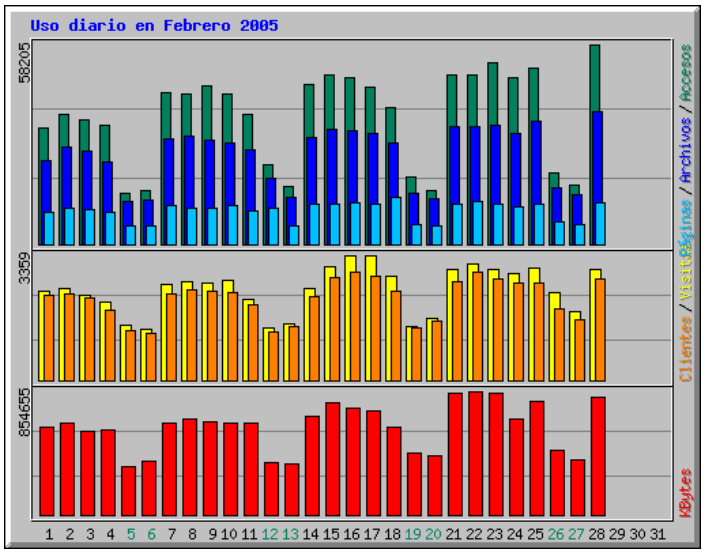
Universidad Rafael Landívar
Tradición Jesuita en Guatemala

UNIDAD DE INVESTIGACIÓN Y PUBLICACIONES
FACULTADES DE QUETZALTENANGO

HERRAMIENTAS ESTADÍSTICAS BÁSICAS PARA LA ELABORACIÓN DE TESIS



MANUAL GUÍA

Ing. Ana Celia de León Sandoval

Quetzaltenango, Diciembre de 2005

De León Sandoval, Ana Celia es catedrática de estadística en la Facultad de Ingeniería, de las Facultades de Quetzaltenango, de la Universidad Rafael Landivar (FQ – URL), donde ha impartido los cursos de Probabilidad y Estadística (Estadística I) y Estadística Inferencial (Estadística II) desde el año 2002. Es Ingeniera Industrial egresada de la Universidad de San Carlos de Guatemala (USAC), Centro Universitario de Occidente (CUNOC), donde publicó su trabajo de tesis, refiriéndose al diseño de las funciones y de la organización de un laboratorio de investigación y desarrollo para una industria de alimentos.

INDICE

CAPITULO **PÁGINA**

I. INTRODUCCIÓN

Las Facultades de Quetzaltenango, de la Universidad Rafael Landivar (FQ – URL), a través de la Unidad de Investigación y Publicaciones (UIP), comprometida con la calidad de las investigaciones universitarias, especialmente con la de los trabajos de tesis, brinda este Manual Guía de Herramientas Estadísticas Básicas, con el objetivo de apoyar a docentes, estudiantes, asesores e investigadores, para que la aplicación de la estadística, lejos de desmeritar su trabajo lo fortalezca.

La presentación se hace de manera breve, sencilla, clara y práctica, para que las personas, con conocimientos básicos sobre estadística, la apliquen fácilmente.

El contenido de esta guía responde a las necesidades encontradas en el Diagnóstico de la Metodología Estadística Utilizada en Tesis Publicadas de Enero 2003 a abril 2005 por las FQ – URL (De León, 2005). Así que se incluye estadística descriptiva, técnicas de muestreo, estadística inferencial, pruebas de hipótesis para una o dos muestras de la media, proporción y varianza y finalmente el análisis de varianza (ANDEVA o ANOVA por sus siglas en inglés).

En ningún momento se pretende presentar los métodos de una manera rígida y mucho menos obligar la aplicación de la estadística en investigaciones universitarias. Lo que sí se pretende es que, si se decide la aplicación de la estadística, se haga de manera correcta, clara, sencilla y con argumentos, para lo que se recomienda que exista interés en profundizar en los temas que se investigan, que se utilice la estadística como herramienta para que los resultados de la investigación sean confiables y así tomar decisiones y conclusiones mas adecuadas a la realidad.

II. SELECCIÓN DEL MÉTODO ESTADÍSTICO

La selección del método estadístico depende directamente del diseño de la investigación. Si se trata de una investigación de tipo descriptivo algunas veces es necesario aplicar **estadística descriptiva** pura, sin necesidad de hacer inferencia de una muestra a una población. Pero si el caso es que sí se está utilizando una muestra para generalizar los resultados en una población, a través de la estimación puntual o por intervalos, se debe utilizar la **estadística inferencial**. La estadística inferencial también es útil cuando se busca la comprobación de hipótesis de tipo experimental en relación a una, dos o mas muestras. Se opta por utilizar la **construcción de modelos** cuando se investiga la relación que pueda existir entre dos o más variables de interés. Es importante recalcar que la utilización de un método determinado no limita la utilización de otro, de hecho es posible utilizar mas de un método estadístico para el análisis de las investigaciones, siempre y cuando el diseño de la investigación lo amerite.

Para ayudar a la selección del método es necesario conocer que Métodos Estadísticos existen y sus características, las cuales se describen en la siguiente sección.

III. MÉTODOS ESTADÍSTICOS

Los métodos estadísticos son herramientas que permiten el análisis de datos, que son recopilados a través de observaciones, experimentos e investigaciones, y así tomar decisiones con bases confiables.

Existen tres categorías de los métodos estadísticos que son: estadística descriptiva, estadística inferencial y construcción de modelos.

3.1. Estadística descriptiva.

La **estadística descriptiva** abarca técnicas analíticas y gráficas para organizar, resumir y presentar información. Esto con el fin de describir o representar visualmente un conjunto de datos que corresponden a una situación de interés. Entre las técnicas más utilizadas están las tablas de frecuencia simple, tablas de frecuencia por intervalos, histogramas de frecuencia, graficas de barras, graficas tipo pastel y otras, que ahora con la ayuda de paquetes de cómputo son de muy fácil aplicación y elaboración automática. Es importante no dejar fuera el cálculo de medidas de tendencia central, de posición y de dispersión, que vienen a ampliar la descripción de la situación en estudio. Esta parte de la estadística sólo se ocupa de describir y analizar un grupo dado, sin sacar conclusiones sobre otro grupo. El grupo que se analiza en la estadística descriptiva generalmente es la población de interés.

Ejemplo 1: Un estudio esta dirigido a conocer el nivel académico de los empleados de una empresa en particular. Esta información es accesible, directamente de los mismos empleados o por medio del archivo de personal. A través de la estadística descriptiva es posible organizar la información de manera que se pueda describir la situación. Esto ayudará a tomar decisiones, posiblemente acerca de promociones internas o apoyo a planes de estudios de los empleados de dicha empresa. No se pretende hacer ninguna inferencia sobre otras empresas aun pertenecieran a una misma corporación, organizaciones o poblaciones independientes a esta.

Entre las técnicas de la estadística descriptiva esta:

3.1.1. Organización y presentación de datos.

La organización y presentación de datos puede darse a través de tablas estadísticas (o cuadros estadísticos) de frecuencia simple, de frecuencia por intervalos y por otras representaciones gráficas.

Generalmente cuando se obtiene la información es de una manera desordenada. En el ejemplo 1, sobre el estudio del nivel académico de los empleados de una empresa, es muy probable enfrentarse a la información como la que se presenta en la Tabla 1.

5	5	15	3	6	8	9	11	12	10	7	7	6	10	15	14	9	13	14	10
7	5	6	6	6	8	12	9	5	12	5	7	13	19	14	8	14	13	16	17
7	7	6	3	11	8	9	4	7	7	9	5	15	10	8	8	9	17	15	13
12	7	6	11	7	4	8	7	7	12	4	9	16	8	13	13	10	14	16	14
12	12	6	4	6	6	11	8	11	12	17	11	13	10	14	14	13	15	13	19

Tabla 1: Resultados en años de estudio de los n = 100 empleados que conforman la empresa. Datos ficticios de elaboración propia.

En esta forma es imposible visualizar las tendencias que puedan existir. Sin embargo a través de una hoja de registro diseñada para recopilar tal información se pueden tener los resultados de una manera mas ordenada como la presentada en la Tabla 2.

