DOCUMENTO TÉCNICO N° 64
Versión 0.1

Consejo de Auditoría Interna General de Gobierno

Gobierno de Chile

MUESTREO ESTADÍSTICO PARA LA AUDITORÍA INTERNA DE GOBIERNO

CONCEPTOS GENERALES

Este documento es parte de una serie de guías técnicas que desarrollan elementos teóricos y prácticos sobre técnicas de muestreo estadístico para auditores internos gubernamentales.

Marzo 2015

MINISTERIO SECRETARÍA GENERAL DE LA PRESIDENCIA

CAIGG
Área de Estudios

TABLA DE CONTENIDOS

PRESENTACIÓN

Como una de las iniciativas tendientes al fortalecimiento de la Auditoría Interna considerado en el Programa de Gobierno de S.E. la Presidenta de la República, Michelle Bachelet; el Consejo de Auditoría Interna General de Gobierno, entidad asesora en materias de auditoría interna, control interno, gestión de riesgos y gobernanza, tiene el rol de promover la mejora continua de la función de auditoría interna gubernamental, y entregar recursos a la red de auditores para la generación de competencias y perfeccionamiento técnico de su trabajo, considerando las últimas tendencias de auditoría interna y las mejores prácticas aceptadas a nivel nacional e internacional.

En este ámbito, se pone a disposición de la red de auditores gubernamentales, el Documento Técnico N° 64, denominado "Muestreo Estadístico para la Auditoría Interna de Gobierno: Conceptos Generales". Este documento está concebido para ser una guía a ser aplicada por los profesionales de auditoría interna de la Administración del Estado, en el desarrollo de sus funciones, especialmente en la ejecución del trabajo del auditor en grandes poblaciones, donde tomar una muestra estadística representativa del universo, es clave para la validez de los resultados y conclusiones del trabajo.

El presente documento, es parte de una serie de guías técnicas que desarrollan elementos teóricos y prácticos sobre inferencia estadística y técnicas de muestreo para auditores internos gubernamentales.

Santiago, marzo 2015.

Daniella Caldana Fulss

Auditora General de Gobierno

I.- INTRODUCCIÓN

El Consejo de Auditoría, en cumplimiento de la Política de Auditoría Interna General de Gobierno, implementada y propiciada por el Ejecutivo para el fortalecimiento y desarrollo de los organismos, sistemas y metodologías que permitan resguardar los recursos públicos y apoyar la gestión de la administración y los actos de Gobierno ha formulado la versión 0.1 del Documento Técnico Nº 64 – "Muestreo Estadístico para la Auditoría Interna de Gobierno: Conceptos Generales".

Este documento es parte de una serie de guías técnicas que desarrollarán elementos teóricos y prácticos, así como también técnicas de muestreo para auditores internos gubernamentales.

En lo principal, este primer documento de la serie considera los conceptos generales sobre la investigación y el método estadístico. Comienza con una definición de conceptos generales de la especialidad, como representatividad, muestreo, población, muestra, parámetro y estadística. A continuación se detalla el método estadístico propiamente tal desde la definición del problema hasta las conclusiones. Se presenta la estadística descriptiva, los distintos tipos de distribución, estimadores, intervalo de confianza y tamaño de muestras. Se tratan también la hipótesis nula y la hipótesis alternativa y los distintos errores o riesgos presentes en un contraste de hipótesis. Por último se tratan los planes de muestreo, las formas de seleccionar una muestra, y los diversos tipos, como son el muestreo aleatorio simple, el sistemático, el estratificado y el por conglomerado.

II.- PROCESO DE OBSERVACIÓN

Un concepto importante que debemos conocer y tener siempre presente para realizar estudios con elementos de Estadística, es el "Proceso de Observación" que consiste en generar los datos que servirán de base para la elaboración de la información. Se debe tener máximo cuidado a la hora de realizar los registros, de forma tal que las características que se desean observar o medir sean totalmente confiables.

