado, se rechaza  $H_0$  y se acepta  $H_1$ .

### Potencia-Eficiencia

Comparada con la prueba paramétrica más poderosa, la prueba F, en las condiciones en que las suposiciones asociadas con el modelo estadístico de la prueba F son satisfechas, la prueba de Kruskal-Wallis tiene una eficiencia asintótica  $3/\pi = 95.5\%$ .

### 2.3. El caso de dos muestras relacionadas

En el caso de dos muestras relacionadas, con el fin de establecer la diferencia entre dos tratamientos o si un tratamiento es “mejor” que otro, fueron estudiadas las siguientes pruebas estadísticas no paramétricas:

- a) La prueba McNemar para la significación de los cambios.

- b) La prueba de los signos.
- c) La prueba de rangos señalados y pares igualados de Wilcoxon.
- d) La prueba de aleatoriedad para pares igualados.

De estas sólo se desarrolló la prueba de Wilcoxon, que es aplicable cuando el investigador puede ordenar las diferencias observadas para los diferentes pares igualados. Cuando la prueba de Wilcoxon se usa para datos que satisfagan las condiciones de la Prueba T, su potencia-eficiencia es de cerca del 95 % para muestras grandes y menor para muestras más pequeñas.

La técnica paramétrica que se utiliza para analizar datos provenientes de dos muestras relacionadas es la ya mencionada, la prueba T, la cual consiste en aplicársela a los puntajes de diferencias; estos puntajes se obtienen de los dos puntajes de los miembros de cada pareja igualada o de los dos puntajes de cada sujeto bajo las dos condiciones; esta prueba supone que estos puntajes de diferencias estén distribuidos normal e independientemente de la población de la cual se tomó la muestra y requieren que sean medidos por lo menos en una escala de intervalos. (Arsula, 1984)

### **2.3.1. Prueba de Wilcoxon de rangos señalados**

La prueba de los signos de Wilcoxon es una prueba no paramétrica para el caso de dos muestras relacionadas o mediciones repetidas en una sola muestra y determinar si existen diferencias entre ellas. Se utiliza como alternativa a la prueba t de Student cuando no se puede suponer la normalidad de dichas muestras. Debe su nombre a Frank Wilcoxon, que la publicó en 1945.

Al igual que la prueba t para muestras relacionadas, la prueba de Wilcoxon implica comparaciones entre las mediciones de las diferencias por pares, por lo que requiere que los datos estén en un intervalo de nivel de medición. Sin embargo, no requieren supuestos sobre la forma de la distribución de las mediciones.

Se utiliza cuando la variable subyacente es continua, pero presupone ningún tipo de distribución particular.

En ocasiones, esta prueba se usa para comparar las diferencias entre dos muestras de datos tomados antes y después del tratamiento, cuyo valor central se espera que sea cero. Las diferencias iguales a cero son eliminadas y el valor absoluto de las desviaciones con respecto al valor central son ordenadas de menor a mayor. A los datos idénticos se les asigna el lugar medio en la serie. La suma de los rangos se hace por separado para los signos positivos y los negativos.  $W$  representa la menor de esas dos sumas. Comparamos  $W$  con el valor proporcionado por las tablas estadísticas al efecto para determinar si rechazamos o no la hipótesis nula, según el nivel de significación elegido.

Propósito:

Compara dos tratamientos a partir de dos muestras relacionadas producidas por estos.

Método:

a) Teoría:

Se tienen  $N$  pares igualados que conforman dos muestras dependientes que se desea comparar. La prueba de los signos analiza sólo el signo de la diferencia entre los elementos de cada par, pero esta prueba tiene en cuenta además la magnitud de esa diferencia.

El procedimiento consiste en hallar la diferencia, eliminar los pares con diferencia cero y luego sustituir esas diferencias por sus rangos (manteniendo el signo). Si no hay diferencia entre los tratamientos debe esperarse que la suma de los rangos positivos sea igual que la de los negativos.

Supóngase que se dispone  $n$  par de observaciones, denominada  $(X_i, Y_i)$ . El objetivo del test es comprobar si puede dictaminarse que los valores  $X_i$  y  $Y_i$  son o no iguales.

1. Si  $d_i = Y_i - X_i$ , entonces los valores  $d_i$  son independientes.
2. Los valores  $d_i$  tienen una misma distribución continua y simétrica respecto a una mediana común  $\theta$ .
3. Las variables  $X_i$  y  $Y_i$  deben ser ordinales.

Si la hipótesis fundamental es cierta, esta suma  $W$  debería ser aproximadamente la mitad de la suma total de los rangos en la muestra completa.

Para verificar la hipótesis, en primer lugar, se ordenan los valores absolutos  $|d_1|, |d_2|, \dots, |d_n|$  y se les asigna su rango  $R_i$ . Entonces, el estadístico de la prueba de los signos de Wilcoxon,  $W$ , es

$$W = \sum R_i$$

Es decir, la suma de los rangos  $R_i$  con el signo menos frecuente ya sean los valores positivos o negativos de  $d_i$ .

Cuando  $N \geq 25$  se utiliza la distribución normal,  $z$ , que viene dada por la expresión:

$$z = \frac{(W - u_z)}{\sigma_z}$$

Donde  $u_z$  y  $\sigma_z$  son la media y la desviación estándar de  $W$  si la hipótesis nula es cierta, y vienen dadas por las siguientes fórmulas:

$$u_z = n(n + 1)/4$$

$$\sigma_z = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}$$

La distribución del estadístico  $W$  puede consultarse en tablas para determinar si se acepta o no la hipótesis nula.

### 2.3.2. Algoritmo de Wilcoxon

Estos son los pasos de la prueba de rangos señalados y pares igualados de Wilcoxon:

1. Para cada para igualado, se determina la diferencia del signo ( $d_i$ ) entre los dos puntajes.
2. Se ordenan estas  $d_i$  sin respetar el signo. Con las  $d$  ligadas, se asigna el promedio de los rangos ligados.
3. Se añade a cada rango el signo (+ ó -) de la  $d$  que representa.
4. Se determina  $W$ , la más pequeña suma de los rangos igualados.
5. Se determina  $N$ , el número total de  $d$  con un signo.
6. El procedimiento para determinar la significación del valor observado de  $T$  depende del lado de  $N$ :
  - a) Si  $N$  es 25 o menor, la tabla  $G$  contiene los valores críticos de  $W$  para diferentes tamaños de  $N$ . Si el valor observado de  $W$  es igual o menor que le dado en la tabla, para un nivel de significación particular y una  $N$  particular,  $H_0$  puede ser rechazada en ese nivel de significación.

- b) Si  $N$  es mayor que 25, se calcula el valor de  $z$  definido por la formula:

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

Se determina su probabilidad asociada conforme a  $H_0$  con la tabla  $A$ . Para una prueba de dos colas se duplica el valor de  $p$ . Si la  $p$  obtenida es igual o menor que  $\alpha$ , se rechaza  $H_0$ .

### Potencia-Eficiencia

Cuando las suposiciones de la prueba paramétrica  $t$  en verdad se satisfacen, la eficiencia asintótica cercana a  $H_0$  de la prueba de rangos señalados y pares igualados de Wilcoxon, comparada con la prueba  $t$  es de  $3/\pi = 95.5\%$ . Esto significa que  $3/\pi$  es la proporción límite de tamaños de muestra necesaria para que las pruebas de Wilcoxon y  $t$  alcancen el mismo poder. Para muestras pequeñas la eficiencia se acerca al 95%. (Siegel, 1988)

## 2.4. El caso de $k$ muestras relacionadas

Para el caso de  $k$  muestras relacionadas se reportan en la literatura las siguientes pruebas estadísticas no paramétricas:

- a) La prueba  $Q$  de Cochran.
- b) Análisis de varianza de Friedman.

La segunda prueba es útil cuando la medida de la variable está al menos en una escala ordinal. Ella determina si las  $k$  muestras relacionadas proceden de la misma población con respecto a las medidas de los rangos.

La prueba de Friedman deberá preferirse a la de Cochran cuando los datos sean adecuados, es decir, siempre que los puntajes estén medidos, por lo menos, en una escala ordinaria. Además, tiene la ventaja de poderse hallar la probabilidad exacta para muestras muy pequeñas mientras que la de Cochran no debe usarse cuando es demasiado pequeña.

La prueba estadística paramétrica para el caso de  $k$  muestras dependientes es la prueba  $F$ , la cual supone:

- a) Que los puntajes u observaciones sean tomados independientemente de poblaciones distribuidas de manera normal.
- b) Que las poblaciones tengan todas la misma varianza, y que las medias en las poblaciones distribuidas normalmente sean poblaciones lineales de “efectos” debido a renglones y columnas.

Requiere, por lo menos, medidas de intervalos de las variables involucradas. (Arsula, 1984)

### **2.4.1. Prueba de Friedman**

La prueba de Friedman es una prueba estadística no paramétrica desarrollada por el economista Milton Friedman. Al igual que la prueba paramétrica de muestras repetidas ANOVA, se utiliza para detectar diferencias en los tratamientos a través de múltiples pruebas. El procedimiento implica el ranqueo de cada fila (o bloques de filas), y se considera también que los valores son ranqueados por columnas.