Años de estudio		Registro de datos	f
Primaria	0		
	1		
	2		
	3	XX	2
	4	XXXX	4
	5	XXXXXX	6
	6	XXXXXXXXXX	10
Básicos	7	XXXXXXXXXXXX	12
	8	XXXXXXXXX	9
	9	XXXXXXX	7
Diversificado	10	XXXXXX	6
	11	XXXXXX	6
	12	XXXXXXXX	8
Universidad (Licenciatura)	13	XXXXXXXXX	9
	14	XXXXXXXX	8
	15	XXXXX	5
	16	XXX	3
	17	XXX	3
Maestría	18	XX	2
	19		
		n =	100

Tabla 2: Resultados en años de estudio de los 100 empleados que conforman la empresa.

La hoja de registro utilizada en esta oportunidad es conocida también como *Histograma de Frecuencia*, el cual es utilizado también como un medio gráfico de presentación de datos. De esta hoja de registro se extrae fácilmente la tabla de frecuencia simple presentada en la Tabla 3.

Años de estudio		f	F	fr	Fr
Primaria	3	2	2	2%	2%
	4	4	6	4%	6%
	5	6	12	6%	12%
	6	10	22	10%	22%
Básicos	7	12	34	12%	34%
	8	9	43	9%	43%
	9	7	50	7%	50%
Diversificado	10	6	56	6%	56%
	11	6	62	6%	62%
	12	8	70	8%	70%
Universidad (Licenciatura)	13	9	79	9%	79%
	14	8	87	8%	87%
	15	5	92	5%	92%
	16	3	95	3%	95%
	17	3	98	3%	98%
Maestría	18	2	100	2%	100%
	n =	100		100%	

Tabla 3: Tabla de frecuencia simple de años de estudio de los 100 empleados que conforman la empresa. Donde años de estudio = posibles categorías en numero de años de educación formal aprobados para los empleados, f = frecuencia o numero de resultados para cada categoría, F = frecuencia acumulada, fr = frecuencia relativa o porcentaje que corresponde a cada categoría y Fr = frecuencia relativa acumulada.

Las **tablas estadísticas de frecuencia simple** son un ordenamiento de los datos recopilados en una investigación que sirven para facilitar el cálculo de valores estadísticos. Comúnmente se utilizan los criterios de mayor a menor o viceversa. Para cada categoría existe un número de casos obtenidos f, una frecuencia acumulada F, una frecuencia relativa fr y una frecuencia acumulada relativa Fr. Esta tabla pudo obtenerse partiendo de la información presentada en la Tabla 1 a través del ordenamiento de la información o directamente de la Tabla 2. En ambos casos se debió calcular los valores para F, fr y Fr como se detalla a continuación.

n = numero total de datos recopilados.

f = frecuencia o número de resultados para cada categoría
Se obtiene directamente al recopilar la información.

F = frecuencia acumulada
Se calcula sumando los valores de f para todas las categorías inferiores y la categoría en cuestión.

fr = frecuencia relativa
Se calcula dividiendo el valor f entre el número total de datos recopilados n, para cada categoría, multiplicado por 100. Se expresa en términos de porcentajes (%).

Fr = frecuencia relativa acumulada
Se calcula sumando los valores de fr para todas las categorías inferiores y la categoría en cuestión, por lo tanto también se expresa en porcentajes (%).

Algunas veces no es conveniente ir tan de prisa y elaborar una tabla estadística de frecuencia simple, cuando ya se tiene un histograma de frecuencia que brinda información importante como la proporcionada en la Tabla 2. En este caso la información ya esta reflejando algunas tendencias, como lo son que hay mayorías en los intervalos de 6 a 8 años de estudio y de 12 a 14 años de estudio. Este tipo de histograma recibe el nombre de *Histograma de Doble Pico*. Generalmente se da este fenómeno cuando dentro de la población existen dos grupos diferentes que fueron incluidos sin hacer diferenciación entre ellos (hombres y mujeres, extranjero y nacional, derecha e izquierda, invierno y verano, máquina 1 y máquina 2, etc.) En este ejemplo el fenómeno de doble pico se debe a que fueron incluidos sin hacer distinción el departamento operativo y el administrativo. Si se hace una división de estos departamentos, desde el momento que se recopila la información, se obtienen dos histogramas de frecuencia de acuerdo a las Tablas 4 y 6.

Años de estudio		Registro de datos	f
Primaria	3	XX	2
	4	XXXX	4
	5	XXXXXX	6
	6	XXXXXXXXXX	10
Básicos	7	XXXXXXXXXXXX	12
	8	XXXXXXXXX	9
	9	XXXXXXX	7
Diversificado	10	XXXXX	5
	11	XXX	3
	12	XX	2
		TOTAL	60

Tabla 4: Resultados en años de estudio de operarios de la empresa.

En este histograma de frecuencia se encuentran registrados los años de estudio de los operarios de la empresa, que refleja un comportamiento *simétrico* en forma de *campana*, o como también es muy conocido en forma aproximada de *distribución normal*.

Para este caso la tabla de frecuencia simple quedaría como el que se muestra en la Tabla 5.

Años de estudio		f	F	fr	Fr
Primaria	3	2	2	3%	3%
	4	4	6	7%	10%
	5	6	12	10%	20%
	6	10	22	17%	37%
Básicos	7	12	34	20%	57%
	8	9	43	15%	72%
	9	7	50	12%	84%
Diversificado	10	5	55	8%	92%
	11	3	58	5%	97%
	12	2	60	3%	100%
	TOTAL	60		100%	

Tabla 5: Tabla de frecuencia simple para años de estudio de operarios de la empresa. Donde años de estudio = posibles categorías en numero de años de educación formal aprobados para los empleados, f = frecuencia o numero de resultados para cada categoría, F = frecuencia acumulada, fr = frecuencia relativa o porcentaje que corresponde a cada categoría y Fr = frecuencia relativa acumulada.

Años de estudio		Registro de datos	f
Diversificado	10	X	1
	11	XXX	3
	12	XXXXXX	6
Universidad (Licenciatura)	13	XXXXXXXXX	9
	14	XXXXXXXX	8
	15	XXXXX	5
	16	XXX	3
	17	XXX	3
Maestría	18	XX	2
		TOTAL	40

Tabla 6: Resultados en años de estudio de administrativos de la empresa.

El histograma de frecuencia mostrado en la tabla 6 registra los años de estudio de los administrativos de la empresa. Este también refleja un comportamiento aproximadamente normal. Generalmente es muy aceptable este tipo de resultado ya que por definición se puede decir que todos los fenómenos tienden a tener este tipo de distribución, esto por supuesto bajo condiciones normales, es decir que no exista una causa que obligue otro resultado.

La Tabla 7 muestra la tabla de frecuencia simple para los años de estudio de los administrativos de la empresa.

Años de estudio		f	F	fr	Fr
Diversificado	10	1	1	2%	2%
	11	3	4	8%	10%
	12	6	10	15%	25%
Universidad (Licenciatura)	13	9	19	22%	47%
	14	8	27	20%	67%
	15	5	32	12%	79%
	16	3	35	8%	87%
	17	3	38	8%	95%
Maestría	18	2	40	5%	100%
	TOTAL	40		100%	

Tabla 7: Tabla de frecuencia simple de años de estudio de administrativos de la empresa. Donde años de estudio = posibles categorías en numero de años de educación formal aprobados para los empleados, f = frecuencia o numero de resultados para cada categoría, F = frecuencia acumulada, fr = frecuencia relativa o porcentaje que corresponde a cada categoría y Fr = frecuencia relativa acumulada.

Luego de dividir la información considerando la diferencia de departamentos es posible visualizar la información de manera más específica, lo que permite considerar las diferencias que pueden darse entre los dos grupos de la población. Es posible entonces tomar decisiones sobre promociones o programas de estudios de acuerdo a las necesidades que se dan en cada uno de los grupos que conforman la empresa.