Esta situación es de real interés para que se pueda detectar lo que realmente se quiere medir. De no ser así, se tendrán resultados válidos pero las conclusiones que se obtengan no serán en ningún caso confiables. Esta situación se podrá lograr si los datos que servirán para la elaboración de la información, los reportes y conclusiones se obtienen de una **"Muestra Representativa"** y cuando una investigación satisface este procedimiento, se dice que la investigación tiene **"Validez Externa"**.

Cuando se satisface la validez externa, se tendrá información de calidad adecuada, con la cual se pueden emitir afirmaciones simples y complejas sobre la característica u objeto de observación pudiendo, incluso, encontrar nuevos hallazgos.

Las afirmaciones que se pueden emitir, pueden derivar en diversas situaciones, una de ellas es la **"Compatibilidad"** con el conjunto de proposiciones aceptadas como válidas, en la característica u objeto que se está estudiando.

En el caso de existir contradicciones o situaciones incompatibles, hay que resolver un nuevo problema, rechazando las afirmaciones emitidas como válidas y buscar las razones que justifican su rechazo, o bien, se replantean las proposiciones aceptadas como válidas buscando razones fundadas que justifiquen el nuevo comportamiento.
Cuando el resultado de esta valoración, crítica de las conclusiones, arroja resultado positivo, se dice que el estudio tiene **"Validez Interna"**.

III.- INVESTIGACIÓN Y EL MÉTODO ESTADÍSTICO

1.- Representatividad

Los estudios basados en métodos estadísticos, se distinguen por que las observaciones se obtienen de muestras probabilísticas, es decir, se eligen aleatoriamente de la población que es motivo de estudio y las inferencias que se realizan tienen naturaleza probabilística. De esta forma las conclusiones que se obtengan, estarán asociadas a niveles de confianza con la componente aleatoria involucrada. En esta situación, una componente adicional, es la **"Representatividad de la Muestra"**, que está asociada a la **validez externa** de la investigación.

Un criterio para valorar la representatividad de una muestra, consiste en valorar dos dimensiones esenciales:

- El mecanismo con el cual se seleccionan las unidades de la muestra (**Forma**), y
- El número de elementos a incluir en la muestra (**Cantidad**).

2. Muestreo

Es el mecanismo o la técnica con el cual se seleccionan las unidades de la muestra, es decir, la **"Forma"**.

Este mecanismo debe conservar la estructura de las características y las relaciones que se desean observar y que situaciones extremas se deban solamente al azar. Este mecanismo se puede resumir del siguiente modo:

"Todas las unidades de la población deben tener la misma probabilidad de ser seleccionadas en la muestra".

En poblaciones que tienen un número finito de elementos, por ejemplo N, la probabilidad de elegir uno de sus elementos es 1/N. Este mecanismo es válido cuando todos los elementos de la población tienen la misma característica, esto es que todas

las unidades de muestreo sean homogéneas y la técnica de selección de la muestra se conoce como **Muestreo Aleatorio Simple (MAS)**.

Cuando los elementos de la población no tienen características similares, las unidades de muestreo son heterogéneas, es decir, no todos los elementos tienen características de estudio similares, se forman subgrupos o estratos con elementos de la población de características homogéneas. En este caso la población quedará particionada en "h" estratos de tamaños N_h ; $h = 1,2,3, \dots, k$ y tal que $N = \sum_{h=1}^{k} N_h$.

De esta manera cada elemento del estrato "h" tiene probabilidad $1/N_h$ de representar al estrato en la muestra y en cada uno de ellos se puede utilizar muestreo aleatorio simple para seleccionar las unidades de la muestra en el estrato "h". Esta técnica de muestreo se conoce como **"Muestreo Estratificado".**

Otra situación se presenta cuando las unidades de la población están agrupadas, por ejemplo, en comunas, regiones, sectores, departamentos, ministerios. Si consideramos por ejemplo los ministerios, por su labor son homogéneos, pero en su interior los departamentos pueden tener características heterogéneas, en cuyo caso se puede recurrir al muestreo por conglomerados y los conglomerados serán los ministerios. La técnica de muestreo consiste en elegir al azar algunos conglomerados y dentro de ellos se seleccionarán departamentos y dentro de los departamentos las secciones.