La prueba de Friedman se utiliza para un solo sentido de medidas repetidas análisis de varianza por rangos. En su uso de los rangos es similar al análisis de varianza por rangos de Kruskal-Wallis.

#### Propósito:

Probar si tres o más muestras pertenecen a la misma población.

#### Método:

##### a) Teoría:

Cuando los datos de  $k$  muestras igualadas están, por lo menos, en escala ordinal, el análisis de varianza de dos clasificaciones por rangos de Friedman es útil para probar la hipótesis de nulidad de que las  $k$  muestras han sido sacada de la misma población.

Puesto que las muestras han sido igualadas, el número de casos es el mismo en cada una de las muestras. La igualación puede hacerse estudiando el mismo grupo de sujeto en cada una de las  $k$  condiciones. O el investigador puede obtener varios conjuntos, compuesto cada uno de los  $k$  sujetos igualados, para asignar al azar un sujeto de cada conjunto a la primera condición, un sujeto de cada conjunto a la segunda condición, etc.

Para la prueba de Friedman los datos se colocan en una tabla de  $n$  filas (sujeto, bloque, etc.) y  $k$  columnas (tratamientos, condiciones, etc.). Las observaciones en diferentes filas son independientes pero las columnas no. Cada fila se reemplaza por sus rangos, así,  $R_{ij}$  es el rango del tratamiento  $j$  en el bloque  $i$ . Entonces  $R_{i1}, R_{i2}, \dots, R_{ik}$  es una permutación de los primeros  $k$  enteros y  $R_{1j}, R_{2j}, \dots, R_{nj}$  es el conjunto de los rangos dados al tratamiento  $j$  en todos los bloques.

Los totales por filas son  $N(N + 1)/2$  pero los totales por columnas son afectados por las diferencias entre tratamientos.

Bajo la hipótesis nula de que los tratamientos son iguales los totales por columnas serán iguales a  $k(N + 1)/2$ .

La prueba de Friedman determina la diferencia significativa de los totales de rango ( $R_j$ ). Para hacer esta prueba se calcula el valor de una estadística que Friedman denota como  $\chi^2$ .

Cuando el número de filas o de columnas o de ambas, no es demasiado pequeño, puede demostrarse (Friedman, 1937) que  $\chi^2$  está distribuida aproximadamente con  $df = k - 1$ :

$$\chi^2 = \frac{12}{Nk(k + 1)} \sum_{j=1}^k R_j^2 - 3N(k + 1)$$

Donde:

$N$ : Es el número de filas.

$k$ : Es el número de columna.

$R_j$ : Suma de rangos en la columna  $j$ .

$\sum_{j=1}^k (R_j)^2$ : Indica sumar los cuadrados de los rangos en todas las  $k$  condiciones.

Se puede realizar una corrección para los valores repetidos dividiendo por:

$$c = 1 - \frac{\sum t}{n(k^3 - k)}$$

Donde  $\sum_t$  indica sumar en todos los grupos de empate de tamaño  $t$ , el valor  $t^3 - 3$ .

## 2.4.2. Algoritmo de Friedman

Estos son los pasos para usar el análisis de la varianza de dos clasificaciones por rangos de Friedman:

1. Se arreglan los puntajes en una tabla de dos clasificaciones de  $k$  columnas (condiciones) y  $N$  hileras (sujetos o grupos).
2. Se ordenan los puntajes de cada hilera de 1 a  $k$ .
3. Se determina la suma de los rangos de cada columna:  $R_j$
4. Se calcula el valor de  $X^2$  con la fórmula:

$$X^2 = \frac{12}{Nk(k+1)} \sum_{j=1}^k R_j^2 - 3N(k+1)$$

5. El método para determinar la probabilidad de ocurrencia conforme a  $H_0$  asociada con el valor observado de  $X^2$  depende de los tamaños de  $N$  y  $k$ :
  - a) La tabla  $N$  contiene las probabilidades exactas asociadas con valores tan grandes como el de una  $X^2$  observada para  $k=3$ ,  $N$  de 2 a 9 y para  $k=4$ ,  $N$  de 2 a 4.
  - b) Para  $N$  o  $k$ , o ambas mayores que las de la tabla  $N$ , la probabilidad asociada puede determinarse por referencia a la distribución chi cuadrada (dada en la tabla C) con  $d_f = k - 1$ .
6. Si la probabilidad obtenida con el método debido es igual o menor que  $\alpha$ , se rechaza  $H_0$ .

## 2.5. Lenguaje de programación PL/SQL

Bajo el nombre de PL/SQL se esconde el Lenguaje de manipulación de datos propietario de Oracle. Conceptualmente, Oracle lo define como una extensión procedimental del SQL, y para entenderlo mejor, se trata de un potente lenguaje de acceso a bases de datos, mediante el cual podemos estructurar y controlar las sentencias SQL que definamos para nuestra base de datos. (Alvarez, 2009)

### ¿Por qué usar PL/SQL y no otro lenguaje de programación?

Por ser además PL/SQL el propio lenguaje de la herramienta Oracle establecida como estándar, sigue la filosofía de los modernos lenguajes de programación, es decir, permite definir y manipular distintos tipos de datos, crear procedimientos, funciones, contempla recursividad, etc. Quizás la diferencia más importante, y algo que debemos tener siempre muy en cuenta, es que la eficiencia de un programa en PL/SQL se mide sobre todo por la eficiencia de los accesos a bases de datos.

La consecuencia más inmediata de lo dicho anteriormente consiste en poder programar de manera óptima en PL/SQL, se debe tener un dominio notable del propio SQL; cumpliendo esta premisa, y algunas otras que veremos más adelante, obtendremos una mejora sustancial en nuestras aplicaciones que interactúen con bases de datos.

### 2.6.1. Integración de PL/SQL con Oracle

Para conectarse con la base de datos solo basta con realizar una consulta, aunque hay que tener en cuenta que el conjunto de filas que devuelve una consulta puede ser 0, 1 o N, y depende de a cuántas filas afecte la condición de búsqueda. Cuando una consulta devuelve múltiples filas, se debe declarar un cursor para procesarlas. Se puede declarar un cursor en la parte de declaraciones de cualquier bloque, subprograma o paquete PL/SQL. También se pueden procesar varias consultas de forma paralela, y declarar y abrir múltiples cursores. (Alvarez, 2009)

A continuación se muestra un ejemplo de unas de las consultas que fueron realizadas sobre el almacén de datos, en este caso para extraer dos muestras relacionadas de una población, es decir, la primera consulta devuelve la cantidad de delitos del mes actual por días mientras que la segunda consulta devuelve la cantidad de delitos también por días pero del mes anterior.

```

Select Ano,Mes,Dia,Count(*) Cantidad
  From Dim_Tiempo p4,Fact_Denun_1 p2
  Where Fechad_Id Between To_number(to_char(add_months(sysdate,-4),'yyyymm')||'01') and
                        To_Number(to_char(add_months(sysdate,-4),'yyyymm')||to_char(last_day(add_months(sysdate,-1)),'dd')) and
                        Dim_id=Fechad_Id and Marca IN (2,6,7,9)

  Group By Ano,Mes,dia
  Order By Ano,Mes,dia;

Select Ano,Mes,Dia,Count(*) Cantidad
  From Dim_Tiempo p4,Fact_Denun_1 p2
  Where Fechad_Id Between To_number(to_char(add_months(sysdate,-5),'yyyymm')||'01') and
                        To_Number(to_char(add_months(sysdate,-5),'yyyymm')||to_char(last_day(add_months(sysdate,-2)),'dd')) and
                        Dim_id=Fechad_Id and Marca IN (2,6,7,9)

  Group By Ano,Mes,dia
  Order By Ano,Mes,dia;
    
```

Figura 2.1 Consultas para extraer la cantidad de delitos

Para realizar todo el trabajo de implementación fue necesario crear dos tablas en el almacén de datos, ellas son **DIM\_NOPARAM\_2GPO** para guardar información con respecto al tipo de análisis, provincia y municipio donde se efectuaron las pruebas no paramétricas; y **VM\_NOPARAM\_2GPO** para guardar otro tipo de información como año, mes, día, cantidad de grupos, ocurrencia del delito, rango promedio y la significación.

Column Name	Data Type	Nullable	Data Default	COLUMN ID	Primary Key	COMMENTS
DIM_ID	NUMBER	Yes	(null)	1	(null) (null)	
ANALISIS	VARCHAR2(60 BYTE)	Yes	(null)	2	(null) (null)	
PROVINCIA	VARCHAR2(30 BYTE)	Yes	(null)	3	(null) (null)	
MUNICIPIO	VARCHAR2(30 BYTE)	Yes	(null)	4	(null) (null)	

Figura 2.2 Definición de columnas en la tabla DIM

Column Name	Data Type	Nullable	Data Default	COLUMN ID	Primary Key	COMMENTS
ANALISIS_ID	NUMBER	Yes	(null)	1	(null)	(null)
ANO	VARCHAR2(4 BYTE)	Yes	(null)	2	(null)	(null)
MES	VARCHAR2(12 BYTE)	Yes	(null)	3	(null)	(null)
DIA	NUMBER	Yes	0	4	(null)	(null)
GRUPO	NUMBER	Yes	(null)	5	(null)	(null)
ORDEN	NUMBER	Yes	(null)	6	(null)	(null)
OCURRENCIA	NUMBER	Yes	(null)	7	(null)	(null)
RANGOMEDIO	NUMBER	Yes	(null)	8	(null)	(null)
SIGNIFICACION	NUMBER	Yes	(null)	9	(null)	(null)

**Figura 2.3** Definición de columnas en la tabla VM

A continuación se muestra como quedan los datos insertados en las tablas anteriormente creadas, que devuelven los métodos no paramétricos después de ser ejecutados.