Las **tablas estadísticas de frecuencia por intervalos** son otra alternativa para presentar los datos ahora refiriéndose a que los datos son agrupados de acuerdo a un intervalo que contiene no una sino varias categorías. Se recomienda utilizar estas tablas para resumir y visualizar de mejor manera los resultados obtenidos. La utilización de las tablas de frecuencia por intervalos tiene relación con el número de observaciones. Algunos autores recomiendan elaborar estas tablas cuando se tienen de 16 a más datos.

Es importante tener presente que el objetivo de las tablas de frecuencia por intervalos es facilitar al investigador y a quienes tengan acceso a la información el análisis de los datos, por lo que posiblemente en el proceso de diseñar una tabla de este tipo se tengan que tomar decisiones claves para que al final la presentación de los datos cumpla con sus objetivos.

El investigador no debe preocuparse demasiado por la construcción de una tabla de frecuencia por intervalos, incluso se recomienda que si el fenómeno en estudio ya cuenta por naturaleza con intervalos que comúnmente son aplicados, que son de fácil lectura y que cumple con el requisito de que sean del mismo tamaño, se deben adoptar. Por ejemplo cuando se tiene información que comúnmente su comprensión es a través de decenas como los años de vida de una persona, las décadas de los años o las calificaciones académicas en la universidad. Para muchas personas por ejemplo existe una gran diferencia entre obtener una calificación en el curso de Estadística de 79 puntos o 80 puntos, esta diferencia no es tan apreciada si se trata de obtener 75 o 76 puntos. Y lo mismo sucede con las edades y con el estudio de la

historia por décadas, los 60´s para muchos tienen características específicas que son diferentes a las de los 70´s.

Esto se refiere a que en el proceso de definición de los intervalos aunque existan procedimientos específicos que se realizan a través de la aplicación de formulas, pero que si los resultados no favorecen la visualización de los datos, es importante que el investigador opte por aplicar otros criterios.

Para los casos en los que no existe ya una clasificación predeterminada como en el caso de las décadas (podrían ser también centenas o siglos) se presenta un procedimiento para diseñar una tabla estadística de frecuencia por intervalos, que consta de los siguientes 10 pasos:

1. Contar el número de observaciones n.

2. Definir el número de intervalos o clases I.

El número de intervalos, I, a utilizar se recomienda que este entre 5 y 20 intervalos. Para calcular este número se puede hacer a través de las formulas siguiente:

Si el valor de n no es muy grande: $I = \sqrt{n}$

En otro caso: $I = 1 + 3.22 \log n$

Por ejemplo si el número de observaciones que se tienen es n=100, un buen criterio es agrupar las observaciones en $I = \sqrt{100} = 10$ intervalos. Pero si el caso fuera que n=1,000,000 calculando a través de la raíz cuadrada de n se obtendrían 1,000 intervalos, para lo cual sería mas razonable si se utiliza I=1+3.22log 1,000,000=20.32 que se aproxima a 20 intervalos.

Otra herramienta para auxiliarse en este cálculo se presenta en la Tabla 8.

Número de datos	Números de intervalos
Menos de 16	Datos insuficientes
16 – 31	5
32 – 63	6
64 – 127	7
128 – 255	8
256 – 511	9
512 – 1023	10
1024 – 2047	11
2048 – 4095	12
4096 – 8190	13

Tabla 8: números recomendados de intervalos para uso en las subdivisiones de datos numéricos en función del numero de datos. Fuente Milton – Arnold 2004. Probabilidad y estadística con aplicaciones para ingeniería y ciencias computacionales, 4ª edición. McGraw – Hill. Página 197.

Es importante que el investigador decida si utilizará o no estos criterios para definir el número de intervalos basándose en el propio conjunto de datos y en sus objetivos personales. Los métodos descritos anteriormente ayudan como referencia, pero es el investigador quien debe decidir el número de categorías que desea emplear para agrupar sus datos en una tabla de frecuencia por intervalos.

3. Localizar las observaciones con valor máximo y mínimo.

4. Calcular el rango de los datos R = valor máximo – valor mínimo.

5. Definir el ancho del intervalo C.

$$C = \frac{\text{valor máximo – valor mínimo + “unidad”}}{I}$$

Cuando se está trabajando con una variable que se cuenta de uno en uno la “unidad” es fácilmente percibida como 1. Sin embargo cuando la variable está expresada con números decimales la “unidad” toma el valor de la mínima subdivisión que se puede dar a la variable.

Por ejemplo si el fenómeno en estudio es la distribución de los salarios de una compañía, es fácil concebir que estos se presentarán incluyendo decimales para presentar la variable hasta unidades de centavos, en este caso la “unidad” toma el valor de un centavo o bien “unidad” = 0.01.

6. Construir los intervalos a través de los límites aparentes y reales.

Los límites aparentes se construyen partiendo desde el valor mínimo hasta llegar a cubrir el total de intervalos definidos en el paso 2, los cuales cubrirán la totalidad de observaciones.

Para los límites reales se considera el espacio que existe entre un límite aparente superior y el límite aparente inferior del siguiente intervalo. El límite real se ubica justamente a la mitad de este espacio.

7. Calcular el punto medio X_i de cada intervalo.

El punto medio es el puntaje que representa al intervalo de clase para efectos de cálculos matemáticos y se ubica en el centro del mismo. Para su cálculo se hace a través de promediar el límite aparente superior con el límite aparente inferior, de la manera siguiente:

$$X_i = \frac{\text{limite aparente superior + limite aparente inferior}}{2}$$

8. Contar los resultados f para cada intervalo en una columna adicional.

9. Completar la tabla con las columnas para: F, fr y Fr.

10. Para calcular la media $\overline{X}$, varianza S^2 y desviación estándar s se debe ampliar la tabla con las columnas:

    11. X_i x fr (la sumatoria da como resultado la media $\overline{X}$)

    12. $(X_i - \overline{X})$

    13. $(X_i - \overline{X})^2$

    14. fr x $(X_i - \overline{X})^2$ (la sumatoria da como la varianza s^2)

Ejemplo 2:

Partiendo de los resultados de la Tabla 1, se elaborará una tabla estadística de frecuencia por intervalos de la manera siguiente:

Paso 1: Contar el número de observaciones n.
La cantidad de datos obtenidos es n = 100, entonces el número de intervalos

Paso 2: Definir el número de intervalos o clases I.
A través de las formulas:

$$I = 1 + 3.22 \log 100 = 7.44 \approx 8$$

Paso 3: Localizar las observaciones con valor máximo y mínimo.

Valor máximo = 18
Valor mínimo = 3

Paso 4: Calcular el rango de los datos R = valor máximo – valor mínimo.

$$R = 18 - 3 = 15$$

Paso 5: Definir el ancho del intervalo C.

$$C = \frac{18 - 3 + 1}{8} = 2$$

Pasos 6 – 8: Construir los intervalos a través de los límites aparentes y reales. Calcular el punto medio Xi de cada intervalo. Contar los resultados f para cada intervalo en una columna adicional.

No.	Límites aparentes		Límites reales		Xi	F
	Li	Ls	Lri	Lrs		
1	3	4	2.5	4.5	3.5	6
2	5	6	4.5	6.5	5.5	16
3	7	8	6.5	8.5	7.5	21
4	9	10	8.5	10.5	9.5	13
5	11	12	10.5	12.5	11.5	14
6	13	14	12.5	14.5	13.5	17
7	15	16	14.5	16.5	15.5	8
8	17	18	16.5	18.5	17.5	5
Sumatorias						100

Tabla 9: Tabla de frecuencia por intervalos de años de estudio de los 100 empleados que conforman la empresa.

Pasos 9 y 10: Completar la tabla con las columnas para: F, fr, Fr, Xi x fr, $(Xi - \overline{X})$, $(Xi - \overline{X})^2$ y fr x $(Xi - \overline{X})^2$ Esta Tabla (9a) se presenta en el Anexo 1.

Existen algunos fenómenos que alcanzan a tener únicamente dos categorías (falso y verdadero, si y no, blanco y negro, de acuerdo y en desacuerdo, masculino y femenino, etc.), siendo estos fenómenos muy frecuentes en trabajos de investigación universitaria. Estos fenómenos en donde se pueden obtener dos tipos de resultados también pueden expresarse a través de tablas estadísticas, que son además muy sencillas pero que pueden expresar información muy importante.