Este procedimiento se conoce como **Muestreo por Conglomerados** trietápico. En el ejemplo, las ponderaciones se definen de acuerdo al número de ministerios (unidades primarias), al número de departamentos (unidades secundarias) y al número de secciones (unidades terciarias).

Cuando las unidades de la población están debidamente ordenadas o numeradas se puede utilizar el **Muestreo Sistemático** y la muestra se obtiene eligiendo las unidades de la muestra a partir de uno o más intervalos iniciales. Cuando las unidades se seleccionan a partir de un intervalo inicial (uno o más arranques aleatorios), el muestreo se llama **Sistemático Simple** y cuando se utiliza una combinación de

selección aleatoria y sistemática, el muestreo se llama muestreo **Aleatorio Sistemático**.

Cuando se establece una técnica de muestreo, como las anteriores, para seleccionar las unidades de la muestra, será una garantía de que la muestra es representativa, por su forma.

La otra dimensión de representatividad de la muestra está relacionada con el tamaño de la muestra, es decir, la cantidad de elementos que se seleccionarán de la población. Este criterio tiene directa relación con el nivel de precisión requerido. Intuitivamente no es difícil darse cuenta, que mientras más precisión se requiera, más grande será el tamaño de la muestra. La precisión de una estimación se puede expresar generalmente a través de dos elementos:

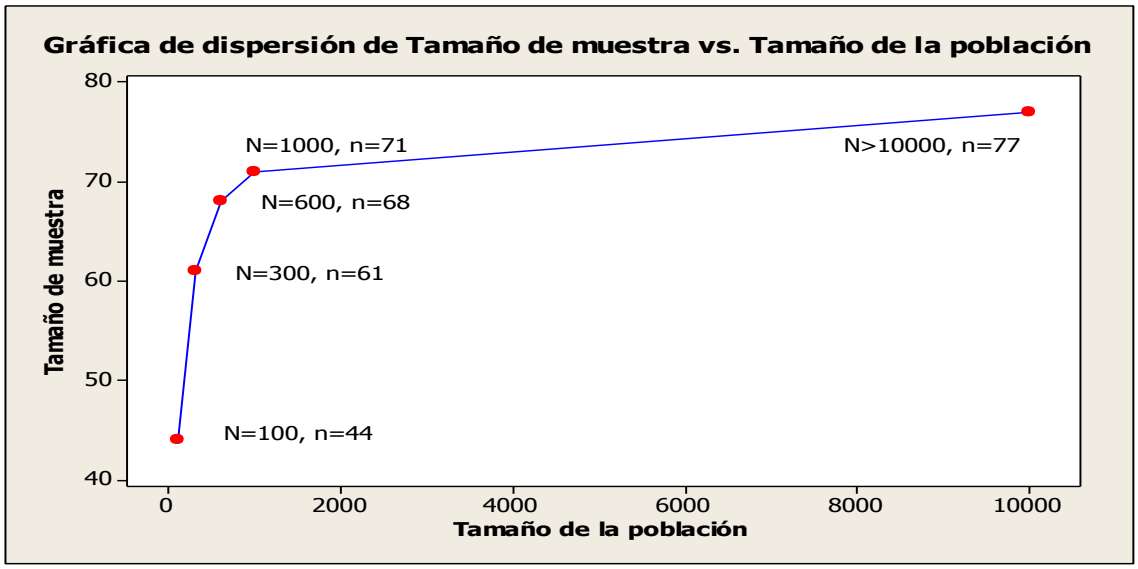
- Un error tolerable (ε).
- Una confianza (1-α), $\alpha \in (0\,;1{,}0)$.

