ESCUELA SUPERIOR POLITECNICA DEL LITORAL

Instituto de Ciencias Matemáticas

Ingeniería en Estadística Informática

"Análisis Estadístico de los sistemas de información gerencial"

TESIS DE GRADO

Previo a la obtención del Título de

INGENIERO EN ESTADISTICA INFORMATICA

Presentada por:

David *Gonzalo* Vera *Alcívar*

GUAYAQUIL - ECUADOR

AÑO 2001

•

AGRADECIMIENTO

A mis amigos.

A mi director de tesis, Ing. Guillermo Baquerizo, por su excesiva paciencia y confianza.

A todas las personas de una u otra forma hicieron posible la culminación de mi carrera.

DEDICATORIA

A Dios.

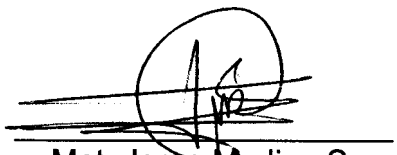
A mis padres.

A mis hermanos.

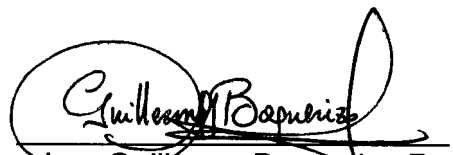
A mi familia, especialmente a Valentín Gonzalo que desde arriba me ayudó bastante.

A mi Sheccid, mi fuerza motivadora

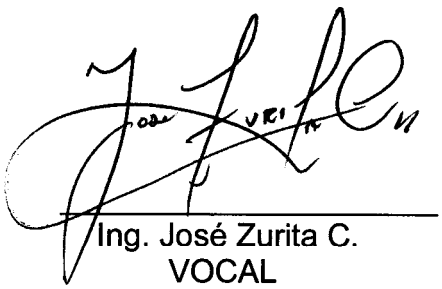
TRIBUNAL DE GRADUACIÓN



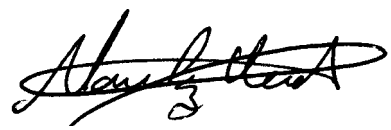
Mat. Jorge Medina S.
DIRECTOR DEL ICM



Ing. Guillermo Baquerizo P.
DIRECTOR DE TESIS



Ing. José Zurita C.
VOCAL



Ing. María de la Paz Vera B.
VOCAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, me corresponden exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL)

David G. Vera Alcívar

RESUMEN

Este estudio trata de medir y explicar el uso actual que se está dando a los sistemas de información gerencia1 en el Ecuador mediante un cuestionario realizado a varias empresas del sector, como herramientas administrativas que, por medio de la informática, son el complemento y soporte a la gerencia de las empresas.

Debido a limitaciones en la logística y a la negativa de atención en varios tipos de empresas, nuestro marco muestral serán todas las empresas de Guayaquil que laboren en el sector industrial de los servicios, un sector que ha tenido un crecimiento bastante alto en los últimos tiempos.

Se empezará resumiendo ciertos conceptos de sistemas de información gerencia1 (SIG), conceptos básicos de estadística y muestreo, determinación del tamaño de la muestra, análisis univariado de las respuestas, para finalizar con el análisis multvariado de las mismas, y las conclusiones y recomendaciones adecuadas que arroje el estudio.

La técnica multivariada utilizada fue el análisis de correspondencias múltiples (ACM), la cual, si bien es relativamente nueva en su desarrollo, tiene una capacidad muy alta para analizar preguntas de carácter cualitativo, como fueron las de nuestro cuestionario.

INDICE GENERAL

INDICE DE GRAFICOS.

INDICE DE TABLAS

ABREVIATURAS

S.I.G.	Sistema(s) de información gerencia1
M.I.S.	Management information systems.
v.a.	Variable aleatoria

SIMBOLOGIAS

μ	Media de la población
σ^2	Varianza de la población
σ	Desviación estándar de la población.
$\bar{X}$	Media de la muestra.
s	Desviación estándar de la muestra
s^2	Varianza de la muestra
θ	Parámetro poblacional
ε	Error de diseño de la muestra

ESPOL
CIB ESPOL

CAPÍTULO 1

1. CONCEPTOS GENERALES SOBRE SISTEMAS DE INFORMACIÓN

Un sistema de información gerencial (SIG), o su equivalente en inglés "management information system (MIS)" es un término general para los sistemas de computadoras en una organización que proveen información sobre sus operaciones de negocios. Es también usado para referirse a las personas que administran estos sistemas. Típicamente, en una gran corporación, el "sistema de información gerencial" o el "departamento de sistema de información gerencial" se refiere a una sistema central, coordinado por expertos en computación y administración, que comúnmente incluyen grandes computadoras (mainframes) pero también incluyen por extensión la red entera de recursos computacionales de la corporación.

incluyen grandes computadoras (mainframes) pero también incluyen por extensión la red entera de recursos computacionales de la corporación.

En un comienzo, las computadoras en los negocios eran usadas para calcular los roles de pagos y llevar control de las cuentas contables de dicho negocio. Al tiempo que las aplicaciones fueron desarrollándose, fueron proveyendo a los administradores con información sobre las ventas, inventarios y otra información que podían ayudar a administrar la empresa, así también el término "SIG" evolucionó para describir estas clases de aplicaciones. Hoy en día, el término es usado ampliamente en un sinnúmero de contextos e incluye (pero no esta limitado) a:

- Sistema de soporte en la toma de decisiones.
- Aplicaciones para administración de recursos y personal.
- Administración de proyectos.
- Aplicaciones de actualización de bases de datos, etc.

La amplia variedad de recursos computacionales para realizar el procesamiento de transacciones, para efectuar el procesamiento de sistemas de información formal y de reportes, y también para brindar apoyo a las

decisiones gerenciales, se clasifican de manera general como el sistema de información gerencia1 para la organización o SIG.

1.1. Definición de Sistema de Información Gerencia1 (SIG)

Un sistema de información gerencia1 (SIG) es un sistema integrado para proveer información que apoye las operaciones, la administración, y las funciones de toma de decisiones en una empresa.

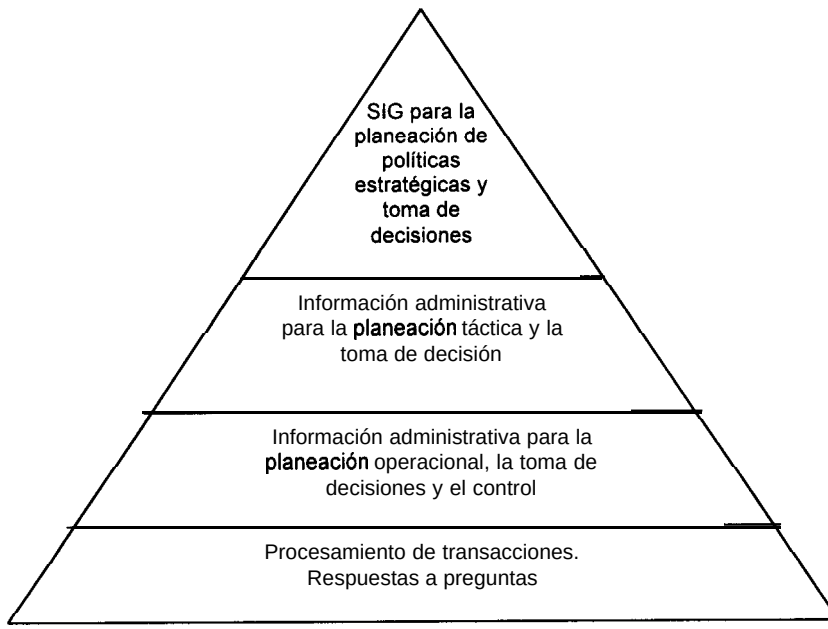
El SIG utiliza equipos de computación y software; procedimientos manuales; modelos para el análisis, la planeación, el control y la toma de decisiones y además una base de datos.

El SIG analizado administrativamente se describe como una estructura piramidal (ver grafico 1 .1) en la cual se lo describe más por la clase de información que cada nivel organizacional aporta al SIG. Cada nivel puede utilizar los datos proporcionados por los niveles más bajos, e incorporar también nuevos datos.

Como vemos, el concepto de SIG parece estar más ligado a las ciencias administrativas, que a las ciencias de la computación. Aquí

hay que aclarar algo, el contenido de la ciencia de la computación es importante, pero los sistemas de información gerencial, como disciplina académica, son en mayor grado una extensión de la administración que de la computación

Gráfico 1 .1. Estructura piramidal de un SIG.



1.2. La integración del SIG.

El principal problema de un SIG a nivel organizacional, es que cada nivel organizacional intenta desarrollar aplicaciones individuales, el inconveniente radica en que las aplicaciones individuales pueden ser

inconsistentes o incompatibles, además de generar ineficiencias en el intercambio de información entre los niveles.

De ahí la necesidad de la integración en el SIG. El primer paso en la integración de diversas aplicaciones del sistema de información es una planeación de conjunto del sistema de información. Aunque algunos sistemas aplicativos se han implementado uno a uno, su diseño puede ser dirigido según la planeación de conjunto, que determina como se integra con otras funciones. En esencia, el sistema de información se diseña como una federación planificada de pequeños sistemas.

La integración del sistema de información también se lleva a cabo a través de estándares, lineamientos, y conjuntos de procedimientos definidos en las funciones del SIG. La fortaleza de tales estándares y procedimientos permiten a las diferentes aplicaciones compartir los datos, cumplir con los requerimientos de auditoria y control, y además pueden ser compartidos por múltiples usuarios.

1.3. Utilización de los modelos.

Usualmente es insuficiente para los receptores humanos recibir solamente datos sin depurar o aun datos resumidos. Los datos necesitan procesarse y presentarse de tal manera que el resultado se dirija hacia la decisión que se va a tomar. Para lograr esto, el procesamiento de datos elementales se basa en un modelo de decisión; por ejemplo, una decisión sobre inversión realizada con gastos nuevos de capital, podría procesarse en términos de un modelo de decisión de desembolsos de capital.

Los modelos de decisión se pueden usar para apoyar las diferentes etapas en el proceso de toma de decisiones. Los modelos de inteligencia se pueden usar para la búsqueda de oportunidades y/o investigación de problemas. Los modelos se pueden utilizar para identificar y estudiar posibles soluciones. La selección de modelos, tales como los modelos de optimización, se puede emplear para hallar la solución más deseable.

En un sistema de información comprensivo, el decisor dispone de un conjunto de modelos generales que se pueden aplicar a muchos análisis y situaciones de decisión, además de un conjunto de modelos particulares para decisiones específicas. Están disponibles modelos

similares para la planeación y el control. El conjunto de modelos es la base de modelos para el SIG.

Los modelos son en general más efectivos cuando el gerente puede utilizarlos a través de un diálogo interactivo para elaborar un plan o volver a repetirlo mediante múltiples alternativas de decisión bajo diferentes condiciones.

1.4. La orientación del SIG.

Como lo habíamos descrito antes, el SIG, antes que un sistema complicado y enorme, es una federación de subsistemas estrechamente relacionados, el SIG no es más que una orientación que guía el desarrollo y la operación de los sistemas de procesamiento de datos.

1.5. Los subsistemas del SIG.

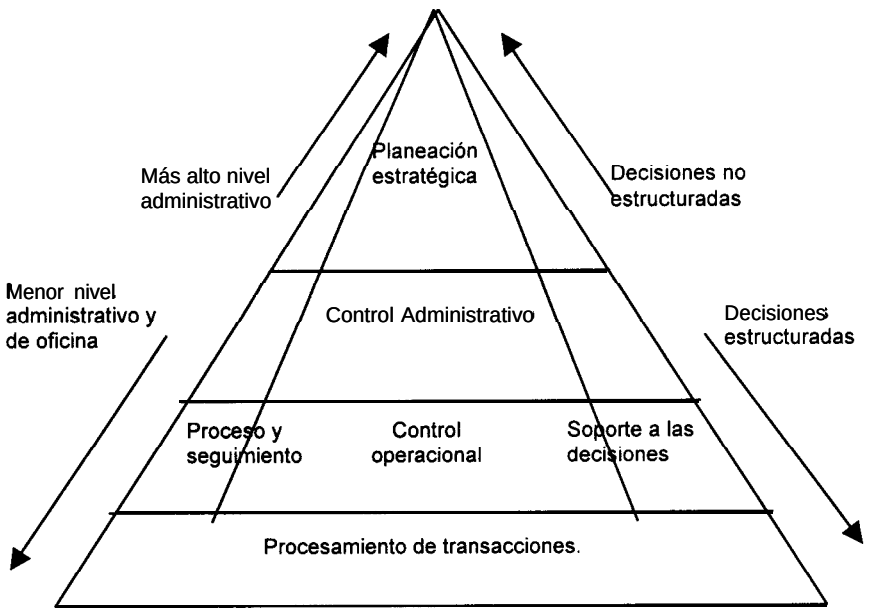
Los subsistemas del SIG se pueden describir en términos de funciones organizacionales (tales como mercadeo o producción) o de actividades (tales como el planeamiento y las transacciones). Cada subsistema funcional se puede ver como un subsistema que contiene actividades relacionadas con dicha función.

1.6. Estructura conceptual de un SIG.

La estructura de los sistemas de información gerencia1 se afecta por las actividades administrativas y las funciones de la organización; pero, en general, sus componentes físicos son el hardware, software, la base de datos, los procedimientos, y el personal.

Un SIG general debería estar estructurado de la siguiente forma:

Gráfico 1.2. Estructuración de un SIG.



CAPÍTULO 2

2. CONCEPTOS BÁSICOS SOBRE ESTADÍSTICA Y MUESTREO.

La Estadística es una rama de las ciencias matemáticas que es muy extensa, en esta primera sección definiremos los conceptos básicos que soportan al Muestreo, rama de la Estadística que estudia las técnicas que permitan tomar subconjuntos representativos de un universo de entes, al cual se les mide una o varias características. Para este capítulo, se supone que el lector tiene nociones básicas sobre Matemáticas, especialmente teoría de conjuntos.

Experimento es el proceso por medio del cual se obtiene una observación.

Cuando se efectúa un experimento, podemos tener uno o más resultados que se denominan eventos. A los eventos que se pueden descomponer en otros eventos se denominan eventos compuestos, mientras que los que no se pueden descomponer se denominan eventos simples. La notación tradicional para los eventos simples será E_i, donde i diferenciará ese evento simple del resto.

Se denomina espacio muestral al conjunto de todos los posibles eventos simples de un experimento.

Los espacios muestrales se subdividen en espacios discretos y continuos. Se denomina espacio muestral discreto al espacio que contiene un número finito o infinito contable de eventos simples, mientras que el espacio muestral continuo tiene un número infinito no contable de eventos simples. La notación tradicional para el conjunto muestral es la letra S.

Nótese que, así un evento sea simple 0 compuesto, siempre será subconjunto del espacio muestral, lo que nos lleva a una definición alternativa de evento como cualquier subconjunto de un espacio muestral, de acuerdo a la teoría de conjuntos.

A continuación, daremos un ejemplo de las definiciones anteriores.

Suponga que lanzamos un dado, dicho lanzamiento es un experimento, ya que por medio de él se pueden obtener una o más observaciones. Supongamos que la observación consiste en ver el número que queda en la cara superior del dado. Un ejemplo de los resultados posibles son:

E_1: Que aparezca el número 1

E_2: Que aparezca el número 2

E_3: Que aparezca el número 3

E_4: Que aparezca el número 4

E_5: Que aparezca el número 5

E_6: Que aparezca el número 6

A: Que aparezca un número par.

B: Que aparezca un número mayor a 3.

Nótese que los eventos E_i, $i = 1,2,...,6$ son indivisibles; mientras que el evento A está compuesto por los eventos E_2, E_4, E_6, y B está compuesto por los eventos E_4, E_5 y E_6. Los eventos E_i son eventos simples y los eventos A y B son eventos compuestos.

Nótese además que los eventos E_i son todos los eventos simples posibles en el experimento, por lo que conforma el espacio muestral S. Lo anterior, descrito en notación de conjuntos sería:

$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$

$A = \{E_2, E_4, E_6\}, A \subseteq S$

$B = \{E_4, E_5, E_6\}, B \subseteq S$

Observe que todos los eventos anteriormente descritos son subconjuntos del espacio muestral S.

Con todos las definiciones anteriores, podemos pasar a definir la función probabilidad.

Sea S un espacio muestral asociado con un experimento, y sea S el conjunto potencia de S; entonces se define la función de probabilidad P(S) como una función con dominio S y rango el intervalo de números reales [0,1], tal que cumple los siguientes axiomas:

Axioma 1: $P(A) \geq 0$, $A \in S$.

Axioma 2: $P(S) = 1$.

Axioma 3: Si $A_1, A_2, \ldots$ son elementos de S, entonces:

$P(A_1 \text{ u } A_2 \text{ u...}) = \sum_{i=1}^{\infty} P(A_i)$ siempre que $A_i \cap A_j = \{\}; i = 1,2,...; j = 1,2,...; i \neq j$

Se le recuerda al lector que se define como conjunto potencia de A al conjunto de todos los subconjuntos posibles de A.

Tomando el ejemplo anterior de los dados, la probabilidad de cualquier E_i es igual a 1/6; y:

$$A = E_2 \cup E_4 \cup E_6 = P(E_2) + P(E_4) + P(E_6) = 1/2$$

- Sean A y B eventos, entonces la probabilidad condicional de A, dado que ya ocurrió B es igual a P(A/B) = P(A ∩ B) / P(B), siempre que P(B) ≠ 0
- Dos eventos A y B son independientes si P(A ∩ B) = P(A)*P(B).