DIM_ID	ANALISIS	PROVINCIA	MUNICIPIO
1	1 (3 MESES) MANN-WHITNEY	TODAS LAS PROVINCIAS	TODOS LOS MUNICIPIOS
2	2 (1 MES) WILCOXON	TODAS LAS PROVINCIAS	TODOS LOS MUNICIPIOS

**Figura 2.4** Inserción de datos en la tabla DIM

ANALISIS_ID	ANO	MES	DIA	GRUPO	ORDEN	OCURRENCIA	RANGOMEDIO	SIGNIFICACION
1	1 2009	04 (ABR)	4	1	4	28	89,83	0,581
2	1 2009	04 (ABR)	5	1	5	29	89,83	0,581
3	1 2009	04 (ABR)	6	1	6	47	89,83	0,581
4	1 2009	04 (ABR)	7	1	7	40	89,83	0,581
5	1 2009	04 (ABR)	8	1	8	33	89,83	0,581
6	1 2009	04 (ABR)	9	1	9	34	89,83	0,581

**Figura 2.5** Inserción de datos en la tabla VM

Finalmente como se puede ver toda la información queda guardada en tablas, con la cual se puede trabajar mediante el uso de consultas al almacén de datos, es decir, se puede extraer, actualizar y modificar.

## **Consideraciones generales del capítulo**

A partir del conocimiento de las pruebas estadísticas no paramétricas, se muestra a modo de resumen que:

- Se expusieron las distintas pruebas no paramétricas para los distintos casos, su fundamento matemático y su algoritmo.
- Se definió y utilizó el lenguaje de programación PLSQL por ser el lenguaje de manipulación de datos propietario de Oracle, lo que permitió implementar procedimientos almacenados con una integridad referencial declarativa bastante potente; además, sigue la filosofía de cualquier lenguaje moderno de programación.

## CAPÍTULO 3. Aplicación de las pruebas estadísticas

En el siguiente capítulo se expone un manual de usuario como ayuda para usar la herramienta. También se muestran los resultados de los métodos no paramétricos, de manera gráfica de algunos análisis, donde se aplicaron las pruebas estadísticas. Además, estos mismos análisis se ejecutan en el SPSS, utilizado para demostrar la validación de los resultados obtenidos.

### 3.1. Manual de usuario

Para explorar la interfaz de usuario de *Oracle BI Enterprise Edition* es necesario utilizar dos herramientas que brindan una funcionalidad similar: *Oracle BI Answers*, para construir, modificar y guardar las consultas analíticas, y *Oracle BI Dashboards*, que permite navegar mediante los cuadros de mando analíticos.

*BI Answer* resulta una poderosa herramienta de visualización con una interfaz Web. Esta es capaz de conectarse con el servidor de BI y permite al usuario realizar consultas de manera sencilla, simplemente arrastrando elementos. Además, ella posibilita la creación de diferentes tipos de gráficos y tablas según las necesidades del cliente.

*BI Dashboards* permite crear pizarrones virtuales donde los usuarios del sistema pueden ingresar y ver consultas a las que necesitan acceder comúnmente, y con un formato establecido que puede ser un gráfico u otra representación. Todo esto se logra ingresando a una aplicación Web desde cualquier navegador.

A continuación se muestra cómo usar la herramienta *Oracle BI Dashboards*; es esta con la que interactúa el usuario final, ya que el interés de este usuario es solo ver los resultados y no modificarlos.

## ¿Cómo navegar por *Oracle BI Interactive Dashboards*?

En esta sección se proporcionan procedimientos para ver un cuadro de mando, desplazarse a una página del cuadro de mando y utilizar métodos abreviados de teclado en un cuadro de mando.

### Para ver un cuadro de mando

- Inicie sesión en Oracle BI.

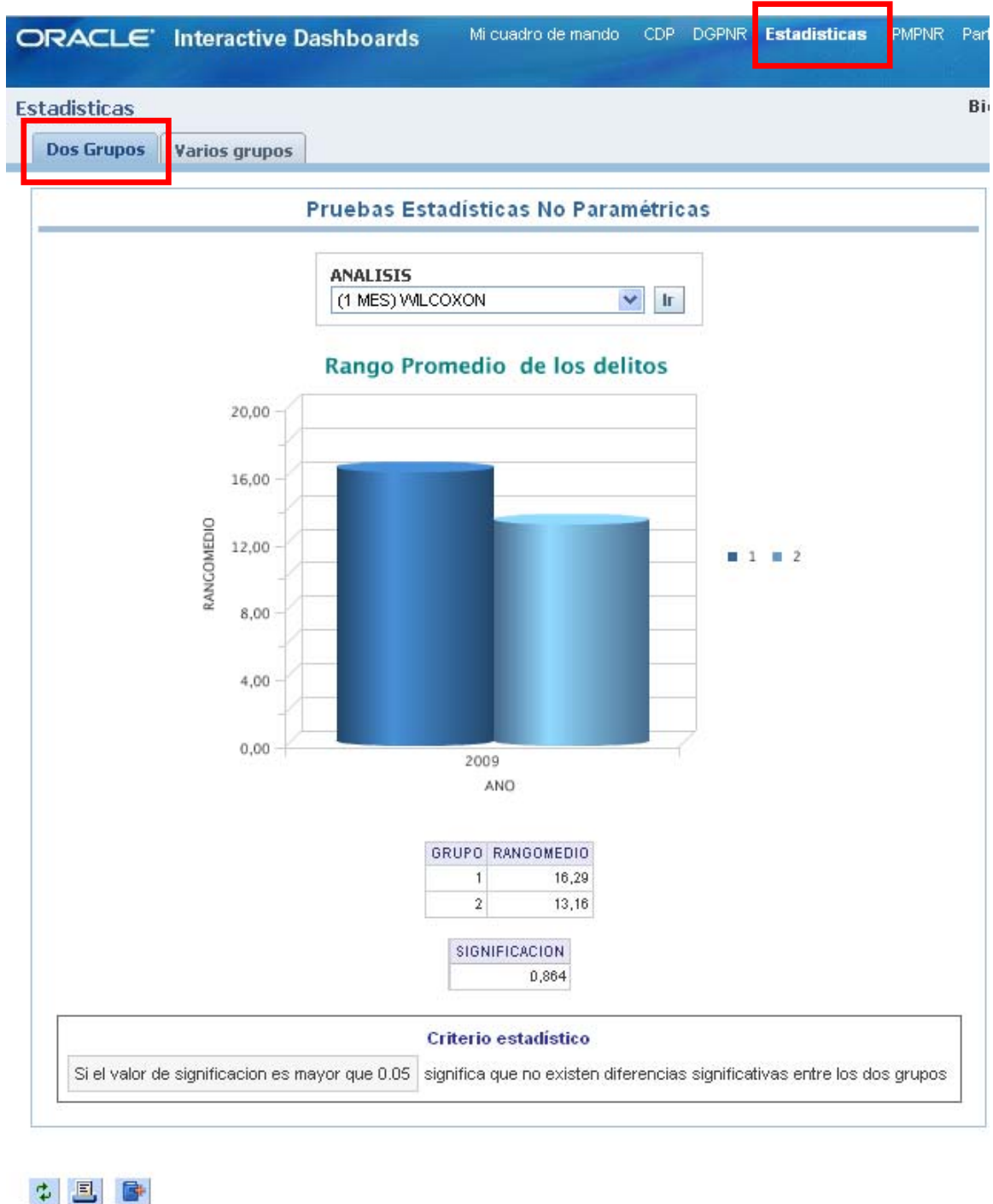
The image shows the Oracle Business Intelligence login interface. At the top, there is a banner with the Oracle logo and the text 'BUSINESS INTELLIGENCE'. Below the banner, the text 'Oracle Business Intelligence' is displayed in a large font. Underneath, there is a prompt: 'Especifique el ID de usuario y la contraseña a continuación y presione el botón Iniciar sesión.' Below this prompt are two input fields: 'ID de usuario' and 'Contraseña'. To the right of the 'ID de usuario' field is a small vertical bar. Below the 'Contraseña' field is a button labeled 'Iniciar sesión'. At the bottom, there is a dropdown menu labeled 'Seleccione un idioma' with 'Español' selected.

**Figura 3.1** Diálogo de autenticación

Después de que el usuario haya sido autenticado se muestra un sitio web como el siguiente, donde se tienen los cuadros de mando en la parte superior y sus distintas fichas de páginas que se muestran en la parte inferior.

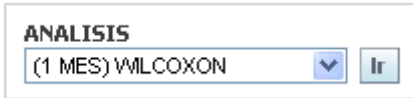
- En la parte superior de la página, haga clic en el nombre de un cuadro de mando para mostrarlo.

Aparecerá el cuadro de mando, en este caso “Estadísticas” y a su vez las distintas fichas de páginas que este posee: “Dos grupos” y “Varios Grupos” como se muestra en la figura 3.2.