Por ejemplo, si la empresa mencionada en el Ejemplo 1 tiene como una de sus intenciones nivelar los años de estudio de los empleados de nivel operativo a que tengan los básicos aprobados, se puede tomar la información de la Tabla 4, para presentar la información de una manera más conveniente para el análisis, como se hace en la Tabla 10.

Trabajador que tiene aprobados los básicos	f	fr
Si	17	28%
No	43	72%
Total	60	100%

Tabla 10: Trabajadores del nivel operativo que tienen y no aprobados los básicos.

En estos casos puede ser muy interesante hacer cálculos de razones y tasas. Las razones es un método común para comparar las dos categorías existentes y se hace a través de una operación muy sencilla dividiendo la frecuencia de una categoría entre la frecuencia de la otra, por ejemplo:

$$\text{Razón} = 43/17 = 2.5$$

De este valor resultante se puede concluir que la cantidad de trabajadores del nivel operativo que no cuentan con los estudios básicos completos es 2.5 veces la cantidad de los que si cuentan con estos estudios terminados.

En cuanto a las tasas es otra forma de expresar a las razones, aunque estas generalmente se utilizan para indicar comparaciones entre el número de casos reales y el número de casos potenciales. Por ejemplo para determinar la tasa de nacimiento para determinada población, la tasa de divorcios, en un tiempo determinado. Refiriéndonos al ejemplo que venimos trabajando, si se sabe que en año anterior fueron 5 los trabajadores del nivel operativo que aprobaron el tercer año de básico,

de 10 que estaban cursándolo, esto significa que 5 de cada 10 operarios aprobaron el tercer año básico el pasado año, claramente para el caso de la empresa en análisis. Y si se quisiera tener una idea de cual sería la cantidad por cada 100 operarios estudiantes del tercer año básico, entonces:

Tasa de aprobación del tercer año básico = 100x5 /10 = 50 estudiantes por cada 100

Aclarando nuevamente que esto es para la empresa en análisis.

Este tipo de cálculos también se puede dar para cuando se agrupan los datos en categorías como: bueno, regular y malo. Aquí hay tres categorías en donde la diferencia con el ejemplo anterior es que se permite una clasificación intermedia. Así como se utilizan tres categorías se pueden utilizar cinco como: pésimo, insuficiente, aceptable, bueno y excelente. Con esto se pretende expresar que las formas de agrupar los datos pueden ser tanto por intervalos numéricos como por categorías cualitativas.

Las **representaciones gráficas** son técnicas visuales para la representación de datos. Los más comunes son: histogramas, graficas de pastel, diagramas de dispersión, polígonos de frecuencia, etc. Estos son de muy fácil aplicación por medio de programas como Microsoft Excel, en donde únicamente al accionar el Asistente para Gráficos en la barra de herramientas y seguir las instrucciones de alimentación de la información se logra elaborar un gráfico de manera automática. Utilizando esta herramienta se han elaborado diferentes graficas para representar el comportamiento del ejemplo que hemos venido citando sobre el nivel académico de los trabajadores de una empresa. Estos gráficos se presentan en el Anexo 2.

3.1.2. Calculo de medidas de tendencia central y de dispersión.

Un fenómeno no está completamente descrito a través de las tablas estadísticas o los otros medios gráficos, para terminar la descripción es necesario contar con las

medidas de tendencia central como la moda, media y mediana, y las medidas de dispersión como la varianza y la desviación estándar.

A continuación se hace una descripción de estas medidas:

Moda:
Es la categoría que cuenta con la mayor frecuencia u ocurrencia dentro de la distribución. Hay algunas distribuciones que pueden tener mas de una moda, es decir que la distribución tiene más de una categoría con presencia muy marcadas.

Media:
La manera más fácil de expresar es como un valor promedio de los resultados. En una lista de puntajes solo basta con elaborar la sumatoria de estos puntajes y dividirlos entre la cantidad de puntajes sumados. También es posible hacer el calculo para distribuciones de frecuencia por intervalos o agrupadas.

Mediana:
Se considera que las observaciones han sido ordenadas de menor a mayor, entonces la mediana tiene el valor que divide exactamente a la mitad la distribución. Es decir el valor en donde el 50 % de los datos se encuentra por debajo de este y el 50 % se encuentra por arriba de este. En el caso de que n sea par los dos valores centrales se promedian para obtener la mediana, mientras que si n es impar la mediana toma el valor del valor central.

Varianza:
Es un grado de dispersión de los datos en unidades cuadradas.

Desviación estándar:

Es un grado de dispersión de los datos en las unidades reales, viene a resolver el problema de manejar la varianza con unidades cuadradas.

Todas estas mediciones se pueden calcular tanto para una muestra como para una población, sin embargo para fines técnicos es importante tener claro que aquellos que se calculan a partir de la muestra reciben el nombre de **estadísticos**, mientras que aquellos que corresponden a la población son los **parámetros**.

En el Anexo 3 se presentan las formulas necesarias para el calculo de la medidas anteriores.

3.2. Estadística Inferencial

La **estadística inferencial** se define como un conjunto de técnicas que ayudan a los investigadores a hacer inferencias de la muestra hacia la población, lo que significa que a partir de una **muestra** se harán los cálculos de **estadísticos** o estadísticas (mediciones que corresponden a la muestra como: proporción muestral, media muestral, varianza muestral y desviación estándar muestral) y estos se infieren en la **población**, que consiste en considerarlos representativos de los **parámetros** (mediciones que corresponden a la población como: proporción poblacional, media poblacional, varianza poblacional y desviación estándar poblacional) a un cierto nivel de confianza. Lo anterior recibe el nombre de **Estimación** (que puede ser puntual o por intervalos). Aunque la estimación ya implica la toma de decisiones; en este caso el aceptar o no un estadístico como representativo de un parámetro, existe una parte aun mas desafiante de la estadística inferencial que es lo concerniente a comprobar hipótesis, que en este caso las decisiones radican en aceptar o rechazar una hipótesis. Partiendo de estas decisiones es necesario tomar otras no menos ni mas importantes, que son las relacionadas directamente con la situación en análisis, que se refieren a la generación de soluciones alternativas, el planteamiento de estrategias y en si la acción misma que se realizará, como respuesta a los resultados

obtenidos en las aplicaciones de estadística. Sin estas últimas decisiones resulta inútil el uso de la estadística, convirtiéndose únicamente en práctica metodológica.

3.2.1. Estimación puntual y por intervalos

La sola estimación consiste en inferir el valor de un estadístico a la población. Esto es posible hacerlo de dos maneras que son: estimación puntual y estimación por intervalos.

La estimación puntual es más exacta, en el sentido que se infiere un valor exacto, sin embargo es menos confiable ya que es menos probable que el valor sea exactamente el que se infirió.

Para hacer inferencias mas confiables es que se utiliza la estimación por intervalos, es decir que el parámetro se estima no por un valor exacto sino por un intervalo, donde el centro de este sería el valor del estadístico. Esto permite que la inferencia no sea precisamente exacta pero si mas confiable, ya que es mas probable que el valor real del parámetro esté en este intervalo.

La estimación puntual únicamente requiere de definir que el valor del estadístico se infiere al parámetro. La estimación por intervalos requiere de definir el tamaño del intervalo para lo cual se requiere de conocer la desviación estándar de la distribución de la variable de interés, de un modelo que corresponda a esta variable y de un nivel de confianza deseado.

En el Anexo 4 se presenta un formulario para utilizarse en la estimación puntual y por intervalos.