De estas definiciones, se pueden deducir los siguientes teoremas:

- Sean A y B eventos, entonces P(A ∩ B) = P(A)*P(B/A). Si A y B son independientes, entonces P(A ∩ B) = P(A)*P(B)

- Sean A y B eventos, entonces $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Si A ***y B son eventos mutuamente excluyentes, es decir que*** $P(A \cap B) = 0$, entonces $P(A \cup B) = P(A) + P(B)$.
- Si A es un evento, entonces $P(A) = 1 - P(A^c)$.

Ahora pasaremos a definir la variable aleatoria.

Sea S un espacio muestral, entonces se denomina variable aleatoria X a cualquier función que tenga como dominio S y como rango cualquier subconjunto de los números reales. Si el rango de X es un número finito o infinito contable de números reales, entonces se dice que la variable aleatoria X es ***discreta,*** de lo contrario X es ***continua.***

A continuación pasaremos a hacer ciertas definiciones sobre variables aleatorias.

Sea X una variable aleatoria continua, se denomina función de distribución de X a:

$F(x) = P(X \leq x)$, $-\infty < x < \infty$.

Pasaremos a revisar ciertas propiedades de la función de distribución.

Si F(x) es una función de distribución, entonces:

1. $\lim_{x \to -\infty} F(x) = 0$, también $\lim_{x \to \infty} F(x) = 1$

2. $F(x_b) \geq F(x_a)$ Si $x_b > x_a$.

Sea X una variable aleatoria con función de distribución F(x). Se dice que X es continua si F(x) es continua para $-\infty < x < \infty$.

Nótese que esta es una manera más formal de definir variable aleatoria continua.

Sea F(x) la función de distribución de una variable aleatoria continua X, entonces la función de densidad de probabilidad de X, denominada por f(x) está dada por:

$f(x) = \frac{dF(x)}{dx}$ = F'(x); siempre que la derivada exista.

Una propiedad importante (tal vez la de más importancia) de las funciones de densidad, es la siguiente:

Sea f(x) una función de densidad, y a y b dos números reales, entonces:

1. $f(x) \geq 0$, para cualquier x.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$

A continuación definiremos el valor esperado y la varianza de una variable aleatoria.

Sea X una variable aleatoria continua, entonces el valor esperado de X, denotado por $E[X]$, μ_x o simplemente μ, es:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Así mismo, la varianza de X, denotada por $Var(X)$, σ^2_x, o σ^2 :

$Var(X) = E[(X - \mu)^2]$, y se puede demostrar que $Var(X) = E[X^2] - \mu^2$.

La desviación estándar de X, denotada por σ_x, es igual a la raíz cuadrada positiva de la varianza.

A continuación definiremos **la** más importante de todas las funciones de densidad, la función de densidad normal.

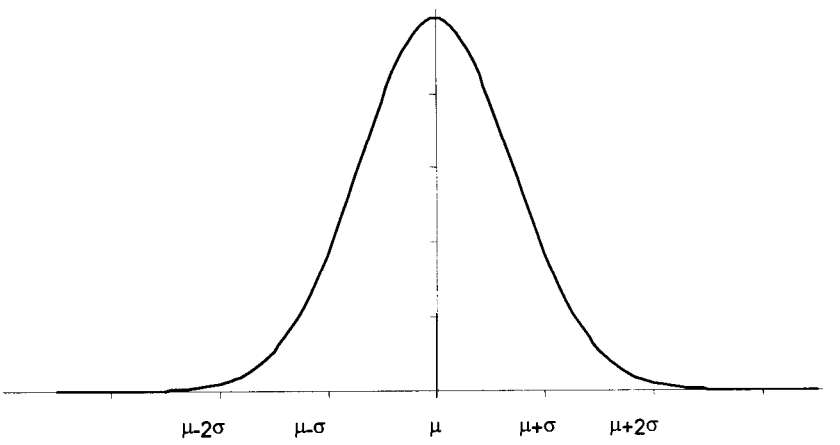
Sea X una variable aleatoria continua, se dice que X tiene una distribución normal, con parámetros μ y σ^2, notado con $X \sim N[\mu, \sigma^2]$, si y solo si:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty \leq x \leq \infty$$

Se puede demostrar que la media de X es igual a μ y la varianza, σ^2

A continuación, el gráfico de la variable aleatoria normal:

Gráfico 2.1. Función de densidad Normal



La función de densidad normal es muy importante en el estudio de la estadística actual, ya que representa con mucha precisión muchos comportamientos de variables económicas, biológicas, industriales, etc. Es más, se cree que el 70% de las variables (que son infinitas a nivel mundial) se pueden modelar usando una función de densidad normal. Además el muestreo (que es el principal objeto de este capítulo) se fundamenta en esta función de densidad.

Una de las densidades normales más importantes es aquella con media 0 y varianza 1 denominada *normal* estándar.

Otra función de densidad importante para nuestro estudio es la densidad γ^2, que se lee ji-cuadrado. Se dice que X tiene una distribución γ^2 con v grados de libertad, notado con $\gamma^2(v)$ si su función de densidad es la siguiente:

$$\frac{x^{\frac{v}{2}-1}e^{\frac{-x}{2}}}{2^{\frac{v}{2}}\Gamma\left(v/2\right)}, \quad x \geq 0,\ v \geq 0,\ \Gamma(n) \approx (n\text{-}1)!$$

Se puede demostrar que E(X) = v y Var(X) = 2v. Esta densidad nos ayuda con la estimación de la varianza en el muestreo.

Hasta aquí, todas las definiciones son para el caso unidimensional, es decir para una variable aleatoria. En la vida real, muchas veces no se tiene una observación, sino que con cada individuo se asocia una seria de observaciones.

Por ejemplo, a un conjunto de personas se les puede medir la variable aleatoria peso, pero si deseáramos medir las variables peso, estatura, y talla de zapato; no lo podemos hacer como si fuesen variables aleatorias separadas ya que entre ellas tienen relación. Es por eso que en la siguiente parte vamos a extender las definiciones y teoremas para vectores de variables aleatorias, conjuntos de variables aleatorias cuyo rango ya no son los reales, sino R^n.

Sean X_1, X_2,...,X_n n variables aleatorias, entonces se define el vector aleatorio n-variado X como:

X= (X_1, X_2 ,..., X,) y se define x = (x_1,x_2 ,..., x,) como un valor que toma el vector.

Sea X un vector aleatorio n-variado, se define la función de distribución conjunta de X como:

$F(x) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x,)$.

De la misma forma, la función de densidad conjunta de X, denotado por f(X) como:

$$f(x) = \frac{dF(x)}{dx_n dx_{n-1} \ldots dx_1}$$

Las funciones de densidad marginales de las variables $X_1, X_2, \ldots, X_n$ se definen como:

$$f_i(x_i) = \int_{-\infty}^{\infty} f(x) dx_i$$

Ahora pasaremos a definir la independencia de variables aleatorias.

Sea X un vector aleatorio n-variado, se dice que las n variables aleatorias son independientes si y solo sí:

$F(x_1, x_2, \ldots, x,) = F(x_1) * F(x_2) * \ldots * F(x_n)$.

Se puede demostrar que cuando las variables aleatorias son independientes se cumple lo mismo con las funciones de densidad de las variables.

Cuando las variables aleatorias no son independientes, se dice que son dependientes. Existe ciertas medidas que nos permiten decidir cuando las variables aleatorias son independientes o dependientes (y el grado de dependencia entre ellas) y no conocemos la función de densidad conjunta de las variables (como es el caso en la vida real). Pasaremos a continuación a definir dichas medidas.

Sean X_1 y X_2 variables aleatorias con medias μ_1 y μ_2 respectivamente, se define la covarianza entre X_1 y X_2, denotada como $\text{cov}(X_1, X_2)$ o σ_{12} como:

$$\text{Cov}(X_1, X_2) = E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right]$$

Se puede demostrar que $\text{Cov}(X_1, X_2) = E(X_1 X_2) - E(X_1)*E(X_2)$.

La covarianza es una medida de la dependencia lineal que existe entre dos variables, pero lamentablemente es sensible a la escala de medición de las variables, es decir, si las variables no están medidas en la misma escala, la covarianza no es una medida confiable. Esto se soluciona estandarizando su valor, obteniendo un coeficiente adimensional, llamado coeficiente de correlación lineal, el cual se lo define como:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}, \quad -1 \leq \rho_{12} \leq 1$$

El teorema que nos ayuda a medir la dependencia de las variables se presenta a continuación.

Si X_1 y X_2 son variables aleatorias independientes, entonces $Cov(X_1, X_2) = 0$

El recíproco de este teorema no siempre se cumple, puesto que, el que la covarianza entre dos variables sea cero dice que no existe dependencia *lineal* entre ellas, pero no descarta la existencia de otro tipo de dependencia.

Una vez definido todo el fundamento teórico, vamos a pasar a los conceptos básicos del muestreo y de la estimación.

Se llama universo a cualquier colección finita o infinita de individuos, entes, o elementos.

Se denomina población a una o más características medidas u observadas a los miembros de un universo.

Se denomina muestra a un subconjunto representativo de una población.

Se denomina muestra aleatoria a una muestra que está compuesta por variables aleatorias independientes e idénticamente distribuidas.

Cabe recalcar en este concepto que, como se supone que todos los valores son tomados de una misma población, las variables deben tener la misma distribución teórica, y como la muestra es “aleatoria” que significa *al* azar, el valor de una observación no influye en la otra, por eso se dicen que son independientes.

Las definiciones anteriores son teóricas y expresan los conceptos básicos del muestreo, a continuación el conjunto de definiciones estadísticas que le dan soporte al muestreo.

Se denomina estadístico a una función de las variables aleatorias que se pueden observar en una muestra y de las constantes desconocidas. Los estadísticos se utilizan para hacer inferencias con respecto a parámetros poblacionales desconocidos. Como los estadísticos son funciones de las variables aleatorias, entonces el estadístico es una variable aleatoria.

Hay muchas clases de estadísticos, pero nos vamos a centrar en unos importantes llamados estimadores. Los estimadores son estadísticos que nos

permiten especular con un cierto grado de certeza sobre el valor de un parámetro poblacional desconocido. Por ejemplo, el estadístico:

$\bar{X} \frac{\sum_{i=1}^{n} X_i}{n}$ (donde las X_i's forman una muestra aleatoria de una población.)

Este estadístico es un estimador de la media poblacional μ, y se lo denomina media muestral. Los estimadores tienen propiedades que nos hacen decidir por el mejor, claro que éste no es el único estimador para la media, existe la mediana, la moda, etc., en fin todas estos estimadores para la media se denominan medidas de tendencia central. De estos estimadores, el mejor es la media muestral, por razones que después se explicarán, pero la principal la da el teorema que a continuación se presenta, llamado teorema del límite central.

Obviamente, la media no es lo único que podemos estimar, también podemos estimar varianza (a través de las medidas de dispersión) y proporciones, es más, esta tesis usa un estudio por muestreo donde las preguntas son cualitativas, lo que indica que la mayoría de inferencias que quisiéramos hacer deberán ser con respecto a las proporciones, pero la

teoría estadística se desarrolla a partir de la medía, y para las proporciones no es más que una particularización.

A continuación definiremos unos de los teoremas más importantes del muestreo, el teorema del límite central

Sean $X_1, X_2, \ldots . X_n$ variables aleatorias independientes e idénticamente distribuidas con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2 < \infty$, entonces:

$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ converge en ley, cuando n $\rightarrow \infty$ a una variable aleatoria normal estándar.

El teorema anterior es la piedra angular del muestreo, y la razón de todo este despliegue teórico en este capítulo para que el lector pueda comprenderlo. Aquí vemos que la media muestral de cualquier población, cuando el tamaño de la muestra es grande, puede aproximarse por medio de la distribución normal estándar.

A continuación definiremos una de las propiedades más deseables de los estimadores.

Sea $\hat{\theta}$ un estimador de un parámetro θ, entonces $E(\hat{\theta}) = \theta + B$, donde B es un número real denominado sesgo. Si $B \neq 0$, entonces se dice que $\hat{\theta}$ es sesgado, pero si $B = 0$ entonces se dice que $\hat{\theta}$ es insesgado.

Que un estimador sea insesgado es una propiedad deseable de un estimador, ya que quiere decir que el valor más probable que puede tomar es precisamente el parámetro que se está estimando.

Sea $\hat{\theta}$ un estimador de θ ,se define el cuadrado medio del error del estimador, denotado por $CME(\hat{\theta})$ como $E[(\hat{\theta} - \theta)^2]$.

Se puede demostrar que $CME(\hat{\theta}) = Var(\hat{\theta}) + B^2$, lo que da a entender que si el estimador es insesgado, entonces el cuadrado medio del error es igual a la varianza.

Se puede demostrar que $\overline{X}$ es un estimador insesgado de μ, esto se haría demostrando que $E(\overline{X}) = \mu$, así mismo se puede demostrar que $Var(\overline{X}) = \sigma^2/n$, donde n es el tamaño de la muestra y σ^2 es la varianza poblacional.

Nótese que a partir de lo mencionado arriba se puede deducir el tamaño de la muestra, pero la varianza poblacional es otro elemento que hay que conocer para deducir n. Esto se conoce en estadística como paradoja de Friedman, que para hacer el muestreo y estimar la media poblacional haya que conocer la varianza poblacional. En los siguientes párrafos resolveremos el problema con un teorema.

Con esto vemos que que el teorema del límite central es un caso particular para el estimador $\overline{X}$. Para hacer un teorema que generalice el teorema del límite central para todos los estimadores, necesitamos hacer una definición.

Sean Z y γ^2 dos variables aleatorias independientes, con distribución normal estándar y $\gamma^2(v)$ respectivamente, entonces se dice que la variable $T = \frac{Z}{\sqrt{\gamma^2/v}}$ tiene una distribución t de Student con n-l grados de libertad.

Esta distribución t es muy parecida a la normal estándar, en su forma, pero su utilidad se ilustra en el siguiente teorema.

Sea $\hat{\theta}$ un estimador para un parámetro θ de alguna población normal, obtenido de una muestra aleatoria de tamaño n, entonces :

$\frac{\hat{\theta} - E(\theta)}{\hat{\sigma}_{\theta}}$ tiene una distribución t con n-l grados de libertad.

Este teorema es de mucha utilidad en el muestreo, ya que el teorema del límite central pone una condición muy dura, que n sea infinitamente grande, aunque con un n > 40, se considera grande. El teorema anterior no pone restricción en el tamaño de la muestra, pero exige que la población sea normal. Con estos dos teoremas se abarcan las mayorías de los casos, pero aún existen poblaciones que no cumplen las condiciones ni del teorema del límite central, ni del teorema anterior. Esos casos caen en otra rama de la estadística llamada estadística no paramétrica que no será objeto de nuestro estudio. Este teorema nos libra también de la necesidad de tener que estimar la varianza poblacional, ya que la varianza del estimador siempre esta en función de parámetros poblacionales, pues vemos que el teorema trabaja con un estimador de la varianza del estimador. Más adelante veremos los estimadores de la varianza.

Este teorema nos sirve también para calcular tamaños de muestras para otros estimadores que no sean la medía muestral. Los resultados que se muestran en la tabla siguiente se pueden demostrar uno por uno, pero en

esta tesis se demostró el caso para la media muestra¡, y los otros se dan por demostrados.

Tabla 2.1. Estimadores más comunes.

Parámetro objetivo	Tamaño de la muestra	Estimador	$E(\hat{\theta})$	$\sigma^2_{\hat{\theta}}$
μ	n	$\bar{X}$	μ	$\frac{\sigma^2}{n}$
P	n	$\hat{p} = \frac{X}{n}$	P	$\frac{p(1-p)}{n}$

El parámetro p es la proporción de entes en una población que cumplen con cierto criterio en la población. Su estimador es X / n, donde X es el número de entes que cumplen con cierto criterio en la muestra. Este parámetro es el que más nos va a interesar en nuestro estudio. Vemos que ambos estimadores son insesgados, ya que el valor esperado de ambos es el parámetro poblacional que están estimando, y que la varianza del estimador siempre es inversamente proporcional al tamaño de la muestra, esto quiere decir que a muestra más grande menos error. Esto daría a pensar que si deseamos una estimación buena, entonces deberíamos tomar una muestra bien grande, lo cual no siempre es posible por aspectos logísticos, económicos, etc. El muestreo trata sobre esto, como tomar la muestra de

manera que se pueda minimizar la varianza del estimador y a la vez no incurrir en costos adicionales.

Sea cual fuere el estimador, la varianza siempre contiene el tamaño de la muestra y un parámetro poblacional, en el caso de la media es la varianza, y en el caso de proporciones es la proporción real de la población.

La varianza es una medida de dispersión y mide el cuadrado de la distancia entre los observaciones de una variable y su media. La varianza de una población para la cual se conoce su densidad o distribución teórica es $E[(X-\mu)^2]$; si no se conociese su distribución teórica, entonces:

$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$, donde N es el tamaño de la población.

Lo lógico sería pensar que el estimador para la varianza, que de ahora en adelante se notará S^2, a partir de una muestra sería lo mismo, pero sustituyendo N por n ,y μ por $\bar{X}$. Pero se puede demostrar que ese es un estimador sesgado para la varianza poblacional, sin embargo el estimador:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n-1}$$

es un estimador insesgado para la varianza poblacional.

Con este resultado, si la población es normal entonces:

$\frac{X-\mu}{S/\sqrt{n}}$ tiene una distribución t de Student con n-l grados de libertad.