**Figura 3.2** Cuadro de mando

Para que el usuario final pueda ver los resultados, debe escoger en un cuadro de selección el análisis que desea y luego dar clic en el botón “Ir”.



**Figura 3.3** Cuadro de selección

### Para desplazarse a una página del cuadro de mando

- En la parte superior del cuadro de mando, haga clic en la ficha de página del cuadro de mando para mostrarla.

Aparecerá el contenido de la página del cuadro de mando. Si no se ha agregado contenido, aparecerá un mensaje de notificación de página vacía del cuadro de mando.

Esta es una página de cuadro de mando vacía.

Haga clic [aquí](#) para añadir contenido.



**Figura 3.4** Página vacía

**Para utilizar métodos abreviados de teclado y desplazarse en un cuadro de mando**

- Utilice los métodos abreviados de teclado que se muestran en la Tabla 3.5 para desplazarse en el cuadro de mando.


Método abreviado	Resultado
CTRL + SHIFT + F2	Accede al primer cuadro de mando que figura en la página. Este cuadro de mando suele denominarse “Mi cuadro de mando”.
CTRL + SHIFT + F3	Accede a la primera página del cuadro de mando actual.
CTRL + SHIFT + I	Accede a la siguiente sección de contenido del cuadro de mando actual. Este método abreviado permite desplazarse primero de izquierda a derecha y después hacia abajo.
CTRL + SHIFT + U	Accede a la sección anterior de contenido del cuadro de mando actual. Este método abreviado permite desplazarse primero de derecha a izquierda y después hacia arriba.
CTRL + SHIFT + F8	Accede al ítem siguiente de una sección de contenido. Si la sección está alineada horizontalmente, este método abreviado permite desplazarse de izquierda a derecha.
CTRL + SHIFT + F7	Accede al ítem anterior de una sección de contenido. Si la sección está alineada horizontalmente, este método abreviado permite desplazarse de derecha a izquierda.

**Figura 3.5** Método abreviado del teclado

**Para imprimir desde cuadro de mando de *Oracle BI***

Puede imprimir en HTML o en un PDF de Adobe (*Document Format*). Se necesita Adobe Reader 6.0 o superior para imprimir en un PDF de Adobe.

**NOTA:** El método HTML de impresión depende de las funciones de manejo de la impresión de cada explorador. Si no obtiene los resultados esperados, seleccione PDF para abrir e imprimir el cuadro de mando o la solicitud.

Para imprimir una página del cuadro de mando, localice y haga clic en el botón “Imprimir”  situado en la parte inferior del cuadro de mando y, a continuación, seleccione HTML o PDF.


- Para HTML, se abre una nueva ventana que muestra el ítem seleccionado sin los enlaces extraños.

Seleccione Archivo > Imprimir, en el menú del explorador.

- Para archivos en formato PDF, utilice las opciones disponibles en la ventana de PDF de Adobe para guardar o imprimir el archivo.

### **Para actualizar desde un cuadro de mando**

Al seleccionar un cuadro de mando, los servicios de presentación de *Oracle BI* comprueban la caché para determinar si los resultados idénticos se han solicitado recientemente. Si ese es el caso, los Servicios de presentación de *Oracle BI* devuelven los resultados más recientes y evitan así que *Oracle BI Server* y la base de datos subyacente realicen procesamientos innecesarios. En caso contrario, la solicitud se envía a *Oracle BI Server* para su procesamiento.

Para actualizar un cuadro de mando, haga clic en el botón “Actualizar” .

Para salir de *Oracle BI*, no cierre la ventana del explorador.

- Desde cualquier pantalla de *Oracle BI*, haga clic en el enlace “Cerrar sesión”.

### 3.2. Resultados de las pruebas estadísticas

Estas pruebas estadísticas están programadas en un paquete estadístico que facilita el trabajo para el investigador. Simplemente, la tarea de decidir por cuál de todos ellos es mediante la consideración que se tengan de los datos, ya que no se requiere de los supuestos acerca de la naturaleza de la población.

Para aplicar cada método correctamente hay que tener en cuenta la hipótesis fundamental y la alternativa. Los datos deben tener un nivel de medición al menos ordinal.

Cada prueba tiene su dominio de aplicación. Se implementaron métodos que comparan dos o más muestras independientes y otros que comparan dos o más momentos en el tiempo. Por eso uno de los objetivos de este trabajo, además de obtener resultados que nos ayudan a la toma de decisiones, es que el usuario aprenda cuándo debe utilizar una prueba estadística en particular.

La variable que se analiza, en este caso, cantidad de delitos, tiene una escala de nivel ordinal.

Los análisis en que se efectuaron las pruebas estadísticas son los siguientes:

- La ocurrencia de delitos diarios entre las provincias de Villa Clara y Cienfuegos comparado el mismo trimestre.
- La ocurrencia de delitos diarios entre las provincias de Villa Clara y Cienfuegos comparado el mismo mes.
- La ocurrencia de delitos diarios entre los municipios de Sagua, Placetas y Caibarién comparando el mismo trimestre.
- La ocurrencia de delitos diarios entre los municipios de Sagua, Placetas y Caibarién comparando el mismo mes.

- La ocurrencia de delitos diarios en un mes comparado con el mes anterior.
- La ocurrencia de delitos diarios en más de dos meses consecutivos.
- La ocurrencia de delitos diarios en un trimestre comparado con otro trimestre del mismo año.
- La ocurrencia de delitos diarios en un mes comparado con el mismo mes del año anterior.
- La ocurrencia de delitos diarios en un año comparado con el año anterior.

El usuario final, que en este caso es un policía, tendrá la posibilidad de seleccionar el método de acuerdo con las consideraciones de los datos. Luego, los resultados se le mostrarán en forma de gráfica para que tenga una idea visual de cómo se comporta la variable cantidad de delitos. También aparece el valor resultante de significación.

El **grado de significación** 'p' o 'significación' es la probabilidad de error al rechazar la hipótesis nula. Cuanto más pequeño sea su valor, más probable será que la hipótesis nula sea falsa.(William H. Press, 2002)

El grado de significación está relacionado con el **nivel de significación**, el riesgo de error que se está dispuesto a asumir en caso de rechazar la hipótesis nula. En el nivel de significación se establece un valor *a priori* que generalmente suele ser 0,01 ó 0,05. El grado de significación se calcula *a posteriori*, es decir cuando se conoce el resultado de haber aplicado una prueba de significación. El grado de significación indica la probabilidad de error calculada al rechazar la hipótesis nula.

A nivel práctico la forma de actuar es la siguiente:

Si  $p \geq 0,05$  se acepta la hipótesis nula.

Si  $p < 0,05$  se rechaza la hipótesis nula

El rechazo o aceptación de una hipótesis nula se basa sobre algún nivel de significación como criterio. Una diferencia se denomina significativa cuando la distancia entre dos medias muestrales señala una diferencia verdadera entre los parámetros de las poblaciones de las que se sacaron las muestras. Por lo cual, el nivel de significación  $\alpha$  se define como la probabilidad de rechazar erróneamente la hipótesis nula.

Los niveles de significación indican la probabilidad de que un resultado se deba al azar. El nivel más frecuente, que se utiliza para indicar que algo es digno de credibilidad, es 0,95. Esto significa que el hallazgo tiene un 95% de probabilidades de ser cierto. Sin embargo, este valor también se utiliza de manera confusa. Ningún paquete de estadísticas mostrará "95 %" o "0,95" para indicar este nivel. En su lugar, aparecerá "0,05", para indicar que el hallazgo tiene un 5 % (0.05) de probabilidades de no ser cierto, que es lo inverso a un 95 % de probabilidades de ser cierto. Para obtener el nivel de significación, se debe restar el número que aparece a uno. Por ejemplo, un valor de "0,01" significa que existe un 99 % ( $1-0,01=0,99$ ) de probabilidades de que sea cierto.

$H_0$  (hipótesis nula) = No hay diferencia entre dos variables.

$H_a$  (hipótesis alternativa) = Sí existe diferencia.

## **Resultados de la prueba U de Mann-Whitney**

### Ejemplo 1:

A continuación se muestran los resultados que se obtuvieron después de aplicar la prueba U de Mann-Whitney, para evaluar si existen diferencias significativas en la ocurrencia del delito; teniendo en cuenta las provincias de Villa Clara y Cienfuegos comparando el mismo trimestre, donde la cantidad de delitos será analizada por días. Además, para este

caso de comparaciones verticales o entre dos grupos independientes esta prueba ofrece los resultados mejores.



**Figura 3.6** Resultados de la prueba U de Mann-Whitney

Los resultados demuestran que el grado de significación 0,000 es menor que el nivel 0,05, por lo que podemos concluir que las ocurrencias del delito difieren cualitativamente entre los dos grupos. También podemos ver cómo las diferencias entre los valores del rango medio difieren. En caso que ocurra lo contrario, es decir, que el grado de significación sea mayor que el nivel 0,05, entonces no existen diferencias significativas.

Ejemplo 2:

A continuación se explica otra aplicación de la misma prueba. Se exponen los resultados que se obtuvieron después de aplicar la prueba U de Mann-Whitney para evaluar si existen diferencias significativas en la ocurrencia del delito; teniendo en cuenta un mes entre las provincias de Villa Clara y Cienfuegos, donde la cantidad de delitos será analizada por días.