3.2.2. Pruebas de hipótesis

Muchos de los problemas de toma de decisiones, pruebas o experimentos en todas las profesiones u ocupaciones, pueden formularse como problemas de prueba de hipótesis. Independientemente de la distribución de probabilidad que se esté tratando el procedimiento para establecer una prueba de hipótesis implica definir claramente lo siguiente:

1. Parámetro o característica de la población, o poblaciones que están en análisis, como proporción ($\mathbf{p}$), media (μ), varianza (σ^2), o una relación entre proporciones ($p_1 - p_2$), medias ($\mu_1 - \mu_2$) o varianzas ($\sigma^2_1 = \sigma^2_2$), al cual se le identifica de manera genérica como θ o si es el caso de dos poblaciones como $\theta_1 - \theta_2$.
2. Valor del parámetro o relación que existe entre ellos, relativo a su realidad, es decir el valor que se supone que ésta característica tiene en la población, el cual se identifica como valor nulo θ_0 (μ_0, σ^2_0, p_0) o para dos poblaciones podría ser $\theta_1 - \theta_2 = \Delta_0$ (donde generalmente Δ_0 obtiene el valor de 0 para expresar que $\theta_1 = \theta_2$, u cualquier otro valor que expresa una diferencia de interés)
3. Hipótesis Nula (H_0) que es una afirmación inicial que favorece que el parámetro tiene realmente el valor de $\theta = \theta_0$ o $\theta_1 - \theta_2 = \Delta_0$. Siempre se expresa como una igualdad.
4. Hipótesis Alternativa (H_a) que es una afirmación que se considerará aceptada si y solo si la evidencia muestral le proporciona un fuerte apoyo. Se puede expresar como una desigualdad a ambos lados con el signo no igual que ($\neq$), una desigualdad a la izquierda con el signo menor que ($<$) o como una desigualdad a la derecha con el signo mayor que ($>$)
5. El estimador al cual se le identifica de manera genérica con el símbolo $\hat{\theta}$, theta sombrero ($\hat{p}$, $\overline{X}$, S^2) que se calculará a partir de los datos muestrales.

6. La distribución de probabilidad que aplica al análisis en cuestión (normal, t, ji-cuadrada, F, etc.), que posiciona al estadístico de prueba (z, t, χ^2, F, etc.) en un lugar de la distribución.
7. Una región de rechazo que contiene los valores a los que H_o será rechazada, y por ende H_a será aceptada.
8. El nivel de confianza al que se hará el estudio (α).

Es recomendable que esta información sea definida antes de realizar la propia investigación, ya que es la guía para determinar que información es útil para el estudio. Esta recomendación no se hace únicamente con el afán de apoyar el proceso de investigación, sino que también para que no exista la posibilidad de sesgar la información a favor de intereses paralelos a la investigación.

Luego de realizada la investigación se deben depurar, organizar y operar los datos, de manera que sea posible hacer el calculo de el estimador $\hat{\theta}$**,** que nos proporcione el valor muestral de la característica en estudio, por ejemplo:

Proporción muestral = p(A) = N(A)/N

Donde:

N(A) es el número de resultados que tienen la característica A

N es el número total de resultados obtenidos

Media muestral = $\overline{\mathbf{X}} = \Sigma (x_i) / n$

Donde:

i = 1, 2, 3, … n

n es el tamaño de la muestra o número de datos disponibles.

Varianza muestral = $\mathbf{S^2} = \Sigma (x_i - \overline{X})^2 / (n - 1)$

Donde:

i = 1, 2, 3, … n

n es el tamaño de la muestra o número de datos disponibles.

$\overline{X}$ es la media muestral

Además es necesario hacer el calculo de la desviación estándar muestral ($S = \sqrt{S^2}$) para luego calcular el estadístico de prueba, y finalmente comprobar la H_o.

En una prueba de hipótesis se puede caer en dos tipos de error que comúnmente son conocidos como: Error Tipo I (que consiste en rechaza H_0 cuando es verdadera) y Error Tipo II (que implica aceptar H_0 cuando es falsa).

En el Anexo 5 se presenta una clasificación de modelos estadísticos que se utilizan mas frecuentemente para la prueba de hipótesis, considerando el o los parámetros de interés y el tamaño de la o de las muestras disponibles, con el objetivo de brindarle una herramienta para que utilice en los estudios en donde se requiera la aplicación de dichos modelos. Esta es una recopilación de modelos que los autores citados en la bibliografía nos brindan es sus obras, puede ser ampliada a medida que las destrezas en la estadística se lo permitan.

3.2.3. Análisis de varianza

El **Análisis de Varianza** (ANDEVA O ANOVA por sus siglas en inglés) es una prueba para la comparación entre grupos a través de la varianza de la variable numérica “y”, en cada grupo de la variable categórica “x”. Básicamente el ANDEVA, se utiliza para corroborar si la significación de diferencias entre medias de dos o mas grupos, son o no debidas al azar. La cifra estadística de prueba obtenida con esta prueba es la razón F.

Suponiendo que se analizan 2 grupos, el ANDEVA, analiza las variaciones entre los dos grupos (inter-grupal) y la compara con la variación dentro de cada grupo (intra-grupal), para obtener mediante una suma de cuadrados el valor de F.

Si las diferencias de varianza entre cada grupo son mayores que las intra-grupales, seguramente existen diferencias significativas entre los grupos que no son debidas al azar.

3.3. Construcción de modelos.

La **construcción de modelos** se refiere al desarrollo de ecuaciones predictivas a partir de datos experimentales. Esto se logra con la aplicación de herramientas matemáticas que permiten dar una explicación teórica del fenómeno en estudio, dando la oportunidad de poder explicarlo en forma verbal.

Lamentablemente este es un método poco utilizado en las investigaciones universitarias, aun cuando existen varios estudios experimentales que pueden requerir de este método. Por ejemplo se puede construir un modelo lineal que relacione la acidez de un líquido con la cantidad dosificada a este de cierto producto.

La aplicación de la estadística requiere además de la comprensión de las herramientas matemáticas llamadas **teoría de la probabilidad**, que en algunas disciplinas como la ingeniería se requiere de un dominio alto de estas. Sin embargo hay otras disciplinas que únicamente necesitan dominar la mecánica de aplicación de estas teorías para darle utilidad a las distintas distribuciones de probabilidad que existen. Algunas veces puede parecer sencilla la aplicación mecánica de la teoría de probabilidad, sin embargo es deseable que cada vez más y más se cuente con el conocimiento de la teoría de la probabilidad para la realización de estudios técnicos.

Es elemental conocer que la probabilidad se puede expresar tanto en porcentajes (%) como en proporciones o fracciones (p), y que lo probable no obliga a su ocurrencia, sino que únicamente es una expresión de las posibilidades que se tienen; unas veces más que otras. Un nivel mas avanzado de conocimiento se refiere a saber interpretar quien es el todo, sus partes y la infinidad de arreglos que se pueden dar, dando lugar a las técnicas de conteo.

IV. SELECCIÓN DE LA MUESTRA

Como ya se mencionó la estadística inferencial parte de los resultados de una muestra para inferirlos en la población a un cierto nivel de confianza, por lo cual es muy importante contar con muestras representativas de la población, para darnos la oportunidad de inferir importantes conclusiones sobre la población a partir del análisis de la muestra.

4.1. Conceptos básicos

Universo:
Conjunto infinito de elementos o unidades (individuos, organizaciones, objetos, etc.) cualesquiera, en los cuales se consideran una o más características o condiciones de admisión en el conjunto, que se someten a estudio estadístico.

Población:
Conjunto de elementos o unidades (individuos, organizaciones, objetos, etc.) de interés, que posee la o las características que resultan básicas para el análisis del problema que se estudia.

En algunos estudios el universo se considera que es el mismo que la población, pero cuando no es de esta manera se define que el universo contiene a la población, por ejemplo: se desea realizar un estudio sobre las empresas de Quetzaltenango, éste es un universo que contiene las poblaciones de: empresas industriales, comerciales y de servicios de Quetzaltenango. Así mismo podría contener otras poblaciones como pequeña, mediana y grande empresa, esto va a depender de la característica que conviene al estudio, en estos casos puede ser tipo de actividad o tamaño respectivamente.

La población puede ser real, es decir que existe en la naturaleza, y también puede ser hipotética (no existe hasta que se construye) que esta formada por todas las

mediciones hechas bajo condiciones experimentales. Por ejemplo la población de resultados obtenidos al lanzar una moneda 1 vez: los posibles resultados son cara o escudo, es decir que la población puede estar conformada por una cara o por un escudo.