El **Error Tolerable**, es la diferencia que estamos dispuesto a aceptar que exista entre el verdadero valor del parámetro θ, que es el valor del parámetro desconocido de la población y su estimador $\hat{\theta}_n$, que es el valor del parámetro estimado a partir de las unidades de la muestra.

La **Confianza**, es la probabilidad con la cual se desea no sea superado el error tolerable:

$$P\left(\left|\theta - \hat{\theta}_n\right| \leq \varepsilon\right) \geq 1 - \alpha\,; \quad n \to \infty$$

En la siguiente figura se muestra la relación entre tamaño de muestra n y tamaño de la población N, para un error tolerable 10% y confianza 95%, para la estimación de una proporción:



En investigaciones que utilizan la metodología Estadística con validez interna, la encargada de realizar y garantizar la **comparabilidad** es la **"Inferencia Estadística".**
La comparabilidad es un tema de cuidado, pues fácilmente se pueden cometer errores a la hora de obtener conclusión sobre una particular inferencia.

Por ejemplo, supongamos que se desea comparar el efecto que produce un aditivo aplicado al combustible de un vehículo con respecto al rendimiento en consumo de combustible en kilómetros por litros. Para tal efecto se registran los rendimientos de un buen número de vehículos y se aplican medidas estadísticas de asociación para detectar el grado de relación entre las dos características. Realizadas las operaciones, se detecta que las características tienen un grado de relación muy escaso. Al revisar los registros se descubre que los vehículos tenían diferentes años de circulación, situación que afectaba la comparación, es decir, no se tenía claridad si el efecto se debía al aditivo o a los años de circulación del vehículo. En este caso, los años de circulación del vehículo, sería la característica que hace la diferencia en las unidades observadas. Estas características se conocen como el **"Factor de Confusión"**.

Entonces, podemos concluir que la validez interna y la comparabilidad se logran controlando los factores de confusión. En nuestro ejemplo, para lograr la validez interna, se pueden agrupar los vehículos de acuerdo a los años de circulación, y de esta

forma dentro de cada grupo se puede realizar la comparación que se desea, salvo que existan otros posibles factores de confusión. Esta solución se conoce como "Construcción en bloques" y es materia del diseño de experimentos. La identificación de los factores de confusión es una tarea de responsabilidad del investigador y no del profesional estadístico.

3. Población

Población es el conjunto de elementos que son motivo de estudio, sobre los cuales se desea tener información y de los cuales se extenderán las conclusiones. El sentido de la palabra Población, no está exclusivamente relacionada con poblaciones demográficas, pues tiene sentido hablar de la población de peces que hay en un lago, de la población de artículos que produce diariamente una máquina, la población de facturas, etc.

En toda investigación, que requiera de métodos Estadísticos, la población debe estar definida con la máxima precisión, de forma tal que no aparezcan, más tarde, elementos mal clasificados, los cuales harán que se pierda la validez de las conclusiones que se puedan emitir.

La precisión con que se defina la población, estará fuertemente ligada al objetivo que se persiga en la investigación. Si el objetivo que se persigue no está bien especificado, lo más probable es que se tenga dificultad para definir la población.

Cuando las poblaciones son finitas, tienen un número determinado de elementos y este caso se dice que la población es de tamaño N. Cuando N no está determinado, la población será de tamaño infinita.

4. Muestra

Muestra es un conjunto de elementos que son parte de la población que es motivo de estudio. Cuando el número de elementos de la muestra coincide con los de la población, estamos en presencia de un Censo. En la práctica es poco frecuente realizar un censo, por costos y tiempo que significa realizarlo, por ello se recurre a una parte de la población, a una muestra.

El tamaño de la muestra se representa con "n" (n<N).