Cuando se toma muestras, siempre se incurre en errores. Uno se lo cuantifica con el sesgo, otro con la varianza del estimador, pero hay un error en que se incurre por el simple hecho de hacer el muestreo, por haber tomado un subconjunto de la población y no toda la población, este error se denomina error de diseño o error de estimación notado por ε, y es la distancia entre el estimador y el parámetro que estima, es decir:

$\varepsilon = |\hat{\theta} - \theta|$

Vemos que en el caso más general de la media, el error esta dado por $\bar{X} - \mu$. Este error se llama de diseño porque se fija al momento de diseñar la

muestra, en otras palabras, de que tamaño debería ser la muestra para obtener el error que el investigador ha fijado.

Otra cantidad que se fija al momento de diseñar la muestra es la confianza, que no es más que la probabilidad de que el estadístico utilizado se encuentre entre dos percentiles de la distribución del mismo. En el caso de la distribución normal y la t, que son asintóticas y simétricas, la confianza se denomina al número $(1 - \alpha)$ tal que:

$P(-k_{\frac{\alpha}{2}} \leq \theta \leq k_{\frac{\alpha}{2}}) = 1 - \alpha$, donde θ representa a un estadístico y k la distribución de θ

La confianza también se fija al momento de diseñar la muestra, en otras palabras, como el estadístico por lo general es una función del tamaño de la muestra, de que tamaño debería ser la muestra para tener la seguridad de, que de 100 valores del estadístico que se calculen a partir de muestras diferentes pero de igual tamaño, $[(1 - \alpha) \times 100]$ de ellas estarán dentro del intervalo arriba señalado. Esta cantidad depende mucho de la investigación que se este realizando, pero por lo general se la fija entre 90% y 95%. Puede ser más 0 menos, pero la confianza es inversamente proporcional al tamaño de la muestra, y trabajar con confianzas grandes y errores pequeños puede llevar a tamaños de muestras muy grandes; muestras que son muy costosas.

El lector se preguntará, entonces si queremos la estimación perfecta, fijemos el error en 0 y la confianza en 100%, pero eso llevaría a un tamaño de muestra infinita o del tamaño de la población.

A continuación, pasamos a ilustrar la fórmula del tamaño de la muestra para la estimación de la media. Fijaremos la confianza en (l-a) y el error en ε y supondremos que la población de la que se muestrea es normal, con estas condiciones:

$$-t_{\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}} \Rightarrow \left|\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}\right| \leq t_{\frac{\alpha}{2}} \Rightarrow \frac{\varepsilon \cdot \sqrt{n}}{S} \leq t_{\frac{\alpha}{2}} \Rightarrow \sqrt{n} \leq \frac{t_{\alpha/2} S}{\varepsilon} \Rightarrow n \leq \frac{t_{\alpha/2}^2 S^2}{\varepsilon^2}$$

Eliminamos el valor absoluto una vez que se le aplicó al error, porque sabemos que S y n son números reales positivos.

Vemos que n es menor o igual que un valor, tomaremos la igualdad ya que con esto tomamos el máximo valor que n podría tomar, finalmente concluímos que:

$$n = \frac{t^2_{\alpha/2} S^2}{\varepsilon^2}$$

Vemos que era lo que esperábamos, que el tamaño de la muestra sea directamente proporcional a la confianza y a la varianza e inversamente proporcional al error.

En los casos que se pueda tomar muestras muy grandes o en los casos que amerite, se puede reemplazar la distribución t por la Z (normal estándar).
En esta fórmula se toma el supuesto que la población es infinita, existe una fórmula que se puede demostrar para poblaciones finitas, pero por lo general trabajaremos con poblaciones muy grandes que se pueden considerar infinitas.

En la fórmula siguiente, N es el tamaño de la población y n_0 el n que calculamos arriba con la presunción de población infinita, entonces de puede demostrar que:

$n = \frac{n_0}{1 + \frac{n_0}{N}}$, vemos que a medida que N es grande, n se acerca a n_0.

Todas estas formas de determinar el tamaño de la muestra se denomina *muestreo aleatorio simple*, ya que al momento de tomar la muestra

asumimos que el muestreo se realiza sin reposición (es decir, no existe probabilidad de tomar un ente dos veces) y tomando elementos de la población al azar. Este muestreo trabaja bien siempre que se cumpla la condición que la varianza sea homogénea, es decir que no existan subconjuntos de la población que tengan distintas varianzas entre sí, ya que si la varianza es heterogénea, esto no podría llevar a una estimación mala o irreal.

En los casos anteriores, existe otra técnica de muestreo denominada muestreo estratificado, el mismo que se caracteriza por dividir a la población en dos o más subpoblaciones denominadas estratos. La división se hace basada en el criterio de que dentro de cada estrato la varianza debe ser homogénea, y entre estratos debe haber bastante heterogeneidad de la varianza. Por ejemplo, un muestreo en que se este midiendo el promedio de ingresos mensuales de las familias en una ciudad. Sabemos que, en una ciudad, puede haber ingresos de $120, así como ingresos de $10,000, es decir con una varianza altísima. De la misma forma, sabemos que las ciudades están claramente sectorizadas en sectores de ingresos altos, medios y bajos, estos sectores determinan el promedio que cada familia tiene de ingresos, los que no da a pensar que en estos sectores la varianza es homogénea, pero entre los sectores hay mucha heterogeneidad de la varianza. Este es un buen ejemplo de estratificación.

CAPÍTULO 3

3. DISEÑO DE LA MUESTRA

En el capítulo anterior vimos el soporte matemático y estadístico para el muestreo. A continuación, aplicaremos las técnicas antes vistas, en conjunto con unas nuevas que aprenderemos, para poder diseñar una encuesta que nos permita analizar el estado actual de los sistemas de información gerencia1 en el Ecuador.

Una de las primeras cosas que hay que definir es el significado de la palabra encuesta. Muchas veces se ha asociado este término solamente con el instrumento utilizado para recopilar los datos, sea éste un cuestionario en papel, un formulario electrónico 0 cualquier otro formulario.

Una encuesta por muestreo es una técnica que permite hacer inferencias sobre la población de la que fue seleccionada la muestra. El diseño de una encuesta comprende varios aspectos íntimamente ligados, ya que el fallo de cualquiera de ellos puede invalidar la encuesta en su totalidad.

La encuesta es un proceso muy complejo que se lo puede dividir en:

- Trabajos preliminares.
  - Objetivo
  - Definiciones
  - Modelo esquemático de tablas y cuestionario

- Diseño de la muestra.
  - Plan de muestreo
    - Formación o actualización del marco
    - Utilización de información complementaria
    - Establecimiento de una jerarquía de unidades de muestreo
    - Tamaño de la muestra
    - Estratificación
    - Selección
  - Métodos de estimación

. Tipos de estimadores y fórmulas para la estimación de los errores debidos al muestreo

. Normas para el tratamiento de la falta de respuesta y otros errores ajenos al muestreo

- Trabajos de campo.
  - Procedimiento de recolección de datos
  - Selección y adiestramiento de agentes y supervisores
  - Redacción de manuales e instrucciones

- Proceso de datos.
  - Proceso y depuración automática de cuestionarios
  - Ajuste de la no-respuesta
  - Control de calidad para los errores de perforación y codificación
  - Estimaciones y preparación de tablas

- Evaluación de resultados
  - Discrepancias entre el diseño teórico y su aplicación
  - Estimaciones de los errores debidos al muestreo
  - Comparación con fuentes externas
  - Comparación con diseños alternativos
  - Análisis de costos

Obviamente que todos esos pasos deben hacerse, aunque dependiendo del tamaño de la investigación, muchos de estos pasos pueden realizarse implícitamente. Además, en ese esquema pueden aparecer muchos términos que no han sido definidos, pero conforme los vayamos alcanzado, los iremos definiendo e ilustrando con el caso particular de nuestra encuesta.

Uno de los primeros pasos es definir el objetivo, el objetivo tiene que ser claro, consistente y realizable, ya que una mala definición del objetivo llevará seguramente a una encuesta con resultados inconsistentes con utilidad nula. En nuestro caso específico, el objetivo es medir el estado actual de la organización de los sistemas en las empresas del medio.

El segundo paso son las definiciones y el diseño del cuestionario, así que se debe decidir que variables debemos medir de tal manera que apoyen a la realización del objetivo. Realizando un análisis utilizando los conceptos del primer capítulo de esta tesis, se definió las variables que eran necesarias medir y se realizó el cuestionario, el mismo que se encontrará en el anexo # 1. A continuación la definición de las variables con sus nombres y codificación:

Tabla 3.7. Codificación de la encuesta.

Variable	*Descripción*	*Valores*
	Aspectos generales de la organización	
X_1	Tamaño de la organización	1. Pequeña 2. Mediana 3. Grande
X_2	Sector industrial	1. Primario 2. Industrial 3. Servicios
X_3	Clase de la organización	1. Estatal 2. Privada 3. Autónoma
X_4	Uso de las computadoras	1. No usa computadoras 2. Soporte de actividades 3. Actividad principal
	Organización de las computadoras	
X_5	Propiedad del servicio informático	1. Propio 2. De terceros 3. Combinación de 1 y 2
X_6	Conexión entre computadoras	1. Redes independientes 2. Redes interconectadas

		3. Por separado
ζ_7	Protocolo de comunicación para redes independientes	1. TCP / IP 2. IPX / SPX 3. NetBios (NetBeui) 4. Otro 5. Desconoce la respuesta
ζ_8	Realización de auditoria de redes	1. Si 2. No
ζ_9	Protocolo de comunicación para red interconectada	1. TCP / IP 2. IPX / SPX 3. NetBios (NetBeui) 4. Otro 5. Desconoce la respuesta
ζ_{10}	Sistema operativo en que están basadas las redes y/o computadoras	1. Windows 95/98 2. Windows NT / 2000 3. Novell Netware 4. Linux 5. Basado en UNIX 6. Otro
ζ_{11}	Uso de las redes	1. Compartir archivos e impresoras

		2. Enviar y recibir mensajes 3. Ejecución de sistemas 4. Compartir conexión a Internet
Organización de los sistemas de información		
X_{12}	Tipos de sistemas usados	1. Sistemas desarrollados por y para la empresa 2. Sistemas desarrollados para la empresa por otros 3. Sistemas generales adaptados para su empresa 4. Utilitarios de distribución masiva
X_{13}	Forma de uso de los sistemas	1. Cada usuario usa un sistema 2. Cada sección usa un sistema 3. Toda la empresa usa ur módulo de un sistema integra
X_{14}	Realización de auditoria de sistemas	1. si 2. No
X_{15}	Uso de bases de datos	1. si 2. No
X_{16}	Formato de las bases de datos	1. Hoja de cálculo

		2. Dbase / Foxbase ***3.*** Access 4. (Visual) FoxPro 5. Servidor de bases de datos (Oracle, SQL Set-ver, etc.) ***6.*** Otro
X_{17}	vledios de facilitación de nformación entre departamentos	1. Reportes impresos 2. Medios magnéticos 3. Consultas entre bases de datos
X_{18}	Jso de los sistemas	1. Proceso y control de transacciones 2. Administración y planeación táctica 3. Planeación estratégica y toma de decisiones
Uso público de los sistemas		
X_{19}	Acceso a Internet	**1. si** ***2. No***
X_{20}	Tipo de conexión a Internet	1. Propia 2. Dedicada

		3. Ocasional DIAL-UP
X_{21}	Empresa en el Internet	1. si 2. No
X_{22}	Tipo de página Web	1. Informativa 2. Consulta de datos específicos 3. Consulta de datos secundarios 4. Comercio electrónico
X_{23}	Usuarios que ingresan a la página Web	1. Empleados 2. Proveedores 3. Distribuidores 4. Cliente o consumidor final
X_{24}	Conexión de página con sistemas	1. Si 2. No
X_{25}	Protección contra intrusos	1. si 2. No
X_{26}	Realización de respaldos	1. si 2. No
X_{27}	Tiempo de recuperación en casos fortuitos	1. Inmediatamente 2. Un día 3. De uno a tres días

		4. De tres días a una semana 5. Más de una semana 6. No sabe

Vemos claramente que no usaremos ninguna variable medible en unidades, sino que nos centraremos en las proporciones de respuesta por uno u otro valor de las variables. Este tipo de variables es de mucha utilidad en el caso de estudios cualitativos, como es el caso del nuestro.

Una vez que hemos definido las variables, pasaremos a diseñar el plan de muestreo. En el se necesita la definición del marco muestral. El marco muestral no es más que la definición física de los entes en la población, por ejemplo; si queremos estudiar la estatura de un grupo de personas, las personas son los entes, pero el marco muestral sería un listado de dichas personas, si queremos medir el promedio del ancho de las calles en Guayaquil, nuestros entes son las calles, pero nuestro marco muestral sería una cartografía de Guayaquil con las calles señaladas o un listado de las calles.

En nuestro caso, los entes que conforman nuestra población objetivo son las empresas, así que nuestro marco muestral debería ser el listado de las empresas a nivel nacional, proporcionado por la Superintendencia de

Compañías. Lamentablemente, por logística sería imposible abarcar todo el territorio nacional en esta investigación, así que tomaremos únicamente la ciudad de Guayaquil como referencia.

A continuación, obtendremos el tamaño de la muestra. Como sabemos, el tamaño de la muestra es una función del error de diseño, la varianza y el tamaño de la población.

Debido a que el tamaño de la población es muy grande (varios miles de empresas con base y operaciones en la ciudad), supondremos que el tamaño de la población es infinito (ya que la contribución de este parámetro es insignificante al momento de obtener el tamaño de la muestra).

Otro detalle que hay que tomar en cuenta es el error de diseño y la varianza. Aquí hay que observar que estos valores son unidimensionales, es decir, tenemos 27 variables, por lo que tenemos 27 varianzas y 27 posibles errores de diseño por decidir.

En este punto hay que elegir una sola de todas las variables, esta variable o pregunta se la denomina variable de diseño. La variable de diseño es la pregunta más importante de toda la encuesta, es la pregunta cuyo resultado debe contribuir en gran parte al cumplimiento de los objetivos de la encuesta,

y es la que debemos asegurar que tenga el menor sesgo, menor varianza y otras propiedades deseables.

La elección de la variable de diseño depende de un análisis, más bien cualitativo que cuantitativo, no se decide por ella a partir de datos o cálculos sino por entrevistas con expertos en el campo de la encuesta, observación, análisis de sectores, etc.

Con nuestro objetivo específico, que es de medir la organización de los sistemas, creo que la pregunta que deberíamos hacernos es: ¿Qué empresas disponen de una mejor organización de sistemas y están más "avanzadas" en este aspecto?. Conversando con ciertas personas conocidas en sistemas, los puntos que más influyen son las preguntas del primer bloque, sobre los aspectos generales. Sobre esta base, procederemos al siguiente análisis cualitativo:

El tamaño de la empresa, a mi criterio, no influye significativamente en la pregunta que nos hacemos, ya que existen empresas pequeñas que tienen una muy buena organización de sistemas con presupuestos modestos, contra grandes empresas de presupuestos exorbitantes con una organización de sistemas modesta.

La clase de organización es un detalle que si influye de manera directa en nuestro objetivo, ya que las empresas de servicios tienden a preocuparse más de su organización de sistemas, además que estos tipos de empresas son relativamente nuevas en el medio, pero aun así, la pregunta siguiente (la cuarta) sobre el uso de computadoras, también es importante ya que las empresas que tengan la computación como actividad principal se preocuparán más de su organización.

Según el criterio de expertos en la materia, la variable de diseño debería ser la cuarta, ya que tiene las características deseables.

La pregunta 4 tiene tres niveles:

1. Soporte para sus actividades.
2. Actividad principal.
3. No utiliza computadoras.

Como el tamaño de la población es bastante grande, debemos tomar una muestra piloto y medir los valores de la variable de diseño elegida para calcular la varianza de la misma y poder obtener el tamaño de la muestra.

Se tomó una muestra de 50 empresas, las cuales fueron encuestadas obteniendo los siguientes resultados:

Porcentaje de empresas que eligieron la opción # 1: 0.5333
Porcentaje de empresas que eligieron la opción # 2: 0.2333
Porcentaje de empresas que eligieron la opción # 3: 0.2333

Como la opción # 1 es la de proporción más alta, la elegiremos para calcular la varianza, es decir será nuestra estimación para la proporción p, nuestro estimador $\hat{p}$.

Supondremos que tenemos suficiente datos para asumir la normalidad de los datos.

Trabajaremos con 95% de confianza y un error de diseño atribuible al muestreo del 9% para la proporción (0.09).

A continuación el cálculo del tamaño de la muestra:

Los parámetros son los siguientes:

$Z_{\alpha/2}$ = 1.96 con 95% de confianza.