Muestra:
Es la lista de individuos u objetos a ser muestreados, o que están disponibles para un investigador, o se le puede construir.

Nivel de confianza:
Se refiere a la probabilidad de que el estadístico obtenido a través de la muestra corresponda al parámetro de la población.

Error experimental:
Es el error máximo permitido en la investigación.

4.2. Definir el tamaño de la muestra

El no determinar el tamaño de la muestra de una manera técnica, puede dar como resultado dos situaciones diferentes: una es que se realice el estudio sin el número adecuado de elementos, con lo cual no se podrá ser precisos al estimar los parámetros y además no se encontrarán diferencias significativas cuando en la realidad sí existen. La otra situación es que se podría estudiar un número innecesario de elementos, lo cual lleva implícito no solo la pérdida de tiempo e incremento de recursos innecesarios sino que además la calidad del estudio, dado dicho incremento, puede verse afectada en sentido negativo.

Aquí se recomienda la aplicación de una formula, para definir el tamaño de la muestra, la cual se explica a continuación:

$$n_0 = \frac{z^2 * p * q}{e^2}$$

Donde:

n_0 = tamaño inicial de la muestra (o definitivo)

z^2 = estimador insesgado para el intervalo de confianza, elevado al cuadrado

p = probabilidad de éxito

q = probabilidad de fracaso (1-p)

e^2 = error muestral al cuadrado

Nivel de confianza	Z
90 %	- 1.64
95 %	- 1.96
95.45 %	- 2.00
99 %	- 2.81

Cuando se conoce el tamaño de la población o universo es posible hacer un ajuste o corrección al tamaño inicial de la muestra con la formula siguiente:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Donde:

N = tamaño de la población o universo

n = tamaño corregido de la muestra de acuerdo al tamaño de la población

4.3. Elegir el tipo de muestra

Los principales **Métodos de muestreo** son:

Muestreo aleatorio simple:
Es un muestreo al azar. Cada individuo de la población tiene la misma oportunidad de ser seleccionado. Si todos los elementos de la población pueden ser distinguidos uno del otro, el numero de muestras de tamaño n posibles en una población N se extraen a través de la formula de combinación ${}_{N}C_{n}$.

Muestreo estratificado o por departamentos:
Cuando se considera el tamaño proporcional de cada grupo por el que se compone la población.

Muestreo sistemático:
Consiste en obtener los elementos de la muestra en una forma ordenada.

Muestreo por racimos o conglomerados:
Útiles para estudios de gran magnitud (Ejemplo: a nivel nacional, se divide en regiones a través de un mapa y se muestrea independientemente),

Muestreo por conveniencia:
Se selecciona la muestra de acuerdo a los intereses que se persiguen.

Depuración de datos:
Se refiere a la anulación de datos que se consideran que no representan la realidad. Muchas veces se puede proceder a la anulación de encuestas basándose en las preguntas de verificación.

4.4. Aplicar el procedimiento de selección

El no contar con un método de selección de la muestra técnico, puede dar como resultado que la muestra no este distribuida uniformemente en la población proporcionando resultados que difieren de la realidad, ya que únicamente representaría a parte de la población y no a la totalidad como se pretende.

La selección se puede hacer a través de la utilización de Tablas de Números Aleatorios, sorteos u otro sistema definido previamente, que contemple el cumplimiento de la objetividad de la muestra.

Hay que tener presente también, que el muestreo en si no es lo importante, pero que sí de este depende el éxito que se tenga al hacer las inferencias deseadas.

V. REFERENCIAS BIBLIOGRÁFICAS

1. Devore, Jay L. "Probabilidad y estadística para ingeniería y ciencias" Editorial Thomson. Quinta Edición

2. Devore, Jay L. "Probabilidad y estadística para ingeniería y ciencias" Editorial Thomson. Cuarta Edición

3. Kume, Hitoshi. "Herramientas estadísticas básicas para el mejoramiento de la calidad" Editorial Norma.

4. Lemus, Jorge "Diseño del tamaño de la muestra en ciencias sociales" Universidad de San Carlos de Guatemala, Centro Universitario de Occidente. DIES.

5. Levin, Jack "Fundamentos de estadística en la investigación social" Editorial Harla. Segunda edición.

6. Mendenhall, W. "Estadística matemática con aplicaciones" Editorial Thomson. Sexta edición.

7. Milton, J. Susan / Arnold, Jesse C. "Probabilidad y estadística con aplicaciones para ingeniería y ciencias computacionales" Editorial Mc Graw - Hill. Cuarta Edición.

8. Walpole & Myers "Probabilidad y estadística para ingenieros" Editorial Prentice Hall. Sexta edición.

9. http://www.bioestadistica.uma.es/libro/node7.htm Texto versión electrónica del manual de la Universidad de Málaga: Bioéstadística: Métodos y Aplicaciones Facultad de Medicina. Universidad de Málaga. Material de apoyo.

10. http//www.fisterra.com/material/investiga/8muestras/8muestras/htm#top. "Determinación del tamaño muestral" Fernández, Pita.

Dibujo de portada tomado de http://www.secyt.gov.ar/estadistica/daily_usage_200502.png

VI. ANEXOS

Anexo 1: Tabla de frecuencia por intervalos

No.	Límites aparentes		Límites reales		Xi	f	F	fr (%)	Fr (%)	Xi x fr	$(Xi - \overline{X})$	$(Xi - \overline{X})^2$	fr x $(Xi - \overline{X})^2$
	Li	Ls	Lri	Lrs									
1	3	4	2.5	4.5	3.5	6	6	6	6	0.21	-6.44	41.47	2.48
2	5	6	4.5	6.5	5.5	16	22	16	22	0.88	-4.44	19.71	3.15
3	7	8	6.5	8.5	7.5	21	43	21	43	1.58	-2.44	5.95	1.24
4	9	10	8.5	10.5	9.5	13	56	13	56	1.24	-0.44	0.19	0.02
5	11	12	10.5	12.5	11.5	14	70	14	70	1.61	1.56	2.43	0.34
6	13	14	12.5	14.5	13.5	17	87	17	87	2.30	3.56	12.67	2.15
7	15	16	14.5	16.5	15.5	8	95	8	95	1.24	5.56	30.91	2.47
8	17	18	16.5	18.5	17.5	5	100	5	100	0.88	7.56	57.15	2.86
Sumatorias						100		100	$\overline{X}$ =	9.94		S^2 =	14.71

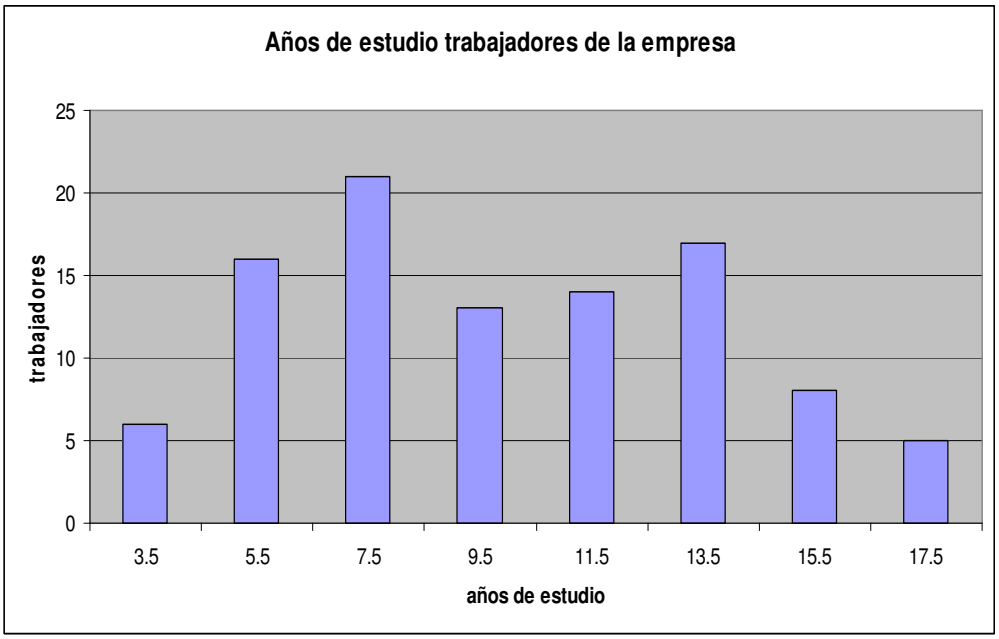
Tabla 9a: Complemento Tabla 9, pasos 9 y 10 del procedimiento para construir la tabla de frecuencia por intervalos, caso años de estudio de los 100 empleados que conforman la empresa.