Existen conceptos de real interés antes de trabajar con una muestra, conceptos que se tratan en ambientes superiores, llamadas **"Técnicas de Muestreo"** o simplemente Muestreo. Conceptos básicos de estas técnicas de interés y fácil de entender, son los conceptos **"Factor de Elevación"** y **"Fracción de Muestreo"**, que se definen respectivamente de la siguiente forma:

$$f_e = \frac{N}{n} \quad ; \qquad f = \frac{n}{N}$$

La **Fracción de Muestreo** representa el porcentaje de elementos de la población que tiene la muestra y el **Factor de Elevación** indica la cantidad de elementos de la población que está representando cada elemento de la muestra.

5. Parámetro

Parámetro es una característica medible de la población, es una constante para la población. Cuando el o los parámetros de una población son desconocidos, la muestra nos permitirá realizar una estimación del o los parámetros de la población.

6. Estadística

Estadística es una característica medible de la muestra, en general es una función de la muestra. Las estadísticas se utilizan para formarse una idea de los parámetros de la

población y cuando esto sucede se llaman **"Estimadores"**. Una estadística varía de una muestra a otra y, en este sentido, se puede mirar como una variable y tratarse como tal.

IV.- MÉTODO ESTADÍSTICO

Consiste en realizar ordenadamente un listado de actividades, que se han previamente planeado, para llevar a cabo una investigación. Las actividades que planean con este método son las siguientes:

1. Definición del Problema

Esta actividad es la más relevante a la hora de iniciar una investigación con método estadístico, pues se debe justificar claramente el estudio que se quiere llevar a cabo. Para lograr este objetivo, se deben tener presente los siguientes pasos:

- Determinar los objetivos del estudio.
- Especificar la bibliografía y experiencias anteriores.
- Formular las hipótesis.
- Definir los parámetros que se desean estimar.
- Especificar la precisión con la que se estimarán los parámetros.

2. Definición de la Población

Para definir la población se deben tener presente todas las indicaciones que se dieron en la definición de población.

3. Definición de la Estrategia de Análisis

En esta actividad se debe planear y definir la forma inicial en que se enfrentará el problema en estudio. Se puede recurrir a algunas técnicas estadísticas que ayuden a un descubrimiento preliminar de la situación.

De esta forma y en la medida que se encuentre información relevante para el análisis, se puede ir modificando el plan inicial e intentar formular un plan definitivo de acción.

4. Determinar las Variables

Esta actividad tiene por objetivo definir las características de la población que entregarán la información necesaria para lograr el éxito de los objetivos formulados en el estudio.

5. Diseño del Estudio

Esta actividad consiste en definir si se tomará un censo o una muestra de la población en estudio. En el caso de optar por una muestra, se deberá definir la forma o técnica de muestreo que se utilizará. Además se debe especificar el error de muestreo y el nivel de confianza para determinar el tamaño de la muestra.

6. Recolección de la Información

Esta es una de las actividades más importantes de la metodología ya que de ella depende la calidad de la información que se tenga para el estudio. Las personas o instrumentos responsables de recolectar la información, deben estar capacitadas adecuadamente o bien calibrados, para que la información carezca de errores importantes, ya sea por observación o por medición.

7. Procesamiento de los Datos

El procesamiento de datos consiste en aplicar las técnicas que proporciona la Estadística descriptiva, la cual organiza la información permitiendo una rápida y mejor comprensión. Para ello proporciona tablas, gráficos e indicadores que facilitan la interpretación de las variables consideradas en el estudio (etapa exploratoria o descriptiva) y permiten hacerse una idea del comportamiento de la población.

8. Inferencia Estadística

Inferencia Estadística es un proceso inductivo que permite hacer proposiciones sobre toda la población, basadas en observaciones y resultados proporcionados por la muestra.

Hacer inferencias acerca de la población tiene un factor de incertidumbre, es por esta razón que la Inferencia Estadística se ampara en la teoría de la probabilidad, la cual no evita los posibles errores pero sí los puede controlar, es decir, los puede cuantificar de acuerdo a un nivel de confianza previamente determinado.