$\hat{p}$ = 0.533

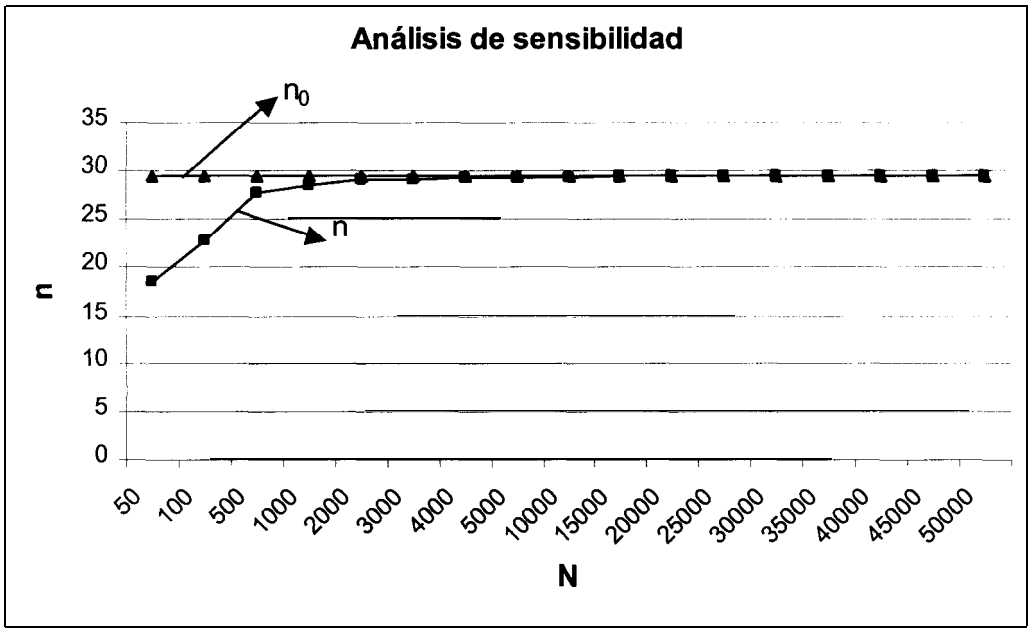
4 = 1 - $\hat{p}$ = 0.47

e = 0.09

$$n = \frac{Z_{\alpha/2}^2 \cdot (P \cdot Q)^2}{e^2} \Rightarrow n = \frac{(1.96)^2 \cdot (0.533 \cdot 0.47)^2}{(0.09)^2} \Rightarrow n = 29\,42 \approx 30$$

Vemos que a pesar del gran tamaño de la población, el tamaño de la muestra es relativamente pequeño. Pasaremos a analizar *n* como una función de *N* (el tamaño de la población) para ver la sensibilidad del tamaño de la muestra con respecto al tamaño de la población, análisis conocido como análisis de sensibilidad:

Gráfico 3.1. Análisis de sensibilidad.



Vemos que el análisis de sensibilidad es muy concluyente, ya que el tamaño de la muestra converge rápidamente al tamaño por nosotros encontrado, denotado por no, lo cual implica que así el tamaño de la población sea de 2,000,000, el tamaño de la muestra permanecerá invariable, lo cual explica el tamaño de la muestra tan pequeño para un población tan grande.

Este resultado muy afortunado nos permite tener la muestra a partir de la muestra piloto, así que elegiremos al azar 30 empresas de las 50 muestreadas solo con la pregunta diseño, y se procedió a realizar las 27 preguntas completas ahora sí a las 30 empresas seleccionadas, en el

siguiente capitulo se procederá a analizar los resultados de la encuesta realizada.

CAPÍTULO 4.

4. ANÁLISIS UNIVARIADO DE LA MUESTRA.

Como vimos en el capítulo anterior, el tamaño de la muestra ya fue seleccionado, y ahora procederemos a hacer el análisis de los datos obtenidos.

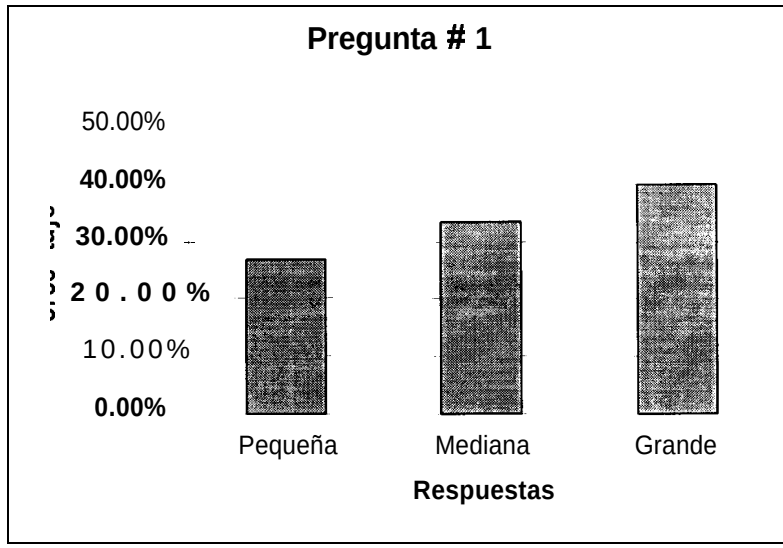
Lo haremos de la siguiente manera, primero haremos un análisis descriptivo para cada una de las variables y después realizaremos análisis de las relaciones o dependencias entre ellas.

Es de recordar que el tipo de muestreo utilizado es aleatorio simple, así que las empresas fueron elegidas al azar.

4.1. Pregunta # 1:

La pregunta 1 averiguaba sobre el tamaño de la empresa encuestada, y se daban tres opciones de respuesta: pequeña, mediana y grande. A continuación el análisis de las respuestas:

Gráfico 4.1. Pregunta # 1.



Al hacer la selección aleatoria, se eligió el 40% de empresas grandes, 33.33% de empresas medianas, y 26.67% de empresas pequeñas.

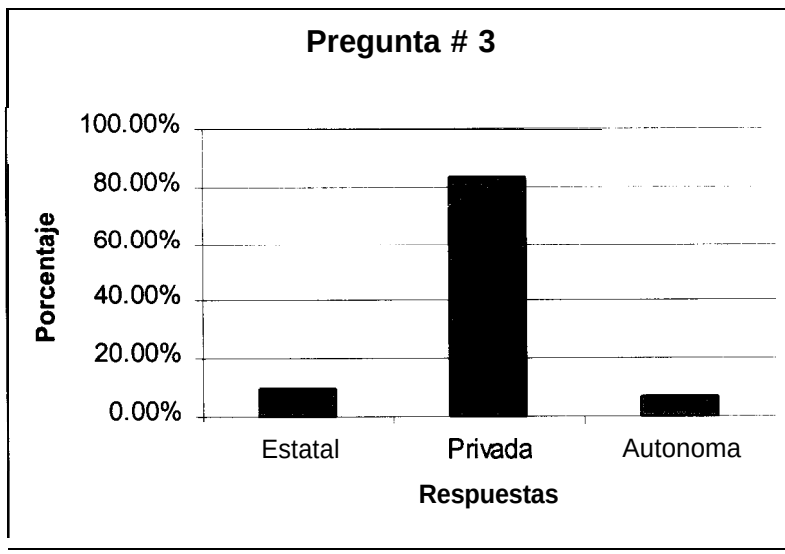
La pregunta # 2 no fue analizada por el motivo que al hacer las encuestas, las empresas del tipo industrial y primaria no colaboraron

con las mismas, por lo que se tomó solo empresa de servicios, dejando sin utilidad esta pregunta ya que una pregunta con varianza 0 no tiene sentido, de aquí y en adelante esta pregunta queda excluida del análisis.

4.2. Pregunta # 3:

La pregunta 3 indagaba sobre la clase de organización encuestada, y tenía tres niveles: Estatal, Privada y Autónoma. Las respuestas encontradas fueron:

Gráfico 4.2. Pregunta # 3.

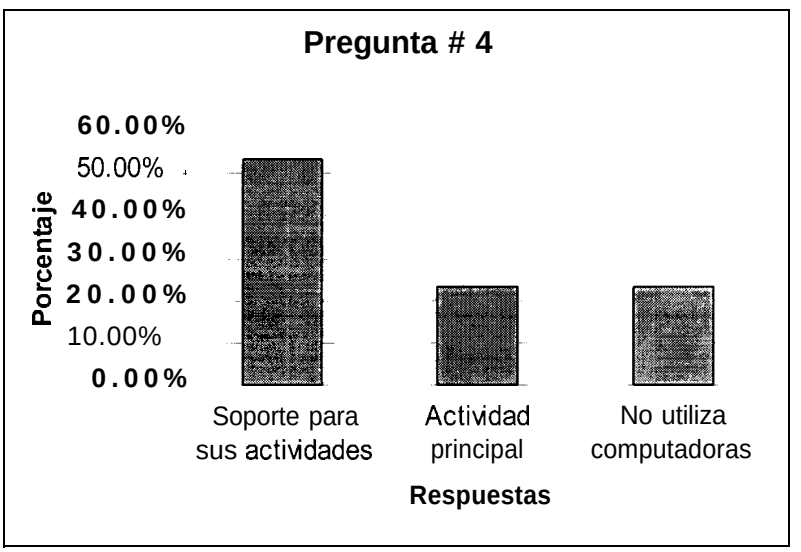


De las empresas elegidas, el 10% eran estatales, el 83.33% privadas, y el 6.67% autónomas.

4.3. Pregunta # 4:

La pregunta 4 es la más importante de esta muestra, ya que es la variable de diseño y preguntaba sobre el uso de computadoras en las empresas, con tres niveles que son: Soporte para sus actividades, actividad principal, y no utiliza computadoras. Los resultados fueron:

Gráfico 4.3. Pregunta # 4.



Vemos que el 53.33% de las empresas utilizan las computadoras como soporte para sus actividades, el 23.33% como actividad principal, y el mismo 23.33% no utiliza computadoras para nada.

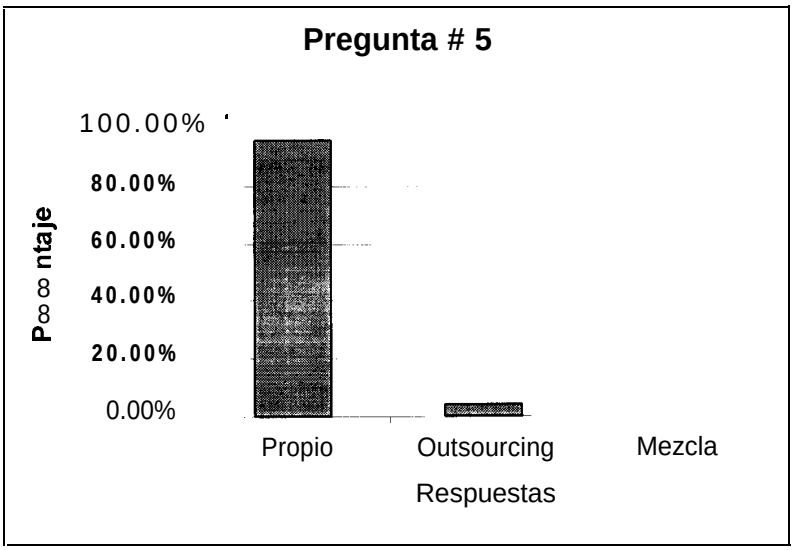
A partir de esta pregunta empiezan las preguntas específicas sobre los sistemas de información, que empieza por tener una buena infraestructura informática. La variable de diseño da información de que en Guayaquil el 23.33% de las empresas ni siquiera utilizan computadoras para realizar sus actividades, lo cual es preocupante en esta época que muchos expertos han calificado como “el siglo de la información”.

4.4. Pregunta # 5:

La pregunta # 5 implica el inicio de un nuevo bloque de preguntas, el cual está condicionado a la respuesta en la pregunta # 4, ya que si la organización no tiene computadoras, los otros bloques de pregunta no tienen razón de ser. Igual, eso va a ser transparente al análisis ya que estamos trabajando con porcentajes, se entenderá en los siguientes porcentajes como “el porcentaje de las empresas que si utilizan computadoras”.

En esta pregunta se indagó por la propiedad del servicio de informática de las empresas que utilizan computadoras, los niveles de esta pregunta son: Propio, outsourcing, y mezcla de los anteriores. Los resultados fueron:

Gráfico 4.4. Pregunta # 5.



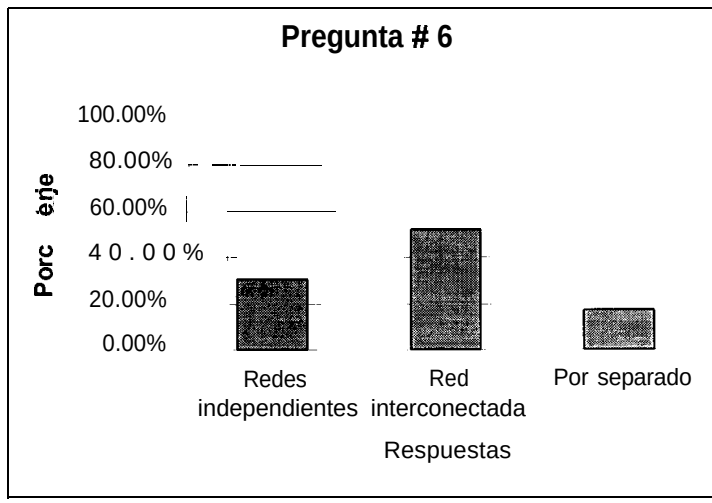
Vemos que el 95.65% de las empresas encuestadas tienen un departamento de Informática propio, mientras que tan solo el 4.35% de las empresas encuestadas lo concesionan, tercerizan o lo entregan en outsourcing a otras compañías, lo cual muestra la poca utilización del outsourcing, que es un modelo administrativo de gran aceptación en otras latitudes.

4.5. Pregunta # 6:

Esta pregunta es otra pregunta condicionada, ya que la respuesta puede alterar el orden natural de la encuesta, por lo que debe entendérsela como se entendió la pregunta # 4.

La pregunta 6 investiga como están conectadas o si no están conectadas en redes las computadoras de la empresa, ésta es otra pregunta de importancia para los sistemas de información ya que de esto depende su organización. Los resultados fueron:

Gráfico 4.5. Pregunta # 6.

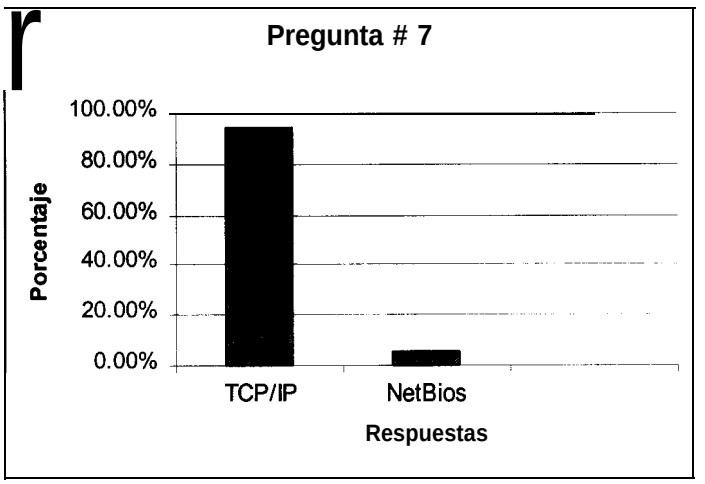


Este histograma indica que las organizaciones que usan computadoras prefieren tener una única red interconectada, lo cual es bueno desde el punto de vista de los sistemas, el 30.43% tiene varias redes independientes en su organización, el 52.17% tienen una única red, y el 17.39% las tienen por separado. A las organizaciones que respondieron por separado, no se les siguió haciendo las preguntas de este bloque.

4.6. Pregunta # 7:

Esta pregunta indaga sobre el protocolo más utilizado en las redes, claro siempre y cuando no se haya respondido en la pregunta 6 que las computadoras estaban por separado. El resultado es:

Gráfico 4.6. Pregunta # 7.

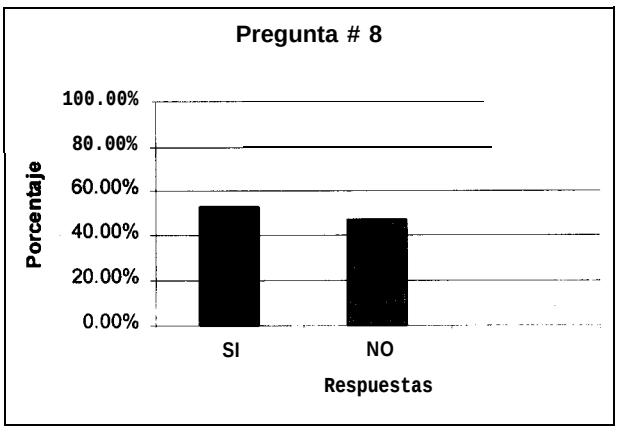


Vemos que entre las empresas que si usan redes, el 94.74% utilizan el protocolo TCP/IP como protocolo de comunicación en la red, lo cual es bueno desde el punto de vista de los sistemas de información ya que es el mejor protocolo al momento de desarrollar sistemas en redes y permite explotar los conceptos de Intranet, Internet, etc., apenas el 5.26% manifestó utilizar NetBios, el cual es un protocolo básico en desuso, y nadie eligió las opciones adicionales.

4.7. Pregunta # 8:

Esta pregunta es de suma importancia ya que analiza si las redes de las organizaciones han sido avalizadas por una empresa externa de certificación si física y lógicamente es apropiada la configuración de la misma. Los resultados fueron:

Gráfico 4.7. Pregunta # 8.



Vemos que en esta pregunta hay casi un empate técnico entre el si y no, 52.63% para el SI, y 47.37% para el NO, lo cual es hasta cierto punto engañoso ya que son pocas las empresas que cumplen los requisitos para llegar hasta esta pregunta, ya que en este momento se han excluido las empresas que no usan computadoras y no tienen una red o redes.

4.8. Pregunta # 9:

La pregunta # 9 es un poco redundante con respecto a la pregunta # 7, se la hizo pensando en respuestas diferentes a ellas, pero no fue así, en la práctica las dos preguntas dieron resultados iguales, así que se omitirá el análisis de esta variable.

4.9. Pregunta # 10:

El análisis de la pregunta 10 es especial ya que en esta pregunta, el encuestado podía elegir más de una opción, es por eso que hay que analizar el porcentaje de respuesta de cada opción dado que eligió otra de ellas también. Esto lo hacemos por medio de una tabla cruzada o “crosstabs”, la cual presentamos a continuación:

Las empresas encuestadas solo usaron tres sistemas operativos para sus respuestas: Windows 95198, Windows NT/2000, y sistemas basados en Unix.