La media es $\overline{X}$ = 9.94 años de estudio

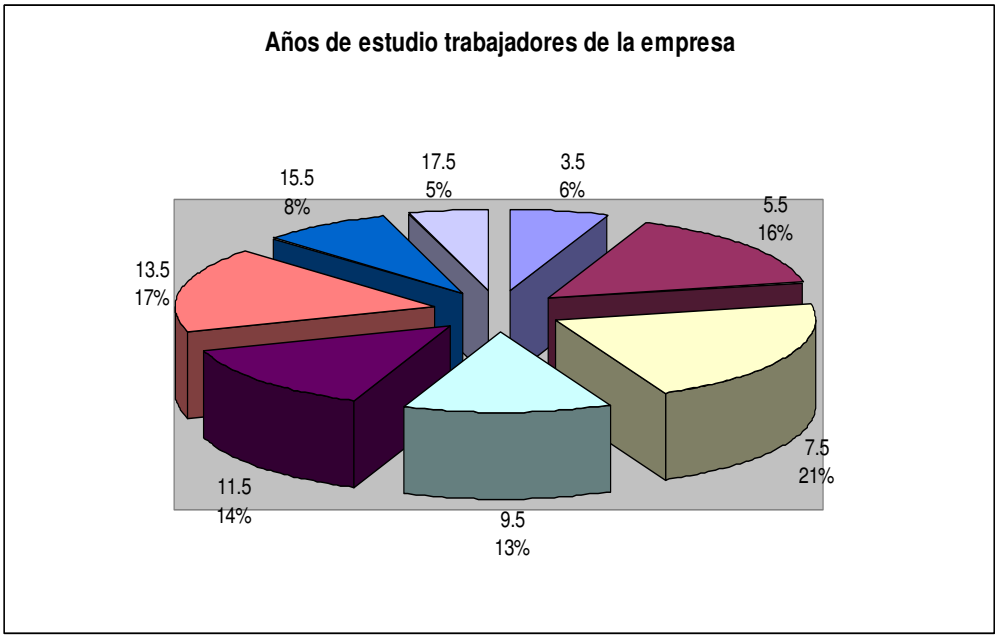
La varianza es S^2 = 14.71 de donde

La desviación estándar es S = 3.84 años de estudio

Anexo 2: Representaciones gráficas



Gráfica 1: grafica de barras para años de estudio de los 100 empleados que conforman la empresa. Elaborados con Microsoft Excel.



Gráfica 2: grafica tipo pie para años de estudio de los 100 empleados que conforman la empresa. Elaborados con Microsoft Excel.

Anexo 3: Formulario medidas de tendencia central y de dispersión

Medida	Estadístico	Parámetro	Distribución de frecuencia simple	Distribución de frecuencia agrupada
Moda	Mo	μo	Valor o categoría con mayor frecuencia f	Identificar la clase modal (la de mayor frecuencia) L1 = limite real inferior de la clase modal D1 = cantidad en que difiere con la clase anterior D2 = cantidad en que difiere con la calase posterior C = ancho del intervalo $L1+\left(\frac{D1}{D1+D2}\right)*C$
Media	$\overline{X}$	$\overline{\mu}$	$\frac{\sum_{i}^{n} x_i}{n}$	$\frac{\sum_{i}(x_i * f_i)}{n}$ $\sum_{i} x_i * fr_i$
Mediana	$\tilde{X}$	$\tilde{\mu}$	Si n es par $\frac{x_{n/2-1}+x_{n/2+1}}{2}$ Si n es impar $x_{\frac{(n-1)}{2}+1}$	Identificar la clase mediana (donde se encuentra el 50 %) L1 = limite real inferior de la clase mediana Fa = frecuencia acumulada anterior f = frecuencia de la clase mediana C = ancho del intervalo $L1+\left(\frac{n/2-Fa}{f}\right)*C$
Varianza	S^2	σ^2	$\frac{\sum_{i}^{n}(x_i-\overline{x})^2}{n-1}$ $\frac{\sum_{i}^{n} x_i^2-\frac{\left(\sum_{i}^{n} x_i\right)^2}{n}}{n-1}$	$\frac{\sum_{i}(x_i-\overline{x})^2}{n-1}$
Desviación estándar	S	σ	$\sqrt[2]{S^2}$	$\sqrt[2]{S^2}$

Anexo 4: Formulario para estimación puntual y por intervalos

Medida	Estadístico	Parámetro	Estimación puntual
Media	$\overline{X}$	$\overline{\mu}$	$\overline{\mu} = \overline{X}$
Varianza	S^2	σ^2	$\sigma^2 = S^2$
Desviación estándar	S	σ	$\sigma = S$
Proporción	$\hat{p}$	p	$p = \hat{p}$

Medida	Estadístico	Parámetro	Condición	Estimación por Intervalos	
				Bilateral	Unilateral
Media	$\overline{X}$	$\overline{\mu}$	n>30	$\overline{X} \pm Z_{\alpha/2} * \frac{s}{\sqrt{n}}$	$\overline{X} + Z_{\alpha} * \frac{s}{\sqrt{n}}$ $\overline{X} - Z_{\alpha} * \frac{s}{\sqrt{n}}$
			n<30 distribución normal	$\overline{X} \pm t_{\alpha/2,n-1} * \frac{s}{\sqrt{n}}$	$\overline{X} + t_{\alpha,n-1} * \frac{s}{\sqrt{n}}$ $\overline{X} - t_{\alpha,n-1} * \frac{s}{\sqrt{n}}$
			n>30 con σ^2 conocida	$\overline{X} \pm Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$	$\overline{X} + Z_{\alpha} * \frac{\sigma}{\sqrt{n}}$ $\overline{X} - Z_{\alpha} * \frac{\sigma}{\sqrt{n}}$
Varianza	S^2	σ^2	Distribución normal	Limite inferior $(n-1)S^2 / \chi^2_{\alpha/2,n-1}$ Límite superior $(n-1)S^2 / \chi^2_{1-\alpha/2,n-1}$	Inferior $(n-1)S^2 / \chi^2_{\alpha,n-1}$ Superior $(n-1)S^2 / \chi^2_{1-\alpha,n-1}$
Desviación estándar	S	σ		Raíz cuadrada del resultado obtenido para la varianza	Raíz cuadrada del resultado obtenido para la varianza
Proporción	$\hat{p}$	p	n>30	$\hat{p} \pm Z_{\alpha/2} * \sqrt{\hat{p}\hat{q}/n}$	$\hat{p} + Z_{\alpha} * \sqrt{\hat{p}\hat{q}/n}$ $\hat{p} - Z_{\alpha} * \sqrt{\hat{p}\hat{q}/n}$

Anexo 5: Modelos Estadísticos para Prueba de Hipótesis (Devore, años 1998 y 2001)