9. Conclusiones

En esta etapa se formulan en forma clara, de acuerdo con los objetivos específicos, todas las conclusiones que logren explicar el cumplimiento de dichos objetivos, señalando alcances, limitaciones y se hacen sugerencias sobre nuevas hipótesis que pudieran surgir del análisis.

V.- ESTADÍSTICA DESCRIPTIVA

La Estadística Descriptiva, constituye una etapa de la metodología Estadística y en ella no se ven involucrados elementos de teoría de probabilidad como herramienta para poder realizar inferencias acerca de la población. La Estadística Descriptiva proporciona herramientas para calcular indicadores, hacer comparaciones, construir tablas y diagramas con el interés de obtener conocimiento de la población de la cual se tomó la muestra.

Los resultados obtenidos, al procesar los datos de una muestra, permiten obtener información que puede ser usada con fines exploratorios, para plantear hipótesis o como elementos para la inferencia estadística.

La Estadística Descriptiva aporta a los análisis de la información indicadores de lugar y de dispersión. Los indicadores más requeridos son la Media Aritmética (Promedio), Mediana, Varianza y Desviación Típica o Estándar, Coeficiente de Variación de Asimetría y Curtosis, y Medidas de Asociación y Correlación.

El promedio o media es un indicador que se ampara en la simetría y la mediana es un indicador más robusto que el promedio y es utilizado cuando existe simetría y más aún cuando existe asimetría.

La varianza es un indicador de dispersión respecto al promedio del cual se desprende la desviación típica, que es el indicador de dispersión usual.

El coeficiente de variación es un indicador que no tiene dimensión y expresa el porcentaje de dispersión relativo al promedio.

La medida de asociación más conocida es la covarianza y la de correlación el coeficiente de correlación.

Fórmulas de algunos de estos indicadores son las siguientes:

Promedio: $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

Varianza: $S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$

Desviación Típica: $S_X = \sqrt{S_X^2}$

Coeficiente de Variación: $CV_X = \frac{S_X}{\overline{X}} 100\%$

Todos los programas estadísticos calculan todos los indicadores descriptivos, incluso el programa Excel®[1] también realiza el cálculo de la mayoría de estos indicadores.

Por otra parte, la **Teoría de Probabilidad** y el cálculo de probabilidades constituyen la base teórica y práctica para el desarrollo de la **Inferencia Estadística,** así como la Inferencia Estadística es la base para otras teorías, en particular para las **Técnicas de Muestreo**.

Los modelos de probabilidades, en estudios paramétricos, son la base de los resultados teóricos de la Inferencia Estadística y de las Técnicas de Muestreo.

El modelo de probabilidades más requerido es el modelo de **Probabilidad Normal**, conocida como la **Distribución Normal de Gauss** (Campana de Gauss).

[1] Excel®, ACL® e IDEA® son productos y marcas comerciales registradas y de propiedad de sus respectivos dueños.