**Figura 3.7** Resultados de la prueba U de Mann-Whitney

Los resultados demuestran que el valor 0,000 es altamente significativo ya que es menor que el nivel 0,05 por lo que podemos concluir que las ocurrencias del delito difieren cualitativamente entre los dos grupos. También podemos ver cómo las diferencias entre los valores del rango medio difieren.

## Resultados de la prueba de Kruskal-Wallis

### Ejemplo 1:

Seguidamente se presentan los resultados al aplicar la prueba de Kruskal-Wallis, para determinar si existen diferencias significativas en la ocurrencia de los delitos enmarcados en los municipios de Sagua, Placetas y Caibarién comparando el mismo trimestre, lo que arroja los siguientes resultados:



**Figura 3.10** Prueba de Kruskal-Wallis

Se puede observar que no existen diferencias significativas entre la ocurrencia de los delitos, ya que el valor resultante 0,189 es mayor que el nivel 0,05, por tanto se acepta la hipótesis fundamental y se rechaza la alternativa (existen diferencias).

Ejemplo 2:

Se presentan los resultados al aplicar la prueba de Kruskal-Wallis para determinar si existen diferencias significativas en la ocurrencia de los delitos enmarcados en los municipios de Sagua, Placetas y Caibarién comparando el mes de diciembre, lo que arroja los siguientes resultados:



**Figura 3.11** Prueba de Kruskal-Wallis

Se puede observar que existen diferencias significativas entre la ocurrencia de los delitos, ya que el valor resultante 0,035 es altamente significativo.

## Resultados de la prueba de Wilcoxon

### Ejemplo 1:

A continuación se muestran los resultados de la prueba de Wilcoxon para evaluar si existen diferencias significativas en la evolución de la ocurrencia diaria del delito, teniendo en cuenta mes actual con mes anterior. En este caso el objetivo del investigador sí es realizar una prueba de antes y después. La prueba de Wilcoxon ofrece los resultados mejores.



**Figura 3.8** Resultados de la prueba de Wilcoxon

Los resultados demuestran que el grado de significación 0,864 es mayor que el nivel 0,05, por lo que podemos concluir que las ocurrencias del delito no difieren cualitativamente entre los dos grupos.

Ejemplo 2:

A continuación se muestra otro ejemplo. Se muestran los resultados de la prueba de Wilcoxon para evaluar si existen diferencias significativas en la ocurrencia del delito; teniendo en cuenta el mismo mes en años distintos, donde la cantidad de delitos será analizada por días.



**Figura 3.9** Resultados de la prueba de Wilcoxon

Los resultados demuestran que el grado de significación 0,357 es mayor que el nivel 0,05, por lo que podemos concluir que las ocurrencias del delito no difieren cualitativamente entre los dos grupos.

## Resultados de la prueba de Friedman

### Ejemplo 1:

Se muestran los resultados al aplicar la prueba de Friedman para analizar si existen diferencias entre la ocurrencia de los delitos entre los meses de octubre contra los dos meses anteriores (septiembre y agosto) de 2009.



**Figura 3.12** Prueba de Friedman

Como se puede apreciar, existen diferencias significativas entre la ocurrencia de los delitos, ya que el valor 0,003 es altamente significativo, por tanto se rechaza la hipótesis

fundamental y se acepta la alternativa (existen diferencias), y además los rangos medios difieren.

Ejemplo 2:

Se muestran los resultados al aplicar la prueba de Friedman para analizar si existen diferencias entre la ocurrencia de los delitos entre el mismo mes en los distintos años 2009, 2008 y 2007.



**Figura 3.13** Prueba de Friedman

Como se puede apreciar, no existen diferencias significativas entre la ocurrencia de los delitos, ya que el grado de significación 0,671 es mayor que el valor 0,05, por tanto se acepta la hipótesis fundamental.

### 3.3. Validación utilizando el SPSS

El paquete estadístico SPSS (*Statistical Package for the Social Science*) es una herramienta muy confiable y profesional para resolver problemas de pequeña o gran envergadura. Se utilizó el SPSS 15 para Windows, el cual proporciona un poderoso sistema de análisis estadístico utilizando menús descriptivos y cuadros de diálogo sencillos que realizan la mayor parte del trabajo. La mayoría de las tareas se pueden llevar a cabo simplemente situando el puntero del ratón en el lugar deseado y pulsando en el botón “Aceptar”.

A continuación se muestran los pasos necesarios para obtener con este paquete los análisis estadísticos empleados en el trabajo. Ello sigue el doble propósito de servir de validación para los métodos implementados, además de constituir una guía para usuarios no expertos en el tema.

#### Resultados de la prueba U de Mann-Whitney en el SPSS

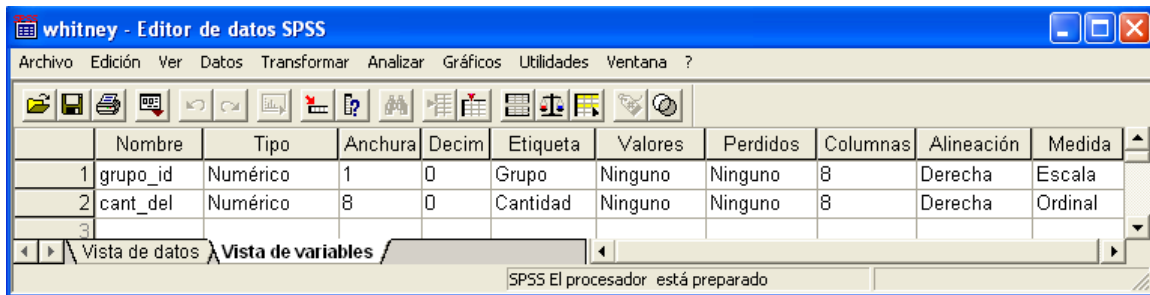
Se ejecutó esta prueba para los distintos análisis:

- La ocurrencia de delitos diarios entre las provincias de Villa Clara y Cienfuegos comparado el mismo trimestre.
- La ocurrencia de delitos diarios entre las provincias de Villa Clara y Cienfuegos comparado el mismo mes.

Construyendo el juego de datos para el SPSS.

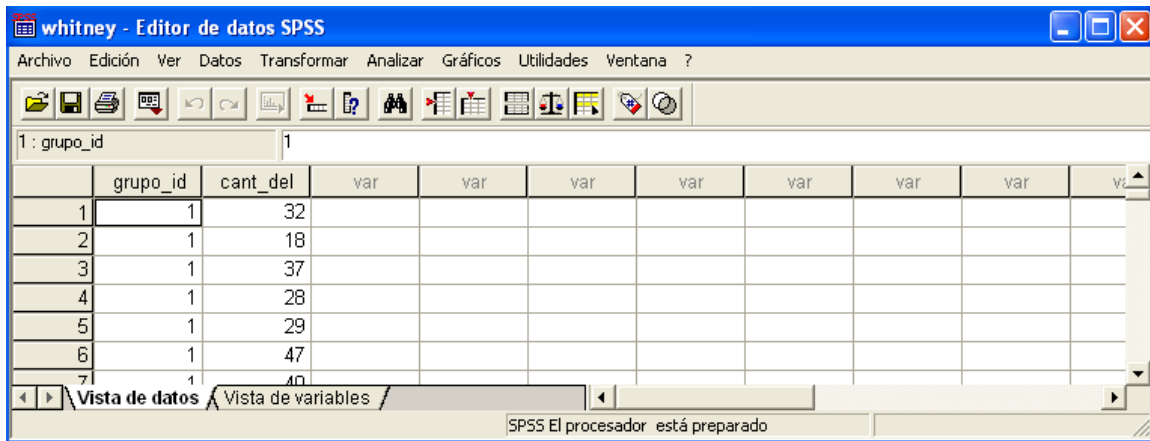
Se definieron 2 variables, todas numéricas: **grupo\_id** (grupo al que pertenece) y **cant\_del** (cantidad de delitos).

Definición de variables.



**Figura 3.5** Definición de variables

Entrada de datos.



**Figura 3.6** Entrada de datos

Prueba M-W para ver si existen diferencias entre dos muestras independientes.