	Proporción	Media	Varianza
Muestras grandes una muestra	Hipótesis nula H_0: $p = p_0$ Estadístico de prueba: $$Z = \frac{\hat{p} - p_o}{\sqrt{p_o(1-p_o)/n}}$$ Hipótesis alternativa: Región de rechazo nivel α: Ha: $p > p_0$ $z \geq z_\alpha$ Ha: $p < p_0$ $z \leq -z_\alpha$ Ha: $p \neq p_0$ ya sea $z \geq z_{\alpha/2}$ o $z \leq -z_{\alpha/2}$	Hipótesis nula H_0: $\mu = \mu_0$ Estadístico de prueba: $$z = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$ Hipótesis alternativa: Región de rechazo nivel α: Ha: $\mu > \mu_0$ $z \geq z_\alpha$ Ha: $\mu < \mu_0$ $z \leq -z_\alpha$ Ha: $\mu \neq \mu_0$ ya sea $z \geq z_{\alpha/2}$ o $z \leq -z_{\alpha/2}$	Hipótesis nula H_0: $\sigma^2 = \sigma^2_0$ Estadístico de prueba: $\chi^2 = (n-1)\,S^2 / \sigma^2_0$ Hipótesis alternativa: Región de rechazo nivel α: Ha: $\sigma^2 > \sigma^2_0$ $\chi^2 \geq \chi^2_{\alpha,\,n-1}$ Ha: $\sigma^2 < \sigma^2_0$ $\chi^2 \leq \chi^2_{1-\alpha,\,n-1}$ Ha: $\sigma^2 \neq \sigma^2_0$ ya sea $\chi^2 \geq \chi^2_{\alpha/2,\,n-1}$ o $\chi^2 \leq \chi^2_{1-\alpha/2,\,n-1}$
Muestras grandes dos muestras	Hipótesis nula H_0: $p_1 - p_2 = 0$ Estadístico de prueba: $$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/m + 1/n)}}$$ $\hat{p}$ = estimador agrupado de $p = (x_1 + x_2) / (n_1 + n_2)$ $= n_1 \hat{p}_1 / (n_1 + n_2) + n_2 \hat{p}_2 / (n_1 + n_2)$ Hipótesis alternativa: Región de rechazo nivel α: Ha: $p_1 - p_2 > 0$ $z \geq z_\alpha$ Ha: $p_1 - p_2 < 0$ $z \leq -z_\alpha$ Ha: $p_1 - p_2 \neq 0$ ya sea $z \geq z_{\alpha/2}$ o $z \leq -z_{\alpha/2}$	Hipótesis nula H_0: $\mu_1 - \mu_2 = \Delta_0$ Estadístico de prueba: $$z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{(s^2_1/m) + (s^2_2/n)}}$$ Hipótesis alternativa: Región de rechazo nivel α: Ha: $\mu_1 - \mu_2 > \Delta_0$ $z \geq z_\alpha$ Ha: $\mu_1 - \mu_2 < \Delta_0$ $z \leq -z_\alpha$ Ha: $\mu_1 - \mu_2 \neq \Delta_0$ ya sea $z \geq z_{\alpha/2}$ o $z \leq -z_{\alpha/2}$	Hipótesis nula H_0: $\sigma^2_1 = \sigma^2_2$ Estadístico de prueba: $f = S^2_1 / S^2_2$ Cómo los valores críticos en tabla α = .1, .05, .01 y .001 entonces la prueba de dos colas se puede realizar solo a niveles α = .2, .1, .02 y .002 Hipótesis alternativa: Región de rechazo nivel α: Ha: $\sigma^2_1 > \sigma^2_2$ $f \geq F_{\alpha,\,m-1,\,n-1}$ Ha: $\sigma^2_1 < \sigma^2_2$ $f \leq F_{1-\alpha,\,m-1,\,n-1}$ Ha: $\sigma^2_1 \neq \sigma^2_2$ ya sea $f \geq F_{\alpha/2,\,m-1,\,n-1}$ o $f \leq F_{1-\alpha/2,\,m-1,\,n-1}$
Muestra pequeña una muestra	Hipótesis nula H_0: $p = p_0$ Por medio de la distribución Binomial P(Error tipo I) = $1 - B(c-1;\, n, p)$ P(Error topo II, cuando $p=p'$) = $B(c-1;\, n, p')$ Hipótesis alternativa: Región de rechazo nivel α: Ha: $p > p_0$ $x \geq c$ Ha: $p < p_0$ $x \leq c$ Ha: $p \neq p_0$ Valores para X pequeños y grandes	**Varianza poblacional conocida** Hipótesis nula H_0: $\mu = \mu_0$ Estadístico de prueba: $$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$ Hipótesis alternativa: Región de rechazo nivel α: Ha: $\mu > \mu_0$ $z \geq z_\alpha$ Ha: $\mu < \mu_0$ $z \leq -z_\alpha$ Ha: $\mu \neq \mu_0$ ya sea $z \geq z_{\alpha/2}$ o $z \leq -z_{\alpha/2}$ **Varianza poblacional desconocida** Estadístico de prueba: $$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$	

	Proporción	Media	Varianza
Muestra pequeña dos Muestras	Modelos mas complejos	**Varianzas poblacionales conocidas** Hipótesis nula H_0: $\mu_1 - \mu_2 = \Delta_0$ Estadístico de prueba: $z = \dfrac{x - y - \Delta_0}{\sqrt{(\sigma^2_1/m) + (\sigma^2_2/n)}}$ Hipótesis alternativa: — Región de rechazo nivel α: Ha: $\mu_1 - \mu_2 > \Delta_0$ — $z \geq z_\alpha$ Ha: $\mu_1 - \mu_2 < \Delta_0$ — $z \leq -z_\alpha$ Ha: $\mu_1 - \mu_2 \neq \Delta_0$ ya sea $z \geq z_{\alpha/2}$ o $z \leq -z_{\alpha/2}$ **La prueba t agrupada (cuando $\sigma^2_1 = \sigma^2_2$)** Hipótesis nula H_0: $\mu_1 - \mu_2 = \Delta_0$ Estadístico de prueba: $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{S_p\sqrt{(1/m + 1/n)}}$ Varianza ponderada: $S^2_p = \dfrac{(m-1)S^2_1 + (n-1)S^2_2}{m+n-2}$ Hipótesis alternativa: — Región de rechazo nivel α: Ha : $\mu_1 - \mu_2 > \Delta_0$ — $t \geq t_{\alpha, m+n-2}$ Ha : $\mu_1 - \mu_2 < \Delta_0$ — $t \leq -t_{\alpha, m+n-2}$ Ha : $\mu_1 - \mu_2 \neq \Delta_0$ ya sea $t \geq t_{\alpha/2, m+n-2}$ o $t \leq -t_{\alpha/2, m+n-2}$ **Prueba cuando $\sigma^2_1 \neq \sigma^2_2$** Estadístico de prueba: $t' = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{(S^2_1/m + S^2_2/n)}}$ Grados de libertad: $v = \dfrac{(S^2_1/m + S^2_2/n)^2}{\dfrac{(S^2_1/m)^2}{m-1} + \dfrac{(S^2_2/n)^2}{n-1}}$ Hipótesis alternativa: — Región de rechazo nivel α: Ha: $\mu_1 - \mu_2 > \Delta_0$ — $t' \geq t_{\alpha, v}$ Ha: $\mu_1 - \mu_2 < \Delta_0$ — $t' \leq -t_{\alpha, v}$ Ha: $\mu_1 - \mu_2 \neq \Delta_0$ ya sea — $t' \geq t_{\alpha/2, v}$	

REFERENCIA DE SIMBOLOS

1. Parámetros de interés θ

$\mathbf{p}$	Proporción
μ	Media
σ^2	Varianza
$p_1 - p_2$	Diferencia entre proporciones
$\mu_1 - \mu_2$	Diferencia entre medias
$\sigma^2_1 = \sigma^2_2$	Diferencia entre varianzas

2. Valores nulos θ_0

p_0	Proporción
μ_0	Media
σ^2_0	Varianza
$\Delta_0 (= \theta_1 - \theta_2)$	Diferencia entre dos parámetros

3. Estimadores $\hat{\theta}$

$\hat{p}$	Proporción muestral
$\bar{X}$	Media muestral
S^2	Varianza muestral

4. Estadísticos de prueba

z	valor normal estándar
χ^2	Valor ji - cuadrado
f	valor F
t y t'	valor t

5. Nivel de confianza

α	Alfa

6. Hipótesis

H_0	Hipótesis nula
H_a	Hipótesis alternativa

7. Otros estimadores

$\hat{p}$	estimador agrupado de p
S^2_p	Varianza ponderada
v	Grados de libertad