1.- Distribución Normal y Distribución Normal Estándar

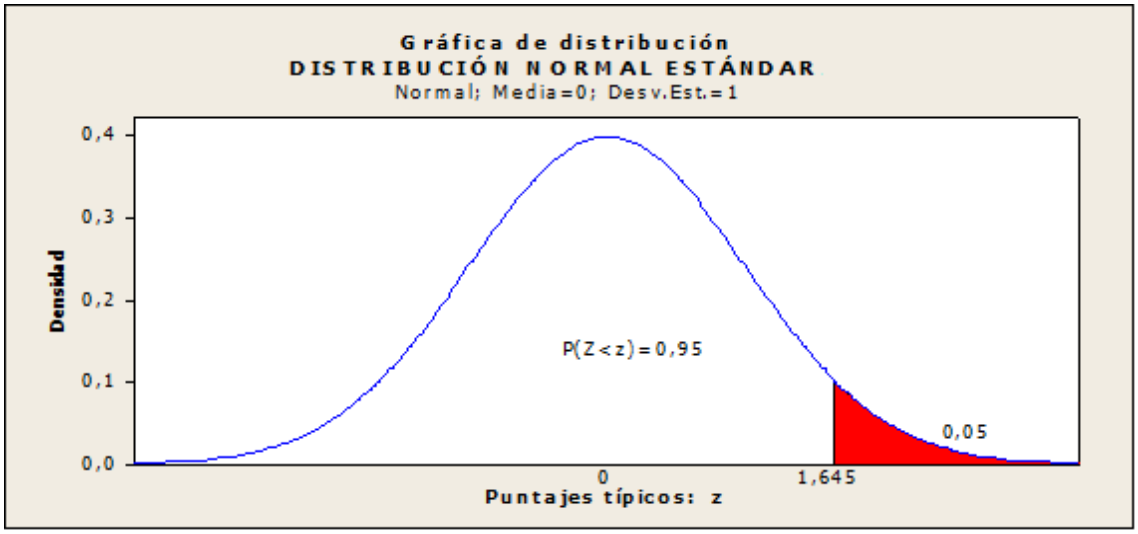
Este modelo de probabilidad se utiliza con mucha frecuencia, sobre todo cuando la cantidad de información disponible de una variable es bastante grande. El modelo depende de dos parámetros, que corresponden a la esperanza y varianza de la variable. La notación usual para estos parámetros es μ y σ^2 para la Esperanza y Varianza respectivamente. Para una determinada característica (Variable) que se distribuya bajo un modelo Normal, la notación usual es $X \approx N(\mu\,;\,\sigma^2)$.

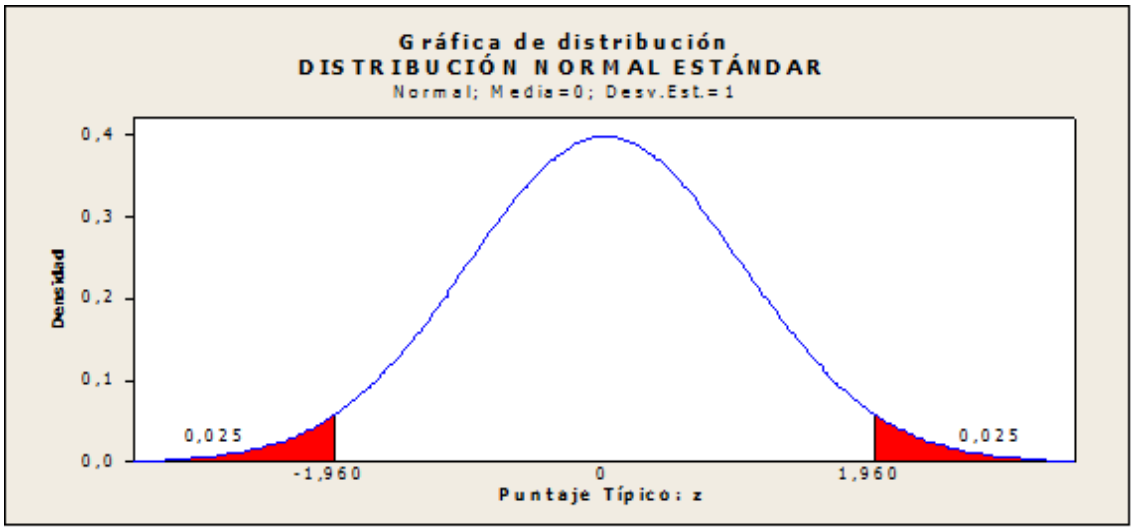
La Distribución Normal Estándar, se obtiene "**estandarizando o tipificando**" la variable X, de la siguiente forma: $Z = \frac{X-\mu}{\sigma} \approx N(0\,;1)$

El cálculo de probabilidades se realiza con la distribución Normal Estándar ya que existen tablas para este propósito. Algunos programas o software estadísticos y también Excel® proporcionan los cálculos de probabilidad para la Distribución Normal y Normal Estándar.