Tabla 4.1. Pregunta # 10.

Pregunta # 10: Sistema operativo

% within X10-1

UNIX			Windows NT/2000 N	Windows NT/2000 S	Total
N	W95/98	N			100.0%
		S		100.0%	100.0%
	Total		45.8%	54.2%	100.0%
S	W95/98	S	16.7%	83.3%	100.0%
	Total		16.7%	83.3%	100.0%

Vamos a analizar las respuestas: el 83.3% de los encuestados usaban Windows 95/98 en unión con Unix y Windows NT/2000, mientras que el 16.7% usaban Windows 95/98 en unión con Unix, y no utilizaban Windows NT/2000, y así se puede analizar todos los valores de la tabla. Vemos que en realidad todos los encuestados usan Windows 95/98, pero varían en los sistemas operativos de los servidores.

4.10. Pregunta # ll :

El manejo de esta pregunta es similar al de la pregunta # 10, ya que el encuestado puede elegir más de una opción. Los resultados fueron:

Tabla 4.2. Pregunta # ll.

Pregunta # ll : Uso de las redes

% within XI l-l

				Mensajes		
Internet	Sistema			.00	1.00	Total
.00	.00	Compartir	.00	100.0%		100.0%
			1.00	100.0%		100.0%
		Total		100.0%		100.0%
	1.00	Compartir	1.00	66.7%	33.3%	100.0%
		Total		66.7%	33.3%	100.0%
1.00	.00	Compartir	1.00	75.0%	25.0%	100.0%
		Total		75.0%	25.0%	100.0%
	1.00	Compartir	1.00	37.5%	62.5%	100.0%
		Total		37.5%	62.5%	100.0%

El 62.5% de las empresas encuestadas utilizaban su red para las cuatro opciones disponibles en esta pregunta que eran: Para compartir archivos e impresoras, envío y recepción de mensajes, ejecución de un sistema integrado y para compartir conexión a Internet, mientras que el 25% la utilizaban para todas las opciones menos para ejecutar sistemas de información, y así se podría analizar cada celda de la tabla.

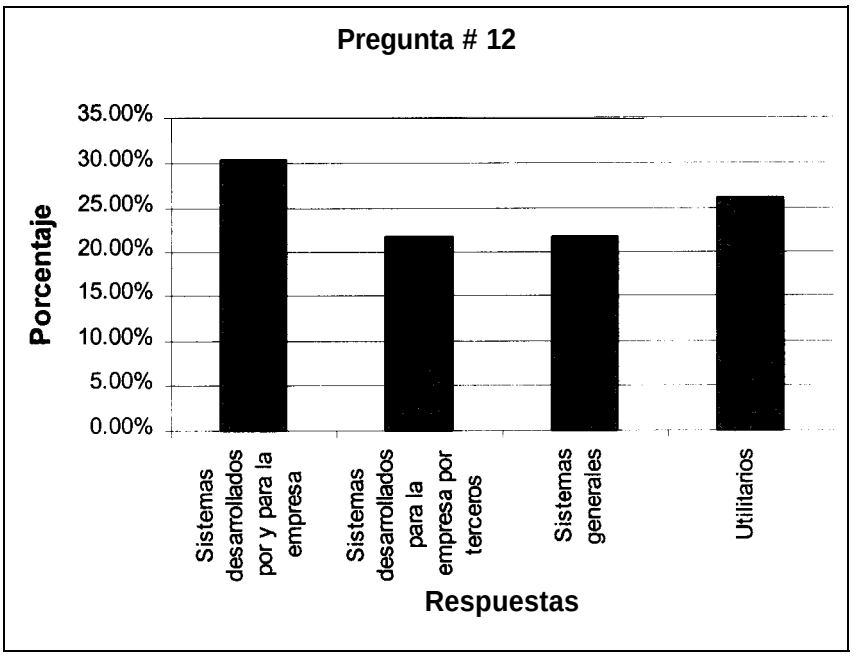
4.11. Pregunta # 12.

La pregunta 12 es el inicio de otro bloque, y de la misma forma que en los bloques anteriores, esta pregunta condiciona la realización del

bloque entero, considerando que los encuestados que saltaron el bloque anterior, pueden contestar éste, ya que trata sobre el software, y el anterior lo hacía sobre las redes.

La pregunta # 12 investiga el tipo de programas que utiliza la empresa para desarrollar sus actividades y las opciones son: Sistemas desarrollados por y para la empresa, sistemas desarrollados para la empresa por terceros, sistemas generales adaptados para su empresa, y los utilitarios de distribución masiva como Word, Excel, etc. A continuación los resultados:

Gráfico 4.8. Pregunta # 12.

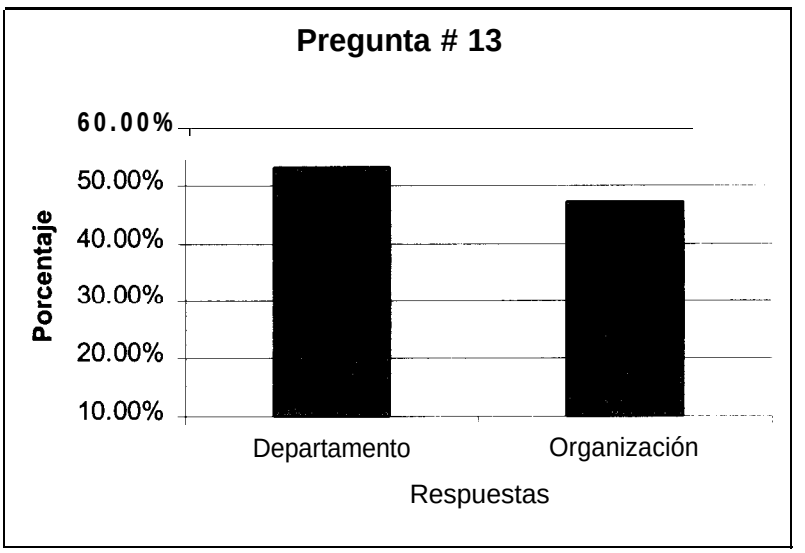


Vemos que los porcentajes son muy cerrados aquí, la mayoría de las empresas encuestadas (30.43%) prefieren los sistemas desarrollados "casa adentro", le sigue quienes solo usan utilitarios para sus operaciones (26.09%), y en un empate (21.74%, lo cual era de esperarse) los sistemas desarrollados por terceros y los sistemas generales. Entiéndase por sistemas generales a sistemas tipo Mónica para Contabilidad, XASS para la agricultura, etc. Los que contestaron utilitarios en esta pregunta se saltaron directamente a la siguiente sección de la encuesta.

4.12. Pregunta # 13:

La siguiente pregunta trata sobre la distribución y alcance de los sistemas de información dentro de la organización, las opciones disponibles son: Cada computadora usa un sistema, cada departamento usa un sistema, y el sistema es integrado a nivel organizacional. Los resultados fueron:

Gráfico 4.9. Pregunta # 13.

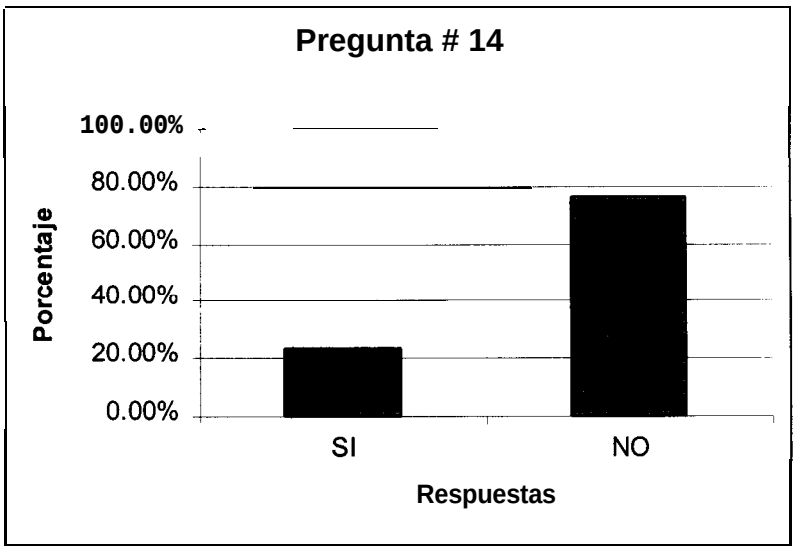


El 52.94% de las empresas que poseen sistemas de información, tienen un sistema independiente por cada departamento, mientras que el 47.06% tienen sus sistemas integrados a nivel organizacional.

4.13. Pregunta # 14:

Esta pregunta trata sobre si las empresas que poseen sistemas de información realizan auditorias periódicas de los mismos. Los resultados fueron:

Gráfico 4.10. Pregunta # 14.



Como era de esperarse, la gran mayoría de nuestras empresas en Guayaquil que poseen sistemas de información (76.47%) no realizan auditorias periódicas de sus sistemas, solo el 23.53% de las mismas realizan éstas tan importantes auditorias.

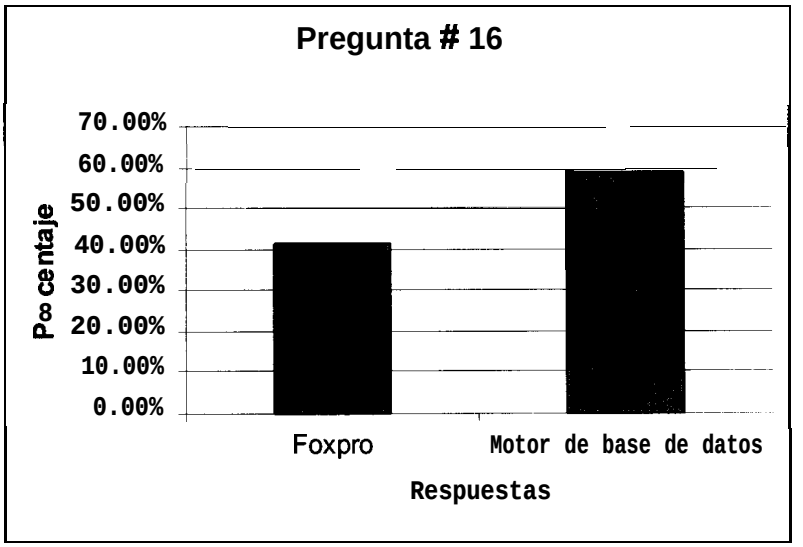
4.14. Pregunta # 15:

La pregunta 15 es un poco obvia, ya que todos los sistemas por lo general usan bases de datos, pero nos referimos también a los sistemas que simplemente usan archivos de texto planos, que a pesar de ser muy raros, todavía existen. Sin embargo, en las empresas encuestadas que usaban sistemas, todas usaban bases de datos, así que como el resultado es 100% que si utilizan bases de datos, no analizaremos esta pregunta.

4.15. Pregunta # 16

Esta pregunta trata sobre el formato de las bases de datos que utilizan las bases de datos. A continuación los resultados:

Gráfico 4.11. Pregunta # 16.

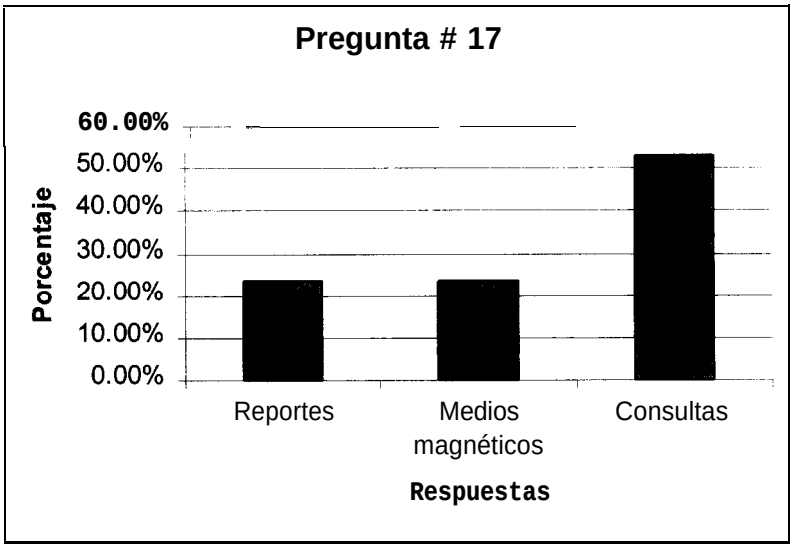


Vemos que los formatos más utilizados para las bases de datos son los motores de bases de datos (SQL Set-ver, Oracle, etc.) con un 58.82% de utilización, y FoxPro con el 41.18%, ninguno de nuestros encuestados eligieron las otras opciones.

4.16. Pregunta # 17:

La pregunta que viene ahora es sobre como se administra la salida y la entrega de información en los sistemas que poseen las empresas. Las opciones disponibles son: reportes impresos, medios magnéticos, y consultas directas al sistema. A continuación los resultados:

Gráfico 4.12. Pregunta # 17.



Vemos que las consultas directas son el método más popular (también el más conveniente) entre los sistemas de información (52.94%), y en un empate técnico las otras dos opciones (las menos convenientes) con un 23.53%.

4.17. Pregunta # 18:

Esta pregunta merece el mismo tratamiento que las preguntas 10 y 11, ya que el encuestado puede contestar a más de una opción, y es una de las preguntas más importantes ya que trata sobre la organización administrativa de los sistemas de información. La pregunta es sobre las actividades que realizan los sistemas y las

opciones son: Proceso y control de transacciones (nivel 1), Administración y planeación táctica (nivel 2), y planeación estratégica y toma de decisiones (nivel 3). Los resultados, presentados en tablas cruzadas son:

Tabla 4.3. Pregunta # 18.

Pregunta # 18

% of Total

Nivel 1			Nivel 2 0	Nivel 2 1	Total
0	Nivel 3	0	%		%
	Total		100.0%		100.0%
1	Nivel 3	0	76.5%	5.9%	62.4%
		1		17.6%	17.6%
	Total		76.5%	23.5%	100.0%

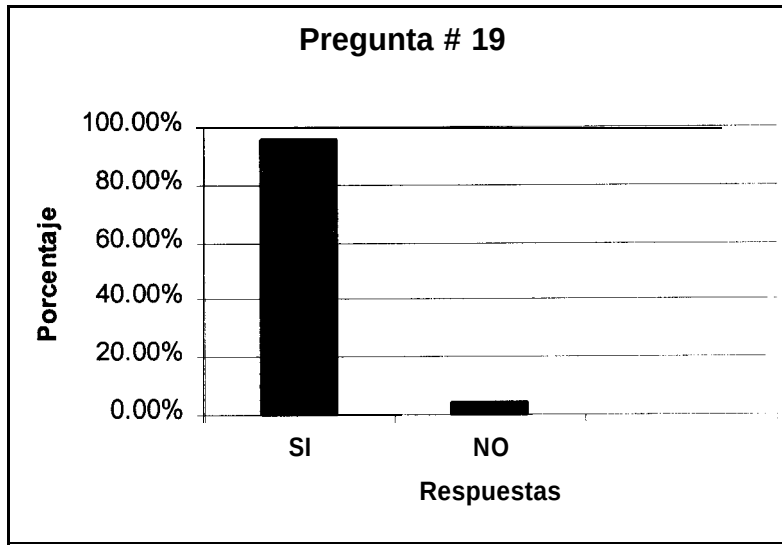
Vemos que el porcentaje de empresas que utilizan los tres niveles son pocas (17.6%), y la mayoría de las empresas solo utilizan sus sistemas para el trabajo netamente transaccional como lo muestra la tabla anterior (76.5%).

4.18. Pregunta # 19:

Esta pregunta es el inicio de un nuevo bloque de la encuesta, y se trata sobre el uso de la Internet en los sistemas, y el resultado de esta

pregunta condiciona el resto de este bloque. La pregunta es que si la compañía tiene acceso 0 no a Internet, las respuestas fueron:

Gráfico 4.13. Pregunta # 19.

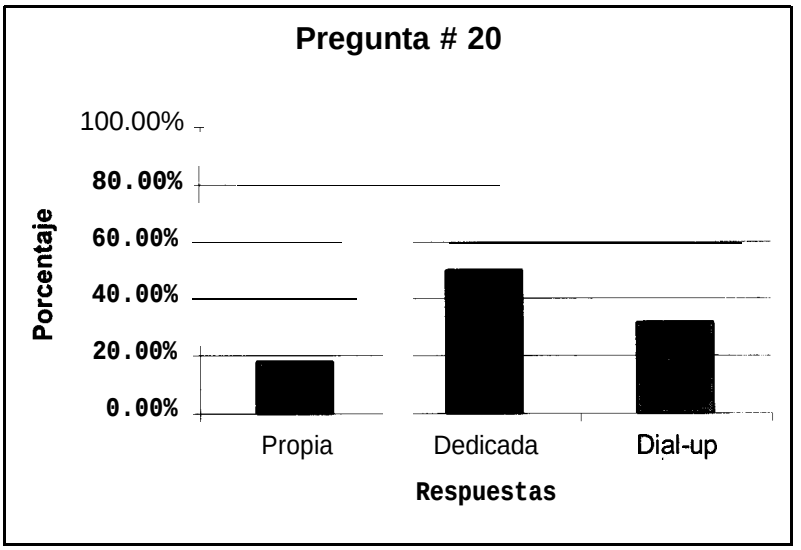


Vemos que la mayoría de compañías tienen algún tipo de acceso a Internet (95.65%), y tan solo el 4.35% no tiene, lo cual asegura una mayor cantidad de empresas consideradas en el presente bloque.