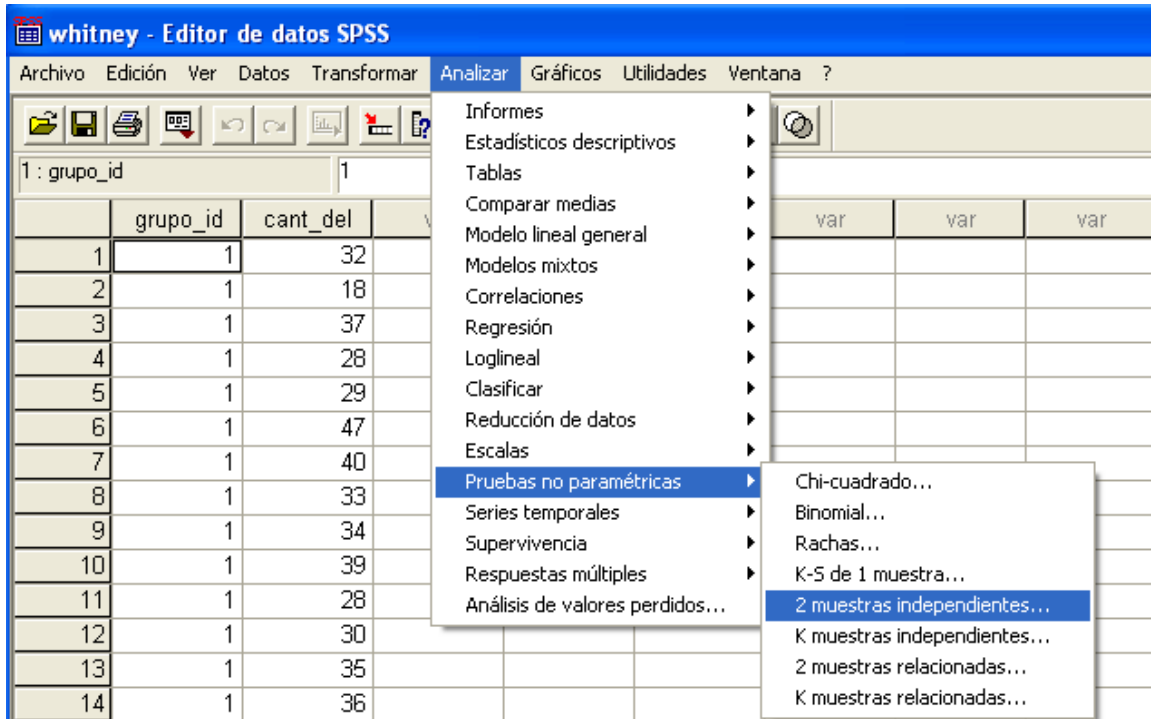


Figura 3.7 Dos muestras independientes

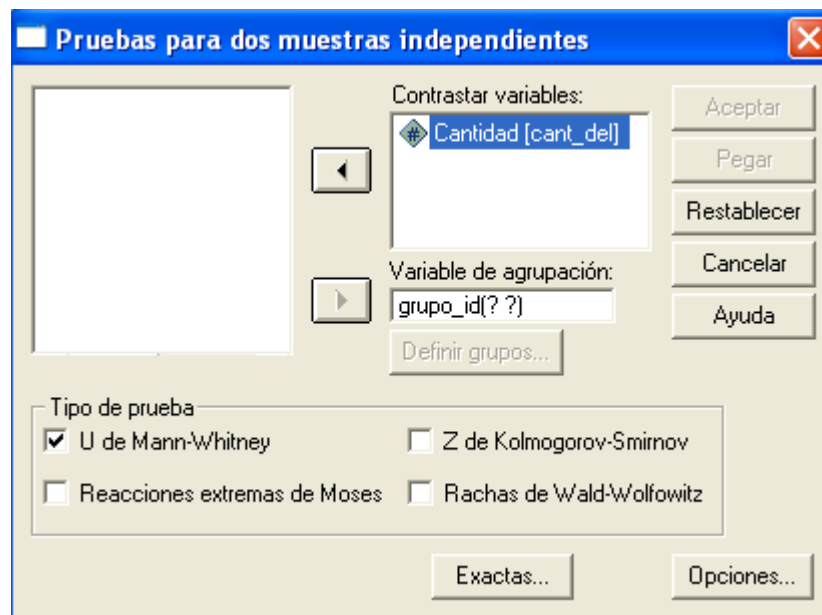
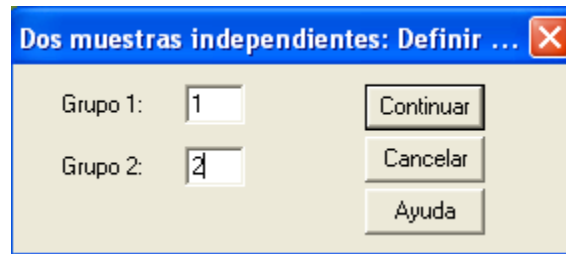


Figura 3.8 Contrastar variables

Se define el intervalo de los grupos que existen en este caso 1 a 2.



**Figura 3.9** Definición de grupos

El resultado comparando dos trimestres entre las provincias de Villa Clara y Cienfuegos fue el siguiente:

### Prueba de Mann-Whitney

#### Rangos

	grupo_id	N	Rango promedio	Suma de rangos
CANT_DEL	1	92	129,37	11902,00
	2	92	55,63	5118,00
	Total	184		

#### Estadísticos de contraste<sup>a</sup>

	CANT_DEL
U de Mann-Whitney	840,000
W de Wilcoxon	5118,000
Z	-9,406
Sig. asintót. (bilateral)	,000

a. Variable de agrupación: grupo\_id

**Figura 3.10** Resultados de la prueba U de Mann-Whitney

La salida del SPSS refleja como “estadística descriptiva” el rango medio de cada grupo, el valor del estadístico U de Mann-Whitney, el valor de la suma de rangos W de Wilcoxon y en el caso de muestras grandes como esta, el valor de la Z común y su

significación. En caso de muestras pequeñas se conforma el test exacto y se da la significación también. En fin, atendiendo a que la significación de este test es menor que 0,05, concluimos que las evaluaciones de la prueba inicial difieren cualitativamente entre los dos grupos.

El resultado, comparando la ocurrencia de los delitos en un mes entre las provincias de Villa Clara y Cienfuegos, fue el siguiente:

### Prueba de Mann-Whitney

**Rangos**

	grupo_id	N	Rango promedio	Suma de rangos
cant_del	1	31	43,77	1357,00
	2	30	17,80	534,00
	Total	61		

**Estadísticos de contraste<sup>a</sup>**

	cant_del
U de Mann-Whitney	69,000
W de Wilcoxon	534,000
Z	-5,717
Sig. asintót. (bilateral)	,000

a. Variable de agrupación: grupo\_id

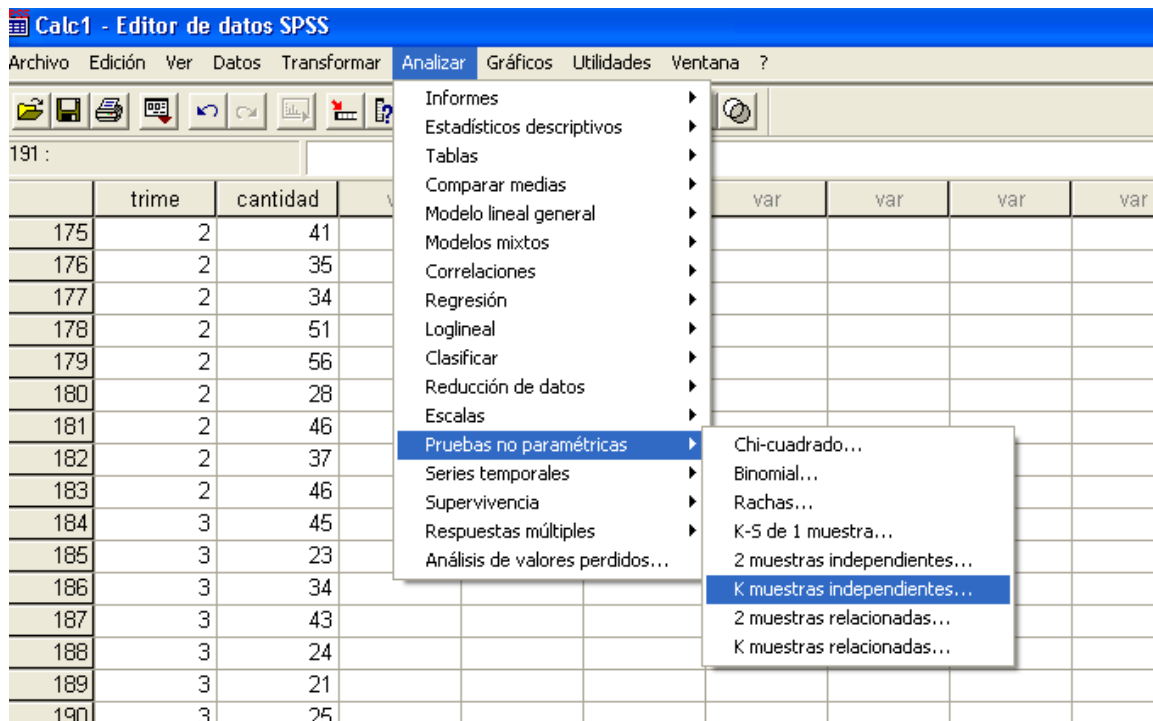
**Figura 3.10** Resultados de la prueba U de Mann-Whitney

Atendiendo a que la significación de este test es, no solo menor que 0,05, sino altamente significativo, concluimos que las evaluaciones de la prueba inicial difieren cualitativamente entre los dos grupos.

## Resultados de la prueba de Kruskal-Wallis en el SPSS

Se ejecutó esta prueba para los distintos análisis:

- La ocurrencia de delitos diarios entre los municipios de Sagua, Placetas y Caibarien comparando el mismo trimestre.
- La ocurrencia de delitos diarios entre los municipios de Sagua, Placetas y Caibarien comparando el mismo mes.