4.19. Pregunta # 20:

Esta pregunta trata sobre el tipo de conexión a Internet que tienen las empresas, y hay tres opciones de elección: propia, dedicada, y ocasional dial-up. Las respuestas fueron:

Gráfico 4.14. Pregunta # 20.

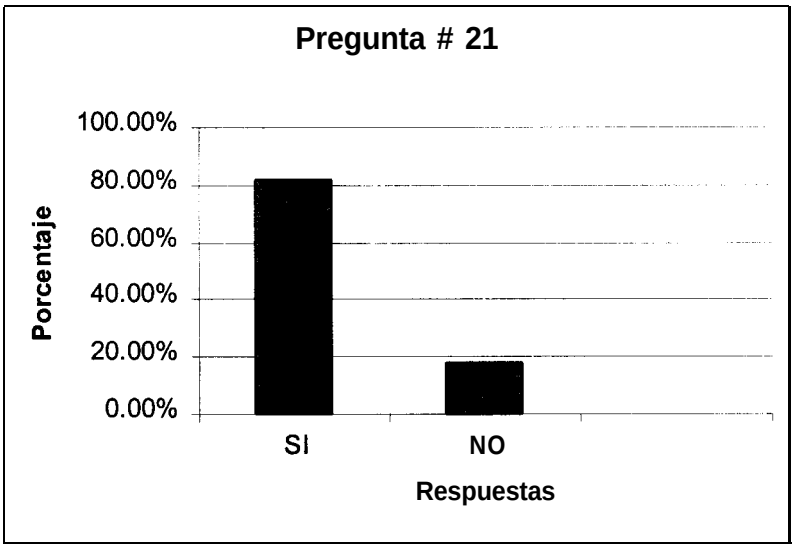


Vemos que las conexiones dedicadas son las más populares (50%) y el resto se dividen las otras categorías, aunque es de esperar que la conexión propia sea la menor (18.18%) por el costo de la misma.

4.20. Pregunta # 21:

Esta pregunta es sencilla, indaga si la compañía tiene o no página Web en Internet, lo cual respondieron:

Gráfico 4.15. Pregunta # 21.

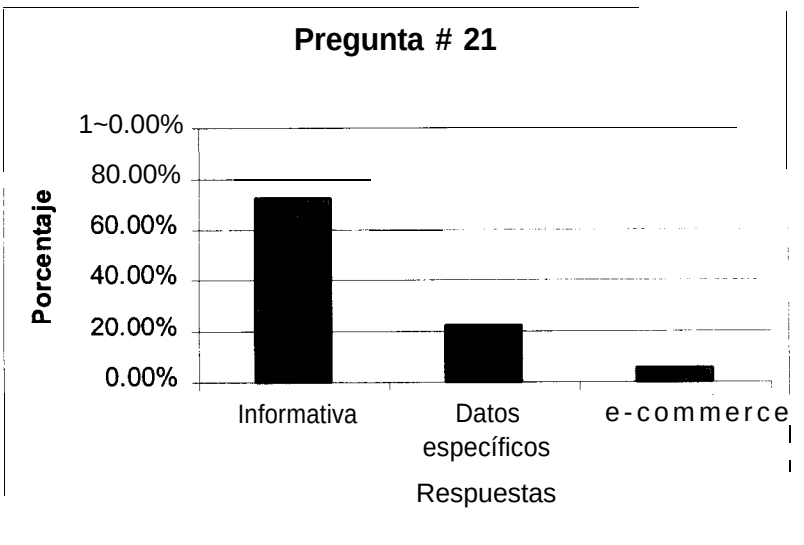


Vemos que la mayoría de las empresas (81.82%) tienen página Web publicada, lo cual es un buen indicativo, pero vamos a ver el uso que le dan en las preguntas subsiguientes.

4.21. Pregunta # 22:

Esta pregunta se refiere al tipo de página Web que tiene la empresa, y tenía las siguientes opciones de respuesta: Informativa, consulta de datos específicos, secundarios, y de comercio electrónico. Los resultados fueron:

Gráfico 4.16. Pregunta # 22



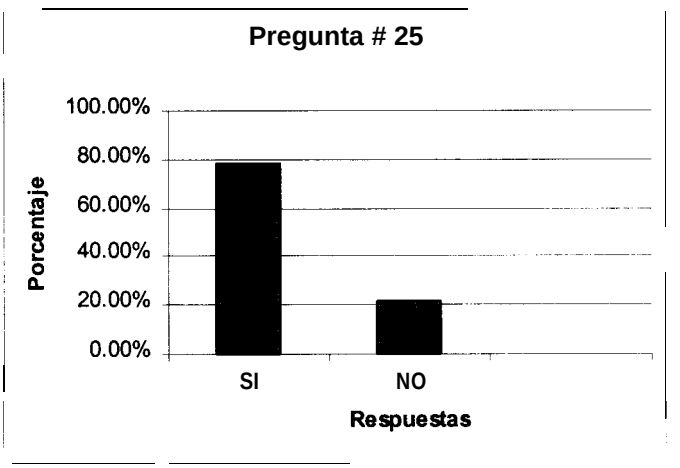
Vemos que la mayoría de las empresas utilizan páginas web simplemente para informar sobre sus actividades (72.22%), pero casi nadie la usa para dar servicios realmente, como los de consultas específicas y comercio electrónico.

Las preguntas 23 y 24 tratan sobre las páginas Web que dan servicios en línea, las cuales son muy pocas y muy poco representativas en nuestra muestra, que dar un análisis de ellas sería inapropiado, dado que no hay los suficientes datos.

4.22. Pregunta # 25:

Las tres últimas preguntas del cuestionario, empezando con ésta, tratan sobre la seguridad de los sistemas de información y la visión que tienen en las empresas sobre la seguridad, la pregunta trataba sobre si creían que la empresa del encuestado estaba protegida contra intrusos (por el web, en una PC descuidada, etc.), los resultados fueron:

Gráfico 4.17. Pregunta # 25

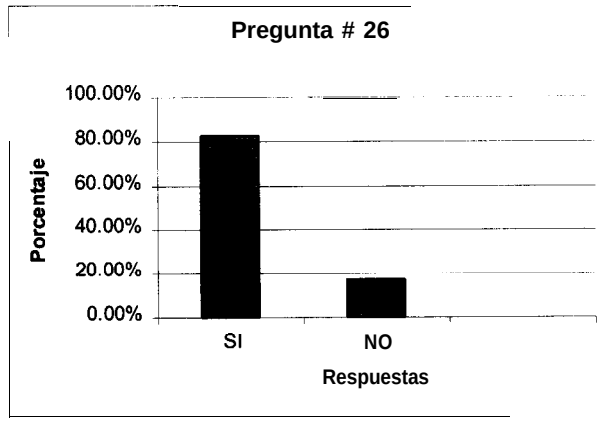


Aparentemente la gran mayoría de empresas (78.26%) creen estar protegidos contra intrusos, pero cuando se les hace las siguientes preguntas sobre parámetros específicos de sus seguridades, las respuestas no tienen la misma contundencia que en esta pregunta.

4.23. Pregunta # 26:

La pregunta 26 indaga sobre la creación de respaldos periódicos de datos críticos para la empresa, la respuesta fue:

Gráfico 4.16. Pregunta # 26.



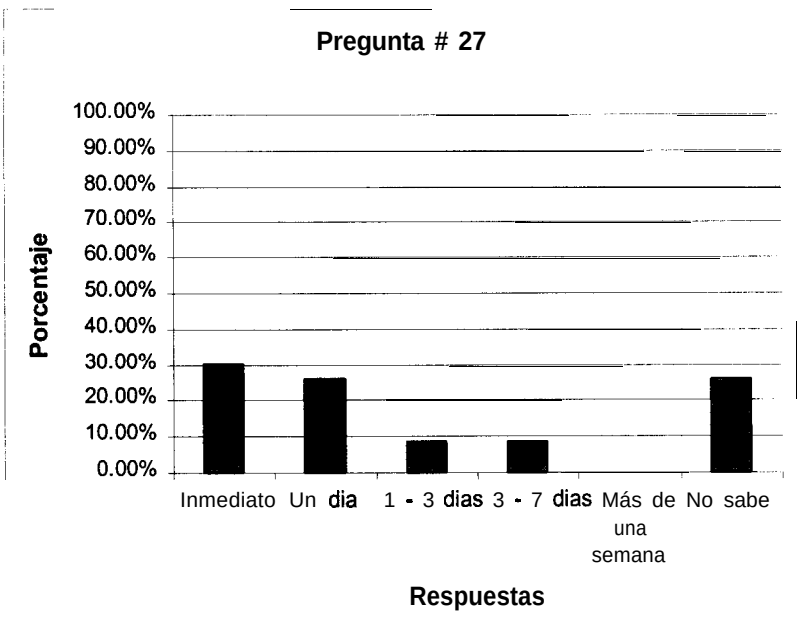
Vemos que la gran mayoría de empresas (82.61%) si realizan respaldos de sus datos críticos, lo cual es una buena señal, aunque faltó un alcance a esta pregunta, ya que en la experiencia personal durante la encuesta, hay personas que respaldan sus archivos de un

disco duro en el mismo disco pero en otra ubicación, la pregunta hubiese sido si realizaba respaldos en medios magnéticos u ópticos. A pesar de la gran cantidad de empresas que si hacen “respaldos”, veamos el resultado de la siguiente y última pregunta.

4.24. Pregunta # 27

Esta pregunta indagaba sobre el tiempo que una empresa demoraría en recuperar sus datos (parte su plan de contingencia) en caso de un desastre mayor, las respuestas fueron:

Gráfico 4.19. Pregunta # 27.



Vemos que a pesar de que la gran mayoría de los encuestados decían tener un sistema de seguridad sobre los datos, la diferencia no es tan grande al momento de preguntarles un parámetro específico de todo plan de seguridad que es el tiempo de recuperación, la respuesta Inmediato (30.43%) casi está en empate con la de No sabe y un día (26.09%), aunque en honor a la verdad solo en tres lugares me dieron un dato específico, en la mayoría de lugares usaban el término “calculo que x días”, lo cual no es tan específico.

CAPÍTULO 5

5. ANÁLISIS MULTIVARIADO

En el capítulo anterior hicimos un análisis univariado de los resultados de la encuesta, lo cual nos da información sobre cada una de las preguntas, pero no nos da información sobre la interdependencia de las variables representadas por cada una de las preguntas, ni tampoco de la relación entre cada una de las modalidades (u opciones) de cada variable.

En el siguiente capítulo aplicaremos el Análisis de correspondencias múltiples, la cual es una técnica de interdependencia para analizar este tipo de encuestas que tienen variables nominales (preguntas donde el encuestado puede responder una opción, pero éstas no tienen un orden

específico), la cual es una técnica que se basa en el Análisis de Factores y en el análisis de componentes principales.

A continuación daremos una breve descripción de cada una de las técnicas arriba mencionadas, y después las aplicaremos para nuestro caso particular de las encuestas sobre el uso y manejo de los sistemas de información.

5.1. **Análisis de componentes principales.**

A un análisis de componentes principales le concierne explicar la estructura de varianzas y covarianzas de un conjunto de variables a través de unas pocas combinaciones lineales de estas variables. Sus objetivos generales son:

- Reducción de datos, e
- Interpretación de datos.

En una muestra de n individuos y p variables observadas, se genera un vector p-variado, que contiene la información de la muestra en p variables aleatorias. Pues bien, mucha de la variabilidad contenida en estas p variables puede ser contabilizada por un número pequeño k, donde obviamente $k < p$, de componentes principales, con el

interesante resultado de que en las k nuevas variables existe casi tanta información como en las p variables originales.

Las k componentes principales pueden entonces reemplazar a las iniciales p variables y el conjunto original de datos se verá reducido de una matriz de n individuos por p variables, a un conjunto de n individuos por k componentes principales.

Un análisis de componentes principales algunas veces revela relaciones que no eran previamente sospechadas y permiten interpretaciones que no resultan de un análisis univariado o bivariado.

El análisis de componentes principales por lo general no son más que un medio para otras investigaciones, porque ellas frecuentemente sirven como intermediario para otros análisis como el análisis de factores y el análisis de correspondencias múltiples.

Algebraicamente hablando, las componentes principales son combinaciones lineales particulares de las p variables aleatorias X_1, X_2, ... x,. Geométricamente, estas combinaciones lineales representan la selección de un nuevo sistema de coordenadas obtenido de rotar el sistema original con $X_1, X_2, ... X_p$ como los ejes

coordenados. Los nuevos ejes representan la dirección con máxima variabilidad y proveen una descripción simple y más parsimoniosa de la estructura de covarianzas.

Sea el vector aleatorio X = [$X_1, X_2, \ldots . X,$] con matriz de covarianzas Σ, las componentes principales Y_i tendrán la forma de las siguientes combinaciones lineales:

$$Y_1 = a_1^{'}X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$Y_2 = a_2^{'}X = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

$$Y_p = a_p^{'}X = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$$

Entonces, por los teoremas conocidos de varianzas y covarianzas, tenemos que:

$$Var(Y_i) = a_i^{'}\Sigma a_i; i = 1,2,\ldots, p$$
$$Cov(Y_i, Y_k) = a_i^{'}\Sigma a_i; i, k = 1,2,\ldots, p$$

Las componentes principales son aquellas combinaciones lineales no correlacionadas $Y_1, Y_2, \ldots . Y_p$ cuyas varianzas, descritas en la fórmula anterior son máximas, es decir lo más grandes posibles.

El siguiente teorema, del cual no daremos la prueba da los valores de a que cumplen esta condición:

Sea Σ la matriz de covarianzas asociada con el vector aleatorio X' = [X_1, X_2, ..., X,]. Tenga Σ el par de valores-vectores propios (λ_1, e_1), (λ_2, e_2), . . . , (λ_p, e_p) donde $\lambda_1 \geq \lambda_2 \geq . . . \geq \lambda_p \geq 0$. Entonces la i-ésima componente principal está dada por

$$Y_i = e_i^{'}X = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p, i = 1,2,..., p$$

Con estas elecciones:

$$Var(Y_i) = e_i^{'}\Sigma e_i = \lambda_i$$
$$Cov(Y_i, Y_k) = e_i^{'}\Sigma e_k = 0, i \neq k$$

Se puede probar también que:

Proporción de varianza total de la muestra explicada por la k-ésima componente principal = $\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}, k = 1,2,..., p$

Se ha visto en casos particulares que la mayoría (80 a 90%) de la varianza total, para valores de p muy grandes, pueden ser atribuidos a las primeras una, dos o tres componentes, entonces estas componentes pueden "reemplazar" las p variables originales sin mucha pérdida de información.

Cada componente del vector de coeficientes $e_i' = [e_{i1}, \ldots . e_{ik}, \ldots \ldots . e_{ip}]$ también amerita inspección. La magnitud de e_{ik} mide la importancia de la k-ésima variable en la i-ésima componente principal, independientemente de otras variables. En particular, e_{ik} es proporcional al coeficiente de correlación entre Y_i y X_k.

Vemos que el análisis de componente principales requieren un nivel de conocimiento medio en álgebra lineal, ya que precisamente son los valores y vectores propios de la matriz de covarianzas los que nos dan los coeficientes y los porcentajes de explicación de las componentes principales.

5.2. Análisis de Factores.

El análisis de factores puede ser considerado una extensión del análisis de componentes principales. Ambos pueden ser vistos como intentos para aproximar la matriz de covarianzas Σ. Sin embargo, la aproximación basada en el modelo del análisis de factores es más elaborada. La pregunta primaria en el análisis de factores es si los datos son consistentes con un estructura prescrita. A continuación, el modelo ortogonal de factores.

Sea el vector aleatorio observable X, con p componentes, y vector de medias μ y matriz de covarianzas Σ. El modelo de factores postula que X es linealmente dependiente de unas pocas variables aleatorias no observables $F_1, F_2, \ldots . F_m$, llamados factores comunes, y p fuentes adicionales de variación $\varepsilon_1, \varepsilon_2, \ldots . \varepsilon_p$, llamadas errores o, algunas veces, factores específicos. En particular, el modelo del análisis de factores es:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p \end{aligned}$$

o, en notación de matrices:

$$\underset{(px1)}{X - \mu} = \underset{(pxm)}{L} \cdot \underset{(mx1)}{F} + \underset{(px1)}{\varepsilon}$$

El coeficiente l_{ij} es llamado la carga de la i-ésima variable en el j-ésimo factor, por lo tanto la matriz L es llamada la matriz de factores de carga. Notesé que el i-ésimo factor específico ε_i esta asociado solamente con la i-ésima respuesta X_i. Las p desviaciones $X_1-\mu_1, X_2-\mu_2, \ldots, X_p-\mu_p$ son expresadas en términos de p + m variables aleatorias $F_1, F_2, \ldots, F_m, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ las cuales no son observables. Con tantas cantidades no observables, una verificación directa del modelo de factores a partir del vector X no tiene esperanzas. Sin embargo, con algunos supuestos adicionales acerca de los vectores aleatorios F y ε, el modelo de factores implica ciertas relaciones de covarianzas que pueden ser analizadas.

Los supuestos a realizar son los siguientes:

- F y ε son independientes.
- E(F) = 0, Cov (F) = I
- E(E) = 0, Cov (E) = ψ.

Donde I es la matriz identidad de mxm y ψ es una matriz diagonal, lo cual significa que las F's son variables aleatorias de media 0 y varianza 1 no correlacionadas, y los E'S son variables aleatorias no correlacionadas de media 0 y varianza ψ_{ii}, i = 1, 2, ... p.

Con estos resultados, la estructura de covarianzas para el modelo ortogonal de factores es:

- Cov (X) = $LL^T + \psi$, 0 lo que es lo mismo:
  - Var (X_i) = $l_{i1}^2 + \ldots + l_{im}^2 + \psi_i$
  - Cov (X_i, X_k) = $l_{i1}l_{k1} + \ldots + l_{im}l_{km}$
  - Cov (X, F) = L, o lo que es lo mismo: Cov (Xi, F_j) = l_{ij}

La porción de la varianza de la i-ésima variable contribuida por los m factores comunes es llamada la i-ésima comunalidad. Esa porción de Var(X_i) = σ_{ii} debido al factor específico es llamada a veces la varianza única 0 específica. Denotando la i-ésima comunalidad por h_i^2, podemos ver que:

$$\underbrace{\sigma_{ii}}_{Var(X_i)} = \underbrace{l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2}_{comunalidad} + \underbrace{\psi_i}_{varianza\cdot especifica}$$

o: $h_i^2 = l_{i1}^2 + l_{i2}^2 + ... + l_{im}^2$ y $\sigma_{ii} = h_i^2 + \psi i, i = 1,2, ... p$

La i-ésima comunalidad es la suma de los cuadrados de las cargas de la i-ésima variable en los m factores comunes.

La estimación de las matrices L y ψ se pueden dar por dos métodos, uno matemático y otro iterativo, el matemático es precisamente por medio de componentes principales y es el que vamos a tratar aquí, y el otro que por lo general tiene menos uso que el matemático que es el de máxima verosimilitud.

La teoría para llegar a la siguiente respuesta es extensa, aquí ilustraremos solo el resultado de la solución de componentes principales para el modelo de factores:

El análisis de factores por componentes principales de la matriz de covarianzas Σ esta dado en términos de sus pares de valores-vectores propios $(\lambda_1, e_1), (\lambda_2, e_2), ... (\lambda_p, e_p)$ donde $\lambda_1 \geq \lambda_2 \geq . . . \geq \lambda_p$. Sea m < p el número de factores comunes, entonces la matriz de factores de carga $L = \{l_{ij}\}$ esta dada por:

$$L = \left[\sqrt{\lambda_1}\cdot e_1 \quad \sqrt{\lambda_2}\cdot e_2 \quad \cdots \quad \sqrt{\lambda_m}\cdot e_m\right]$$

La matriz de varianzas específicas está provista por los elementos de la diagonal de la matriz $\Sigma - LL^T$, y sus demás elementos 0, así que:

$$\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2$$

De la misma manera que en componentes principales, tenemos que:

(Proporción de la varianza total dado por el j-ésimo factor) =

$$\frac{\lambda_j}{\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp}}$$

Vemos que el análisis de factores es muy parecido al de componentes principales, la diferencia es que los factores son nuevas variables que explican a las variables originales, la diferencia con la regresión multivariada es que estas nuevas variables no son observables; mientras que las componentes principales son nuevas variables que son explicadas en términos de las variables originales. El análisis de factores es una técnica estadística de interdependencia muy poderosa en muchos campos del conocimiento humano:

psicología, marketing, sociología, biología, etc. La limitación de los factores es que la teoría antes descrita solo funciona con variables cuantitativas, medibles, lo cual es una limitación para la aplicación en la explicación de encuestas como la de esta tesis. A continuación veremos una técnica que complementa el análisis de factores para poder utilizarlo con variables nominales (cualitativas).

5.3. Análisis de correspondencias múltiples.

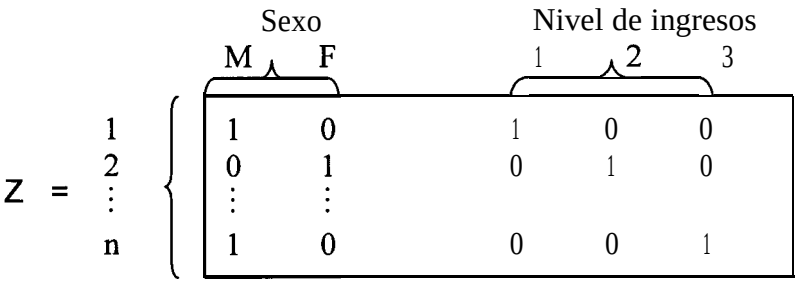
El análisis de correspondencias múltiples es una técnica multivariada muy útil en la investigación comercial y de marketing. Generalmente, en una encuesta se formulan preguntas cerradas. Cada una tiene diversas modalidades de respuesta excluyentes, y el encuestado debe elegir una. Por ejemplo, preguntas como:

Sexo: Masculino 0 Femenino 0

Nivel de ingresos: menos de $150 0 entre $151 y $300 0

mayor a $300 0

Pueden analizarse por medio de este análisis. La información proporcionada por esta encuesta se recoge en una tabla disyuntiva del tipo:



Como vemos, se ha transformado cada modalidad u opción de la pregunta en una variable aleatoria binomial que puede tomar el valor de l(suceso) o 0(falla), dado si la opción ha sido o no elegida respectivamente. La matriz Z será una matriz de ceros y unos. Una tabla disyuntiva completa Z queda descrita mediante:

. Un conjunto de individuos (el tamaño de la muestra), generalmente notado por n.

. Un conjunto de variables o preguntas $J_1, J_2, \ldots J_Q$.

. Un conjunto de modalidades u opciones a cada pregunta 1, $2, \ldots m_k$.

El análisis de correspondencias múltiples está diseñado para analizar tablas disyuntivas completas. Son tablas de variables cualitativas; sin embargo siempre es posible transformar una variable métrica en cualitativa, dividiendo su intervalo de variación en clases de equivalencia sucesivas, como hicimos con el nivel de ingreso.

El objetivo del análisis es obtener una representación simultánea, en un espacio de dimensión reducida R^q, de

- Las alternativas (modalidades) de todas las preguntas.
- Los individuos.

Se trata de estudiar las relaciones entre todas las modalidades, no entre las preguntas. Este análisis se podría realizar utilizando factores codificando cada opción de pregunta con un número, por ejemplo masculino = 1 y femenino = 2, nivel de ingreso mayor a $ 300 = 3, y tratando a las variables como cuantitativas, el problema es que los factores solo analizan las relaciones entre variables, pero las correspondencias múltiples, además de esto, analiza las relaciones entre cada nivel, opción o modalidad de las variables, lo cual, como veremos ya mismo, dará unos resultados muy importantes y

significativos que, utilizando factores, no se hubiese podido llegar a ellos.

Vamos a transformar la matriz Z de unos y ceros a una matriz a la cual podamos calcularle los factores, la diferencia es que en esta matriz que vamos a obtener cada modalidad de una pregunta será considerada una variable.

La técnica, cuya justificación es muy extensa y complicada, consiste en obtener la matriz V de la siguiente manera:

$$V = \frac{1}{Q} D^{-1} B$$

donde Q es el número de preguntas, B es igual a $Z^T Z$, y se la conoce como matriz de Burt. Es una matriz simétrica formada por Q^2 bloques. Esta matriz es la más importante, ya que es muy fácil observar que:

- Los bloques de la diagonal son tablas diagonales que cruzan una pregunta con ella misma (ya que $Z^T Z$ es una forma

cuadrática). Los elementos de la diagonal son los efectivos de cada modalidad.

- Los bloques fuera de la diagonal son verdaderas tablas de contingencia obtenidas cruzando las preguntas de dos en dos. Sus elementos son las frecuencias de asociación de las dos modalidades correspondientes.

Vemos la capacidad de la matriz de Burt para transformar la matriz de 0 y 1 a una matriz numérica que mide las frecuencias de asociación de todas las modalidades.

La matriz D es una matriz diagonal cuyos elementos diagonales son los de la matriz de Burt, los efectivos de cada modalidad. El resto de los elementos son ceros.

A la matriz V se procede a obtener los factores y hacer el análisis correspondiente, como tradicionalmente se realiza con los factores.

Vemos que el análisis de correspondencias múltiples realiza una pequeña transformación de los datos antes de recurrir a los factores, y al análisis de componente principales, el cual puede ser considerado como el “padre” de las técnicas de interdependencia.

A continuación realizaremos la aplicación de todas las técnicas antes descritas para nuestro caso específico.

5.4. Análisis Multivariado de la encuesta sobre sistemas de información.

Como sabemos, la encuesta esta dividida en 26 preguntas cualitativas, además hay preguntas en que el encuestado podía elegir más de una respuesta, las cuales se las dividió en tantas preguntas como opciones para poder ajustarlas al modelo del análisis de correspondencias múltiples, lo cual generó 66 variables.

Como ha de recordar el lector, la encuesta se la codificó dándolo un número a cada uno de las opciones de las preguntas para realizar el análisis univariado, lo cual no es compatible con el análisis a realizarse, por lo cual una vez concluido el análisis univariado se procedió a realizar la adecuación de la matriz de datos y transformarla a una matriz de 0 y 1 por medio de un algoritmo realizado en Excel que tomaba el número de opciones de cada pregunta y dividía a la pregunta en tantas variables de unos y ceros como modalidades tuvieran las variables, y a la modalidad elegida en

la observación le asignaba un 1, y al resto 0. Las modalidades que nunca fueron elegidas en la encuesta por los encuestados fueron eliminadas. Al final de toda esta transformación resultó una matriz Z de 30 filas (el tamaño de la muestra) por 66 columnas (66 modalidades de las preguntas).

Después, con la ayuda de Excel, ya que no teníamos disponibilidad de un paquete estadístico que realice todos los pasos descritos automáticamente, ejecutamos un algoritmo que calcula la matriz $V = \frac{1}{Q} D^{-1} B$.

Una vez calculada la matriz V, se procedió a realizar el análisis de factores de dicha matriz, ahora si con la ayuda de un paquete estadístico al que se importó la matriz resultante en Excel, a continuación presentaremos los resultados obtenidos y su respectivo análisis:

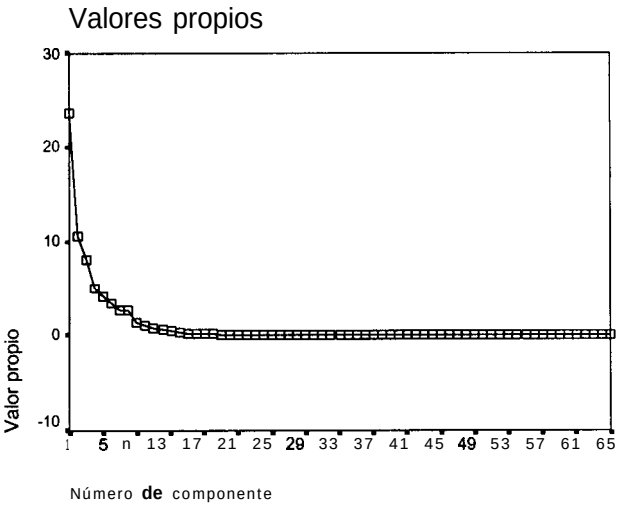
La salida del paquete estadístico es bastante extensa, así que presentaremos lo netamente esencial:

Tabla 5.1. Varianza total explicada por los factores

Componente	Valores propios iniciales			Sumas de extracción de los cuadrados de las cargas		
	Total	% de varianza	% acumulativo	Total	% de varianza	% acumulativo
1	23.852	36.696	36.696	23.852	36.696	36.696
2	10.539	16.214	52.910	10.539	16.214	52.910
3	8.082	12.434	65.343	8.082	12.434	65.343
4	4.947	7.611	72.954	4.947	7.611	72.954
5	4.064	6.252	79.207	4.064	6.252	79.207
6	3.340	5.138	84.345	3.340	5.138	84.345
7	2.685	4.131	88.476	2.685	4.131	88.476
8	2.575	3.962	92.438	2.575	3.962	92.438
9	1.379	2.122	94.560	1.379	2.122	94.560
10	1.008	1.550	96.110	1.008	1.550	96.110
11	.717	1.104	97.214			
12	.530	.816	98.030			
13	.436	.671	98.701			
14	.250	.385	99.086			
15	.205	.315	99.400			
16	.136	.209	99.609			
17	9.629E-02	.148	99.757			
18	8.419E-02	.130	99.887			
19	3.558E-02	5.474E-02	99.942			
20	2.374E-02	3.652E-02	99.978			
21	8.508E-03	1.309E-02	99.991			
22	3.207E-03	4.934E-03	99.996			
23	1.653E-03	2.543E-03	99.999			
24	7.663E-04	1.179E-03	100.000			

Método de extracción: Componentes principales

Gráfico 5.1. Valores propios.



Esta primera tabla que arroja el análisis de factores no es más que el análisis de componentes principales, el cual ayuda a determinar el número de factores con el que vamos a trabajar. Por lo general el paquete estadístico toma el criterio del valor propio mayor a uno para considerar el número de factores, pero siempre es interesante hacer este análisis.

Vemos que de las 65 variables analizadas (una variable se eliminó ya que al hacer la operación de cálculo de la matriz V y el pivoteo su fila y columna estaban llenas de ceros), podemos obtener 24 nuevos factores que expliquen el 100% de la varianza contenida en la matriz original V, pero solo 10 de estos nuevos factores tienen valores

propios mayores a uno, con lo cual y siguiendo el criterio del paquete estadístico deberíamos trabajar con 10 factores, pero analizando el gráfico de los valores propios, vemos que el último factor propio significativo antes que la curva tome una forma asintótica alrededor del 0 es el factor # 8, con 10 factores tendríamos el 96.11% de la varianza explicada en los nuevos factores, y con 8 factores el 92.44%.

El lector podría decir que para que tomar tan poco factores si con 24 podemos explicar el 100%, pues bien, esta técnica a utilizar, como ya lo describimos anteriormente sirve para:

- Reducción de datos, e
- Interpretación de los mismos.

La utilidad de reducción de datos se la dejo de utilizar hace tiempo ya que el poder computacional de que se dispone actualmente no requiere una reducción de datos, las computadoras actuales pueden manejar un gran volumen de datos, pero la segunda utilidad de interpretación no busca obtener la mayor cantidad de varianza explicada sino una correcta interpretación de los mismos, y a veces

entre menos factores se tengan es mejor, sin renunciar a la cantidad de varianza explicada para ayudar a una mejor interpretación.

Vemos además que mucha de la variabilidad (53%) se concentra alrededor de las dos primeras componentes, para ayudar a una mejor interpretación, se debe distribuir mejor esta variabilidad, lo cual se logra por medio de una rotación ortogonal.

La rotación de factores no es más que rotar ortogonalmente el nuevo sistema coordenado de factores con el fin de obtener una mejor distribución de las comunalidades sin afectar a la varianza explicada, uno de los métodos más usados de rotación es la rotación Varimax.

A continuación, procederemos a correr el análisis de factores, pero ahora con 8 factores y una rotación varimax, los resultados fueron:.

Tabla 5.2. Varianza total explicada con rotación

Componentes	Sumas rotadas de los cuadrados de las cargas		
	Total	% de varianza	% acumulativo
1	15.936	24.516	24.516
2	12.512	19.250	43.766
3	10.543	16.220	59.986
4	5.201	8.001	67.987
5	5.020	7.724	75.711
6	3.709	5.706	81.417
7	3.597	5.533	86.950
8	3.567	5.488	92.438

Método de extracción: Análisis de componentes principales.

Vemos aquí que la rotación permitió redistribuir mejor los porcentajes de explicación sin afectar al porcentaje de explicación que dan las 8 componentes. A continuación veremos la construcción de los factores y su análisis.

Tabla 5.3: Matriz de componentes (Sin rotación)

Variables	Factores							
	1	2	3	4	5	6	7	8
X6_2	.940		.153				.101	
X9_1	.935	-.130	.125			.145		.147
X17_3	.926		.151	-.247		.106	-.116	
X16_5	.910		.217	-.190	.147			
X13_3	.870		-.255	-.112	.242		-.215	.160
X27_1	.867	-.164	.105	-.190		.300	.193	
X10_5	.849		-.397	-.131	.202			.119
X8_1	.846		-.375	.216				.243
X26_1	.839		.228	.388	.124	-.153		
X7_1	.825	.334	.132			-.318	.168	
X27_6	-.822	.340		-.220	.294	.155	.118	.123
X11_3	.816	.343	.118	-.287		-.257	.132	
X21_1	.786	.340	.144	.386		.222	-.175	
X12_1	.779	.130	-.420	.118	.203	-.146		
X20_3	-.778			.186	.489	.250		
X18_2	.767	-.126	.115	-.293		.409	.281	-.147
X26_2	-.765	.353		-.291	.194	.234		.232
X18_3	.747		-.457	-.136	.175	.237	.216	
X24_1	.721	-.121	.498	-.270			-.137	.233
X14_1	.717		-.519			.204	.301	
X20_2	.715		.471	.122	-.183		-.227	.262
X25_1	.706		.221	.202	.398	.351	-.230	
X22_2	.706				.174	-.265	-.316	.427
X10_1	.688	.670	.115		-.131			
X11_1	.688	.670	.115		-.131			
X11_4	.684		.398	.384	-.323		.165	.157
X4_2	.679	-.137	-.400		.411		-.115	-.129
X12_3	-.652	.635		-.248	.146		-.193	
X6_3	-.635	-.362			.579	.105	.220	.152
X21_2	-.621	-.347	.103		.576		.280	.110
X25_2	-.619	.344				-.332	.473	.189
X20_1	.573	.200	-.508	.172			.330	-.255
X12_4	-.528	-.433		.403	.192		.244	.368
X16_4	-.519	.780	-.104				-.102	
X10_2	.504	.759	.179	.124	-.178			

X6_1	-.553	.752					-.131	
X15_1	.593	.707	.162	-.192	.162			-.134
X18_1	.593	.707	.162	-.192	.162			-.134
X17_1	-.576	.682	-.188	-.134		.184		.106
X1_3	.428	.681	-.470	-.105	.145			
X13_2	-.433	.653	.427		-.122		.210	-.297
X14_2		.637	.603	-.196	.100	-.195	-.290	
X3_2		-.595	.517	.278		-.314		-.202
X4_3	-.269	-.583	-.264	-.225	-.524	-.101	-.275	-.273
X23_4	.244	-.102	.761	-.284	-.243	.336	.162	-.149
X22_4	.244	-.102	.761	-.284	-.243	.336	.162	-.149
X1_2	-.280	-.369	.730		.367			
X12_2	.544		.697		-.207		.201	
X5_1	-.157	.455	.608	.295	.257	-.134		.447
X8_2	-.387	.549	.569	-.178				-.248
X5_2	.517		-.566	-.169	.214	.300	.233	-.330
X4_1	-.438	.506	.535			.100	.284	.295
X22_1	.111	.480	-.346	.679		.214		-.233
X24_2		.535	-.352	.633		.214		-.270
X3_1	.271	.344	-.484	-.609	.175			
X19_2	-.360	.479	-.129	-.601		-.296	.257	.149
X19_1	.460	.122	.296	.593	.466	.285		.109
X27_2			.294	.548	.244	-.524		-.284
X27_4		.190	-.249	.547	-.504		.261	.325
X1_1	-.245	-.493	-.309	.144	-.714		-.158	
X7_3	-.444	.477				.582	-.327	
X3_3	-.424	.399	-.134	.329	-.296	.531	-.174	.249
X17_2	-.168	.247	.210	.413	.170	-.518	.126	-.511
X27_3	.259	.130		-.145	.230	-.265	-.693	.204
X11_2	.372	.363	-.237		-.301	-.353	.367	.461

Método de extracción: Análisis de componentes principales.