**Figura 3.16** K muestras independientes

El resultado de comparar la ocurrencia del delito del mismo trimestre entre Sagua, Placetas y Caibarien fue:

### Prueba de Kruskal-Wallis

**Rangos**

	TRIME	N	Rango promedio
CANTIDAD	1	73	124,83
	2	85	114,65
	3	71	105,32
	Total	229	

**Estadísticos de contraste<sup>a,b</sup>**

	CANTIDAD
Chi-cuadrado	3,328
gl	2
Sig. asintót.	,189

a. Prueba de Kruskal-Wallis  
b. Variable de agrupación: TRIME

**Figura 3.17** Resultados de la prueba de Kruskal-Wallis

La salida del SPSS al aplicar la prueba de Kruskal-Wallis refleja el rango medio de cada grupo, se arrojó un grado de significación de  $0,189 > 0,05$ , por lo que podemos llegar a la conclusión de que no existen diferencias significativas entre la ocurrencia de los delitos en los municipios analizados.

El resultado de comparar la ocurrencia del delito entre los municipios de Sagua, Placetas y Caibarien en el mes de diciembre fue:

### Prueba de Kruskal-Wallis

Rangos			
	GRUPO	N	Rango promedio
CANTIDAD	1	25	46,74
	2	29	40,67
	3	24	30,54
	Total	78	

Estadísticos de contraste <sup>a,b</sup>	
	CANTIDAD
Chi-cuadrado	6,727
gl	2
Sig. asintót.	,035

a. Prueba de Kruskal-Wallis  
b. Variable de agrupación: GRUPO

**Figura 3.18** Resultados de la prueba de Kruskal-Wallis

Arrojando un grado de significación de  $0.035 < 0.05$  podemos llegar a la conclusión de que existen diferencias significativas entre la ocurrencia de los delitos en los municipios analizados.

### Resultados de la prueba de Wilcoxon en el SPSS

La prueba de Wilcoxon se ejecutó sobre los siguientes análisis:

- La ocurrencia de delitos diarios en un mes comparado con el mes anterior.
- La ocurrencia de delitos diarios en un mes comparado con el mismo mes pero del año anterior.

Se definieron 2 variables, todas numéricas: **cant\_gr1** (cantidad de delitos de la muestra 1), **cant\_gr2** (cantidad de delitos de la muestra 2).



Figura 3.11 Definición de variables

Prueba W para ver si existen diferencias entre dos muestras relacionadas.

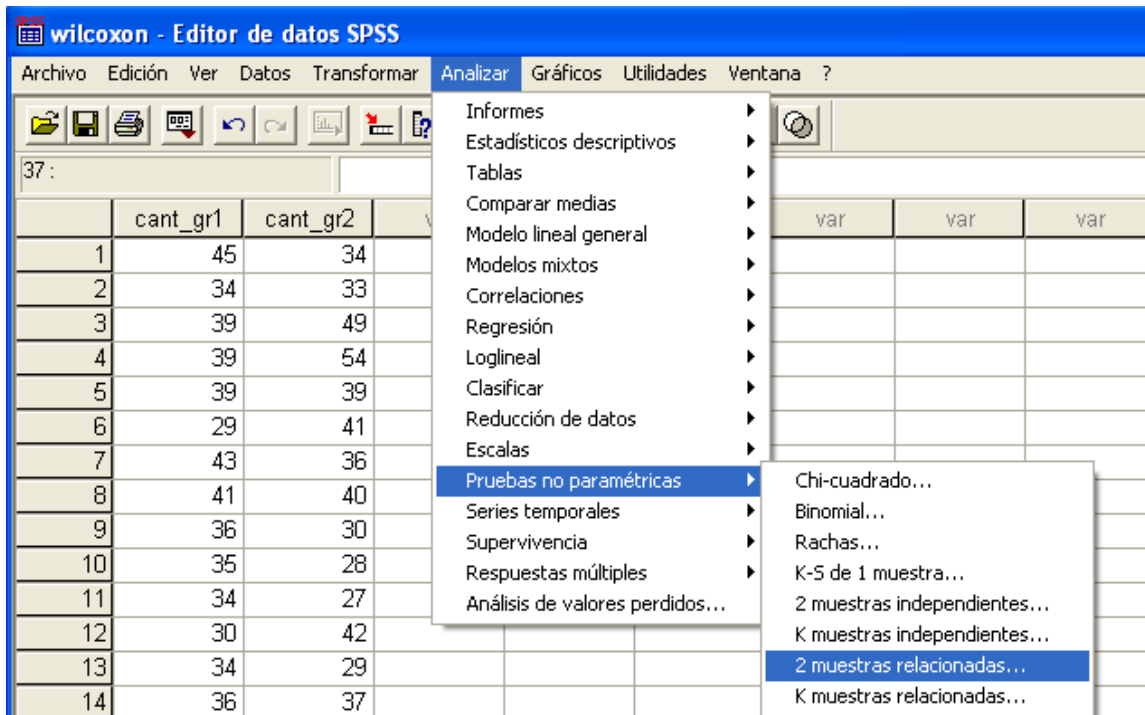


Figura 3.12 Dos muestras relacionadas

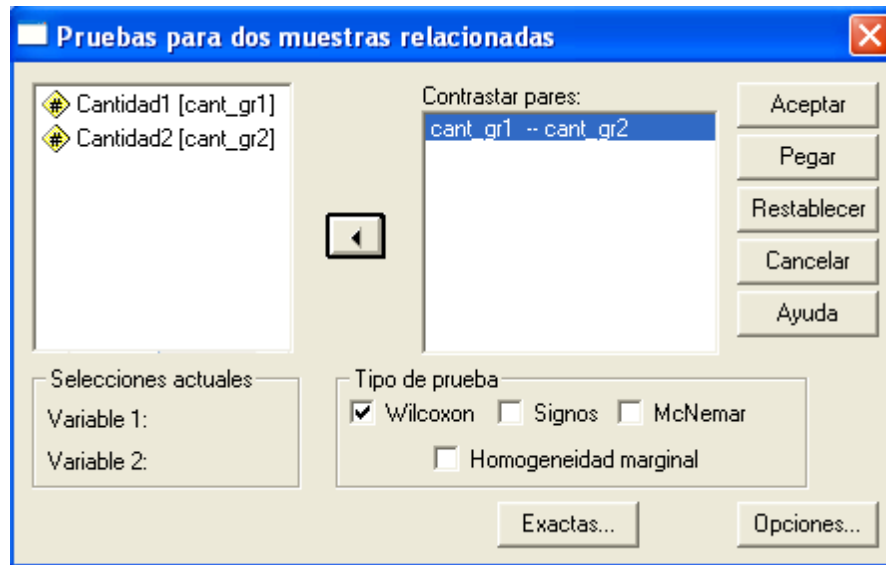


Figura 3.13 Contrastar pares

El resultado de comparar la ocurrencia del delito de un mes con el anterior fue:

### Prueba de los rangos con signo de Wilcoxon

**Rangos**

		N	Rango promedio	Suma de rangos
Cantidad2 - Cantidad1	Rangos negativos	16 <sup>a</sup>	13,16	210,50
	Rangos positivos	12 <sup>b</sup>	16,29	195,50
	Empates	2 <sup>c</sup>		
	Total	30		

- a. Cantidad2 < Cantidad1
- b. Cantidad2 > Cantidad1
- c. Cantidad2 = Cantidad1

**Estadísticos de contraste<sup>b</sup>**

	Cantidad2 - Cantidad1
Z	-1,71 <sup>a</sup>
Sig. asintót. (bilateral)	,864

- a. Basado en los rangos positivos.
- b. Prueba de los rangos con signo de Wilcoxon

Figura 3.14 Resultados de la prueba Wilcoxon

La salida del SPSS refleja como “estadística descriptiva” el rango medio de cada grupo, en ella obtenemos que de una muestra de 30 casos, 16 de ellos tienen una cantidad de delitos menor que la que presentaban en el mes actual, 12 aumentaron en el mes anterior y hubo 2 empates. En el caso de muestras grandes como esta, el valor de la Z común y su significación. En fin, atendiendo a que la significación de este test es mayor que 0,05, concluimos que las evaluaciones de la prueba inicial no difieren cualitativamente entre los dos grupos.

El resultado de comparar la ocurrencia del delito de un mes en dos años fue:

### Prueba de los rangos con signo de Wilcoxon

		Rangos		
		N	Rango promedio	Suma de rangos
grupo2 - grupo1	Rangos negativos	17 <sup>a</sup>	17,35	295,00
	Rangos positivos	14 <sup>b</sup>	14,36	201,00
	Empates	0 <sup>c</sup>		
	Total	31		

a. grupo2 < grupo1

b. grupo2 > grupo1

c. grupo2 = grupo1

#### Estadísticos de contraste<sup>b</sup>

	grupo2 - grupo1
Z	-.922 <sup>a</sup>
Sig. asintót. (bilateral)	.357

a. Basado en los rangos positivos.

b. Prueba de los rangos con signo de Wilcoxon

**Figura 3.15** Resultados de la prueba Wilcoxon

Atendiendo a que la significación de este test es mayor que 0.05, concluimos que las evaluaciones de la prueba inicial no difieren cualitativamente entre los dos grupos.

## Resultados de la prueba de Friedman en el SPSS

La prueba de Friedman se ejecutó sobre los siguientes análisis:

- La ocurrencia de delitos diarios en tres meses consecutivos.
- La ocurrencia de delitos diarios en el mismo mes, pero de tres años consecutivos.