8 componentes extraídas.

El lector habrá podido notar que la notación a usar una vez hecho el análisis de correspondencias múltiples es XI_J para denotar la j-ésima alternativa o modalidad de la i-ésima pregunta o variable.

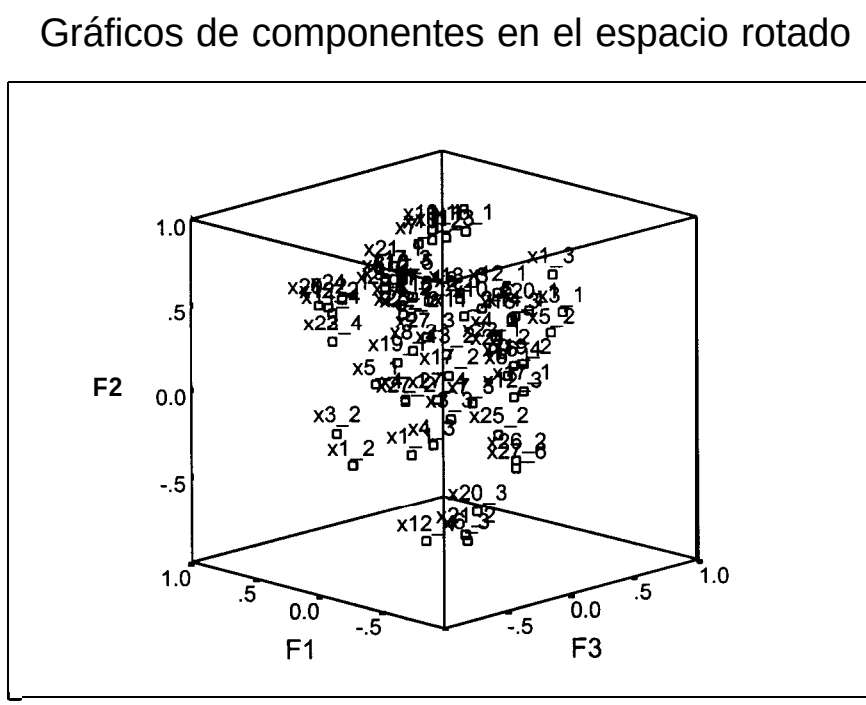
Los espacios en blanco que vemos en la tabla anterior son porque le pedimos al paquete estadístico que oculte las cargas menores en valor absoluto a 0.1, ya que nos interesen las cargas mayores y significativamente distintas de cero. La matriz anterior es la matriz de cargas sin rotación, a continuación veremos como la rotación concentra las mayores cargas alrededor de los primeros factores.

Tabla 5.4: Matriz de componentes (Con rotación)

Variables	Factores							
	1	2	3	4	5	6	7	8
X12_3	-.862	.166	-.313	.171			.148	-.167
X27_6	-.832	-.255	-.260	.307		.149	.172	
X17_1	-.814	.173	-.189	.113	.232	.152	.266	.129
X11_4	.812	.305		.112	.142	.108		.350
X20_2	.809	.379	-.182	.115		-.223		
X26_2	-.806	-.213	-.238	.259		.128	.346	
X6_1	-.801	.287	-.294	.114	.273			
X9_1	.799	.366	.344	.108			.208	
X16_4	-.796	.325	-.252	.104	.300			
X6_2	.789	.415	.336	.110				.119
X24_1	.721	.402		.177	-.349	-.142	.246	
X26_1	.721	.367	.263	.285	.106	-.109	-.340	
X27_1	.706	.323	.465		-.150	.189	.251	
X19_2	-.705	.290			-.475	.112	.151	.266
X12_2	.694	.392	-.208	.154	-.209	.370		
X16_5	.689	.436	.378	.215	-.264			-.133
X17_3	.681	.463	.354	.149	-.225	-.182	.229	-.156
X25_2	-.665		-.232	.113	-.195	.247	-.148	.496
X18_2	.597	.311	.473	.103	-.177	.352	.312	-.175
X3_2	.574	-.312	-.271		-.215		-.532	-.135
X25_1	.569	.279	.255	.489	.244	-.206		-.336
X15_1		.874	.209	.373				-.112

X18_1		.874	.209	.373				-.112
X10_2		.859		.239	.271			.176
X10_1	.229	.853	.146	.265	.210			.215
X11_1	.229	.853	.146	.265	.210			.215
X12_4		-.842	-.192	.221	.110			.243
X6_3	-.286	-.806	-.111	.397	-.165			-.153
X21_2	-.287	-.766	-.118	.395	-.230		-.185	-.108
X11_3	.327	.758	.345	.198	-.336			
X7_1	.444	.712	.256	.165	-.124		-.193	.313
X14_2	-.212	.638	-.544	.317	-.110		-.138	-.268
X1_3	-.292	.618	.567	.198	.111	-.166		.202
X21_1	.565	.581	.176	.244	.477	-.113		
X20_3	-.526	-.557	-.267	.407	.189	.104		-.279
X5_2		.137	.908			.145	.120	-.171
X18_3	.328	.182	.861				.204	
X14_1	.281	.212	.852		.107		.149	.187
X20_1	.100	.300	.777		.190	.101	-.195	.211
X10_5	.432	.251	.739			-.343	.198	
X4_2	.327	.101	.734			-.372	-.134	-.161
X12_1	.285	.375	.716			-.320	-.149	
X4_1	-.299	.111	-.607	.512		.343		.238
X8_1	.470	.328	.601		.153	-.325		.384
X8_2	-.360	.405	-.600	.171		.358		-.268
X13_3	.456	.362	.581	.112	-.114	-.501	.123	
X3_1	-.368	.408	.571		-.373	-.133	.248	
X1_2	.234	-.393	-.496	.382	-.267	.180	-.151	-.405
X4_3		-.309		-.910	-.118			-.142
X1_1	.109	-.368	-.154	-.811	.202		.134	.257
X5_1		.144	-.577	.732			-.126	.232
X19_1	.487		.123	.717	.423		-.146	
X24_2	-.262	.207	.196		.860		-.213	
X22_1	-.124	.212	.266		.854		-.219	.112
X3_3	-.344		-.294		.715		.424	.150
X7_3	-.476		-.251	.124	.591		.410	-.244
X27_3		.308				-.763		-.223
X22_2	.474	.292	.210	.147	-.240	-.658	.116	.126
X23_4	.513	.254	-.359		-.236	.550	.227	-.274
X22_4	.513	.254	-.359		-.236	.550	.227	-.274
X13_2	-.475	.380	-.457	.209	.110	.535	-.144	
X17_2	-.179	.138	-.138	.103	.103	.125	-.877	
X27_2	.119		-.184	.178			-.842	

Gráfico 5.2. Componentes en el espacio rotado



A simple vista el gráfico es un poco confuso por el número de variables, pero si se lo analiza al detalle con un acercamiento bastante alto puede dar conclusiones interesantes.

En el ler factor tenemos 3 cargas negativas altas en las 3ª alternativa de la pregunta 12, la 6ª alternativa de la pregunta 27, y la 1ª opción de la pregunta 17 (X12-3, X27-6, y XI 7_1), lo cual se lo puede interpretar como que las personas que respondieron que usaban sistemas generales adaptados para su empresa, no sabían como recuperarse de un desastre a su información, y además

XI 1_2		.383	.196					.832
X 2 7 4				-.100	.526	.126		.723

Método de extracción: Análisis de componentes principales.

Método de rotación: Varimax con normalización Kaiser.

La rotación convergió en 21 iteraciones.

A continuación procederemos a realizar el análisis de los factores, a partir de dos métodos:

- La contribución de las variables en cada factor, y
- Los puntos cercanos en el espacio F_1-F_2-F_3.

La contribución de las variables en cada factor las podemos analizar de la tabla, así como los puntos cercanos en el plano F_1-F_2-F_3, pero para este último método nos ayudaremos con el gráfico a continuación que ilustra todas las alternativas en el espacio antes mencionado.

usaban reportes impresos para intercambiar información entre departamentos.

A veces el análisis de factores arroja resultados que hasta cierto punto pudieran considerarse obvios, pero los arroja con suficiente evidencia estadística para afirmarlo. Un ejemplo de ellos son dos cargas positivas altas entre las opciones XI 14 (Uso de la red para conexión a Internet), y X20-2 (conexión dedicada a Internet), lo cual puede interpretarse como que las compañías que comparten conexión a Internet en la red, tienen una conexión dedicada, lo cual era de suponerse, pero ahora esta respaldado con análisis estadístico.

El tamaño de la organización no tiene casi representatividad en los resultados excepto la opción 3 (Grande) cuyo punto en el sistema F_1-F_2 esta muy cercano con la variable X14-2 (No realizan auditorias) con lo cual podemos concluir que las empresas grandes no se preocupan de realizar auditorias periódicas de sus sistemas.

Vemos una carga alta positiva, y aparentemente iguales entre las opciones X24-1 y X26-1, lo cual significa que las empresas que tienen conectada su página Web con sus sistemas de información

son las que mejor están defendidas contra desastres, ya que se podrían recuperar de uno de ellos inmediatamente.

Una carga alta en X16-5, X12-2, y XI 7_3 se podría concluir como que las empresas que tienen sus bases de datos en motores de bases en un servidor, con las que mejor manejan las salidas, ya que cualquier departamento puede consultar información, y prefieren usar sistemas desarrollados exclusivamente para la empresa por terceras personas.

Otras cargas altas entre X10-1 y X10-2 implica que las empresas que utilizan Windows 95/98, también prefieren utilizar Windows NT/2000, y una carga alta entre X10-5 y X4_2 implica que las compañías que utilizan computadoras como actividad principal prefieren usar sistemas operativos basados en Unix.

Dos cargas antagónicas, muy parecidas en magnitud, pero con signos distintos entre ellas, dan a notar que las variables X19-2 y X27-1 están inversamente relacionadas, es decir que los que no tienen conexión a Internet son los que menos están preparados para defenderse de un desastre a sus datos críticos.

En la pregunta sobre el uso de los sistemas de información (pregunta # 18) podemos ver que están muy cercanos los puntos X5_2 y X18-3 en el plano F_2-F_3, lo cual nos indica que las empresas que tercerizan o administran sus departamentos de informática por medio del outsourcing, son las que mejor utilizan sus sistemas de información, como es para la planeación estratégica y toma de decisiones.

Las cargas de X10-1 y XI 1_1 son idénticas en todos los ejes coordenados F_i, lo que demuestra que se puede casi asegurar que casi todas las compañías que basan sus redes en sistemas operativos Windows 95/98 solo utilizan la red para compartir archivos e impresoras.

Las cargas de X9-1 y X16-4, las cuales tienen signos cambiados demuestran que las empresas que todavía no utilizan TCP/IP como protocolo de comunicación, siguen trabajando con FoxPro para mantener sus bases de datos.

Este es el análisis más representativo que se puede realizar, el análisis de correspondencias múltiples tiene una capacidad de análisis increíble que pudiéramos seguir analizando y encontrando

relaciones de interdependencia entre las opciones de cada una de las variables.

Se hubiera podido realizar simplemente un análisis de factores utilizando la matriz de datos que se usó para el análisis univariado, pero ahí solo hubiésemos podido encontrar relación entre variables, y no un análisis tan completo y minucioso entre las alternativas como aquí se ha realizado, lo cual demuestra la efectividad del análisis de correspondencias múltiples para este tipo de encuestas, que se pueden dar en marketing, sondeos de opinión, censos, etc.

Además vemos como dos o tres variables obtenidas a partir de 65 nos pueden dar toda la información arriba descritas.

CONCLUSIONES Y RECOMENDACIONES.

A partir de este estudio vemos que el análisis multivariado es una herramienta muy poderosa para analizar conjuntos de datos obtenidos mediante encuestas, utilizando la técnica apropiada.

Como el objetivo de esta tesis era ver el uso actual que se le esta dando a los sistemas de información gerencia1 en el Ecuador, el análisis univariado sólo nos daba información puntual mientras que con el análisis multivariado hemos obtenido como se agrupan las variables alrededor de los factores.

Es interesante también observar el uso de la técnica de correspondencias múltiples para analizar este tipo de encuestas cualitativas, muy útil también en otras áreas del conocimiento.

Como conclusiones de esta tesis podemos rescatar lo siguiente:

1. A los sistemas de información gerencia1 les falta mucho camino por recorrer en el país, ya que muchos empresarios todavía tienen ideas equivocadas sobre el uso de la informática y la información como herramientas de negocios y soporte.

2. Las empresas todavía no tienen un concepto bien desarrollado sobre seguridad e integridad de los datos, y no tienen planes de contingencia y recuperación en caso de un desastre.

3. Se sigue usando el medio impreso para compartir información, lo cual genera multiplicidad del trabajo (es decir, doble trabajo para ingresar la misma información), y es uno de los medios más obsoletos en la actualidad, además de inseguro.

4. El uso de Internet en las organizaciones todavía es para navegación, más no para hacer negocios, comercio electrónico, etc.

5. El tamaño de las empresas no influye en los resultados, lo cual rompe el paradigma de que sólo las empresas grandes invierten en

tecnología, es más, las empresas grandes son las que menos se preocupan de realizar auditorías a sus sistemas de información.

6. Un resultado interesante fue que las empresas que mejor utilizan sus sistemas de información, es decir hasta el nivel más alto de planeación, son las que administran por medio del outsourcing su departamento de informática, lo cual deja entrever que este modelo administrativo realmente está dando resultados a las empresas que lo utilizan.

7. Las redes informáticas en las compañías se utilizan, en la mayoría, para compartir archivos e impresoras y compartir conexión a Internet, muy pocas empresas la utilizan para ejecutar sistemas.

Las recomendaciones que se pueden dar para mejorar este panorama pesimista en la ciudad de Guayaquil son:

1. Concientizar a los empresarios, por medio de las universidades, a invertir más y de manera conveniente en sus sistemas informáticos.

2. Este estudio debería ser el punto de partida antes de tomar cualquier medida sobre el problema que aquí se ilustra, para tomar los

BIBLIOGRAFÍA

1. William Mendenhall, Dennis D. Wackerly, Richard L. Scheaffer, Estadística Matemática con Aplicaciones (4ta Edición, Grupo Editorial Iberoamérica S.A. de C.V., 1990)

2. Amitava Mitra, Fundamentals of Quality Control and Improvement (2da Edición, New Jersey, Prentíce Hall, 1993)

3. Joseph F. Hair, Jr., Rolph E. Anderson, Ronald L. Tatham, William C. Black, Multivariate Data Analysis (5ta edición, Prentice-Hall Inc., 1998)

4. Henry C. Lucas, Jr., Sistemas de información (Ira edición, Mc-Graw Hill Inc, 1987)

5. John E. Freund, Ronald E. Walpole, Estadística Matemática con Aplicaciones (4ta edición, Prentice-Hall Hispanoamericana S.A., 1990)

6. Francisco Azorín, José Luis Sánchez-Crespo, Métodos y aplicaciones del muestreo (2da edición, Alianza Editorial, 1990).

correctivos necesarios, y después de cierto tiempo volver a realizar el estudio para ver si las medidas tomadas dan los frutos deseados.

3. Decía el experto japonés en calidad Ishikawa “Lo que no es medible, no es mejorable”, tal vez antes se creía que la percepción de las personas ante cierto tema no era medible, pero ahora que las técnicas multivariadas dan lo necesario para medir impresiones de las personas, debería trabajarse más en mejorarlas. Que interesante que en la carrera de Ingeniería en Estadística Informática hubiese un estudio de esta categoría, con respecto a la Estadística, para saber que hay que atacar y como promocionar una carrera que tiene mucho futuro en el Ecuador, y después de 10 años volver a realizar el mismo estudio y decir “De aquí partimos y aquí llegamos”.