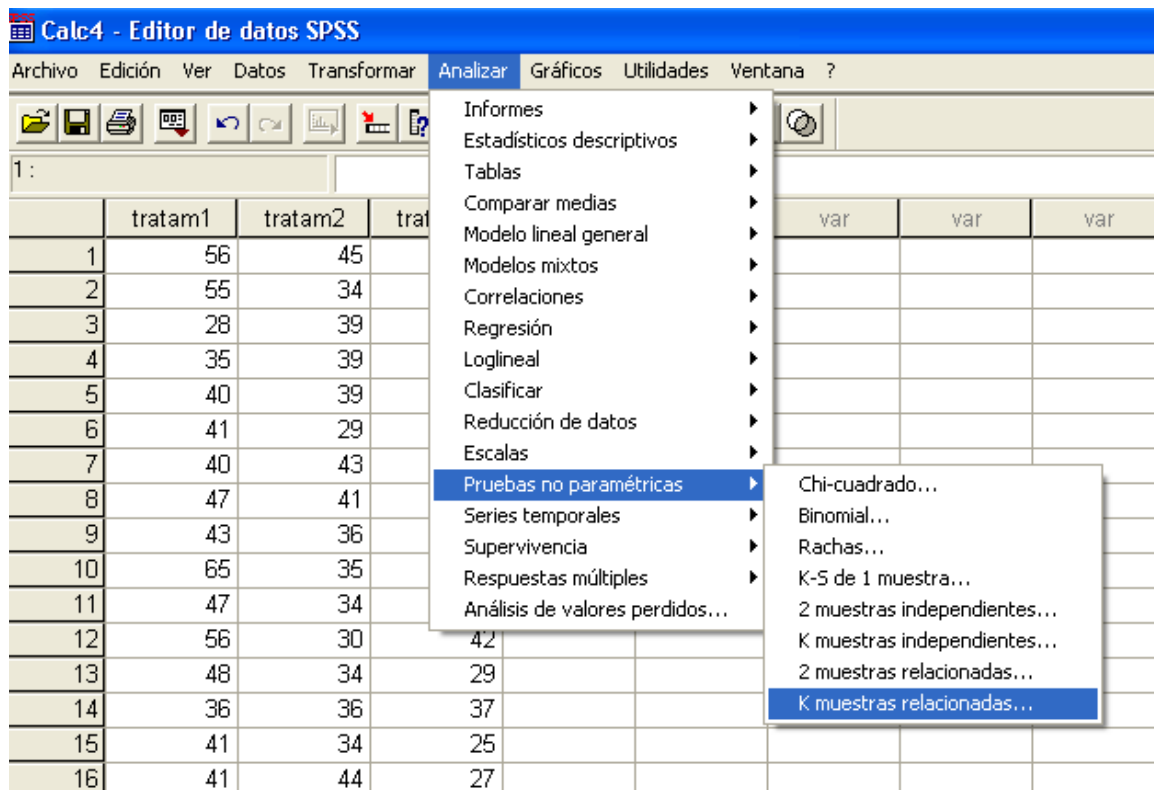


Figura 3.19 k muestras relacionadas

El resultado de comparar la ocurrencia del delito en tres meses consecutivos fue:

## Prueba de Friedman

	Rango promedio
TRATAM1	2,48
TRATAM2	1,83
TRATAM3	1,68

N	30
Chi-cuadrado	11,322
gl	2
Sig. asintót.	,003

a. Prueba de Friedman

**Figura 3.20** Resultados de la prueba de Friedman

La salida del SPSS al aplicar la prueba de Friedman refleja el rango medio de cada tratamiento, en esta obtenemos que existen diferencias significativas, ya que el grado de significación  $0,003 < 0,05$ .

El resultado de comparar la ocurrencia del delito en el mismo mes pero de tres años consecutivos fue:

## Prueba de Friedman

	Rango promedio
TRATAM1	2,11
TRATAM2	2,00
TRATAM3	1,89

N	31
Chi-cuadrado	,797
gl	2
Sig. asintót.	,671

a. Prueba de Friedman

**Figura 3.21** Resultados de la prueba de Friedman

No existen diferencias significativas ya que el grado de significación 0,671 es mayor que el nivel 0,05.

### Consideraciones generales del capítulo

A partir de aplicar las pruebas estadísticas no paramétricas a distintos análisis desarrollados en este capítulo podemos concluir:

- Se realizó un manual de ayuda que permite el uso fácil y adecuado de la herramienta.
- Se aplicaron las pruebas estadísticas no paramétricas adecuadas para cada tipo de análisis y se obtuvieron resultados fiables para la toma de decisiones.
- Se utilizó el SPSS como paquete estadístico complementario del análisis realizado, lo que permitió la validación de la implementación de las pruebas estadísticas.

## CONCLUSIONES

Como resultado de esta investigación se desarrolló la implementación de pruebas estadísticas no paramétricas empleando el lenguaje PL/SQL de bases de datos propietario de Oracle, lo cual permitió ayudar a evaluar el comportamiento del delito de forma confiable.

Este trabajo ha servido de punto de partida para realizar investigaciones estadísticas en la evaluación del delito, ya que:

- Se realizó un estudio profundo del almacén de datos DW\_SAJO, donde se encuentran almacenados todos los datos relacionados con el comportamiento del delito.
- Se implementaron cuatro pruebas no paramétricas en un paquete estadístico mediante la herramienta *Oracle SQL Developer*: las pruebas U de Mann-Whitney, Wilcoxon, Kruskal-Wallis y Friedman.
- Se efectuaron las pruebas estadísticas no paramétricas sobre los análisis requeridos por parte del Minint.
- Quedaron graficados los resultados de estas pruebas estadísticas, aplicadas sobre los distintos análisis en una interfaz web utilizando la herramienta *Oracle Business Intelligence*.
- Se realizó la validación de los resultados de las pruebas no paramétricas utilizando la herramienta SPSS.
- Se realizó un manual de ayuda, para facilitarle el uso fácil de la herramienta *Oracle Business Intelligence Dashboards* al usuario final.

## **RECOMENDACIONES**

- Efectuar estas pruebas estadísticas sobre otros tipos de análisis, ya sea por tipos de delitos, provincias, municipios, etc.
- Realizar otras técnicas de estadística como son las series de tiempo, técnicas de regresión, entre otras, para pronosticar la ocurrencia de los delitos.

## BIBLIOGRAFÍA

- ALVAREZ, S. 2009. *Tutorial de Oracle, Introduccion a PL/SQL* [Online]. Available: [www.DesarrolloWeb.com](http://www.DesarrolloWeb.com) [Accessed].
- ARSULA, J. A. M. 1984. *"Paquete de programas de pruebas no parametricas"*. Universidad Central de las Villas.
- CORDER, G. W. A. D. I. F. 2009. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*, New Jersey, Wiley.
- DAVID PLANEAUX, A. D. 2007. Oracle Business Intelligence Standard Edition One Available: [www.oracle.com](http://www.oracle.com).
- ELLIOTT, T. 2004. Implementing Business Intelligence Standards.
- GANCZARSKI, J. 2009. Data Warehouse Implementations: Critical Implementation Factors Study.
- MANN, H. B., D. R. WHITNEY. 2007. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Available: <http://www.jstor.org>.
- INMON, W. H. 2005. *Building the Data Warehouse*, Indianapolis, Wiley Publishing, Inc.
- INTERIOR, Ó. D. I. D. M. D. 2004. Manual de Procesamiento Penal.
- J.SHESKIN, D. 2000. Handbook of Parametric and NonParametric Statistical Procedures. 2da Edition ed. Washington, D.C, New York: Chapman & Hall / CRC.
- LUMPKIN, G. 2007. Oracle11g para Data Warehousing e Inteligencia de Negocios.
- MARTÍNEZ, J. 2002. ¿Qué es la Inteligencia de Negocios y que beneficios ofrece?
- SIEGEL, S. A, N. J. C. J. 1988. *Nonparametric statistics for the behavioral sciences*, New York, McGraw-Hill, Book Company.
- SPIEGEL, M., R., JOHN J.SCHILLER AND ALU SRINIVASAN 2009. *Probability and Statistics*, , New York, McGraw-Hill, Companies Inc.
- SPSS 2007. SPSS, Inc.
- VALLEJOS, S. J. 2006. *"Minería de Datos"*. Universidad Nacional del Nordeste.
- WALLIS, W. H. K. W. A. 2007. Use of Ranks in One-Criterion Variance Analysis. Available: <http://www.jstor.org>.

WESKE, M. 2007. *Business Process Management: Concepts, Languages, Architectures* Springer.

WILLIAM H. PRESS, S. A. T., WILLIAM T. VETTERLING, BRIAN P. FLANNERY  
2002. *Numerical Recipes in C, The Art of Scientific Computing*. Second Edition  
ed. New York

WITTSCHEN, L. 2004. *Why Business Intelligence?* [Online]. [Accessed].

YGLESIAS, R. 2008. Oracle vs Oracle.

YORTY RUIZ HERNÁNDEZ, E. L. R. 2008. *"Descubrimiento de Conocimiento en Bases de Datos Delictivas"*. UCLV.