De acuerdo a las leyes de convergencia de la Teoría de Probabilidades, cuando las muestras son grandes $(n \to \infty)$ las distribuciones que tienen Esperanza y Varianza finita convergen al modelo de probabilidad normal.

Para encontrar los puntajes típicos de la variable Z, se puede utilizar la tabla de la función de distribución de esta variable. La función de distribución, para cada puntaje típico entrega la probabilidad acumulada hasta el puntaje típico, como se ilustra en las siguientes figuras para los puntajes típicos $z = 1{,}645$ y $z = 1{,}96$:





La distribución normal estándar proporciona puntajes típicos que son de bastante utilidad en la práctica. Estos puntajes son simétricos respecto del origen y se obtienen de la tabla de la función de distribución de Z.

Puntaje Típico	Probabilidad Centrada
$z = 0{,}97$	0,668
$z = 1{,}645$	0,90
$z = 1{,}96$	0,95
$z = 2{,}65$	0,992
$z = 3{,}03$	0,9976

Z	0	1	2	3	4	5	6	7	8	9
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
Z	0	1	2	3	4	5	6	7	8	9

2.- Distribución Bernoulli, Binomial y Poisson

La distribución Bernoulli, es un modelo de probabilidad discreto bastante sencillo y es la base de otros modelos variables aleatoria discreta. Este modelo se caracteriza por tener sólo dos resultados posibles: Éxito o Fracaso.

La probabilidad de éxito se designa con la letra p y la probabilidad de fracaso con la letra q, donde $q = 1 - p$.

La función de probabilidad de este modelo está definida de la siguiente manera:

$$P(X = x) = \begin{cases} 1 - p & si \;\; x = 0 \\ p & si \;\; x = 1 \end{cases}$$

El valor esperado de esta variable es $E(X) = p$ y su varianza es $V(X) = p(1 - p)$. Cuando se realizan "n" experimentos independientes Bernoulli X, se obtiene el modelo de probabilidad Binomial $Y = \sum_{i=1}^{n} X_i$. La función de probabilidad del modelo Binomial es la siguiente:

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \;\; ; \;\; y = 1,2,3, \dots, n$$

Y= Número de éxitos "y" en las "n" pruebas independientes Bernoulli. La esperanza y la varianza de esta variable, son respectivamente $E(Y) = np$, $V(Y) = np(1 - p)$.

Otro modelo de probabilidad relacionado, es el modelo de probabilidad Poisson que involucra todos los sucesos de Poisson. Los sucesos de Poisson, son sucesos cuya ocurrencia se cuenta en un intervalo de tiempo o espacio. La función de probabilidad de este modelo es:

$$P(T = t) = \frac{e^{-\theta} \theta^t}{t!} \;\; ; t = 1,2,3, \dots, \infty$$

$T =$ Números de sucesos de Poisson que ocurren en el instante t. La esperanza y la varianza de esta variable son respectivamente, $E(T) = \theta$; $V(T) = \theta$

El modelo de probabilidad Normal es un modelo continuo, los modelos Bernoulli, Binomial y Poisson son modelos discretos. Estos modelos, son modelos que trabajan bajo incertidumbre, por tal razón los resultados que se obtienen de ellos son resultados probables por tanto pueden estar cercanos o alejados de la realidad, todo dependerá si el modelo se ajusta a la variable que se está estudiando.

En las pruebas de control en auditoría, se utiliza el modelo Bernoulli considerando que el control es correcto o incorrecto, donde el control de auditoría corresponde a un "Fracaso" si este es correcto, y por el contrario corresponde a un "Éxito" en el caso que el control sea incorrecto. Esto dado a que el auditor como principal objetivo debe encontrar desviaciones o excepciones para aceptar o rechazar el control de procesos.

En cambio la Distribución Normal se relaciona mayormente con las pruebas sustantivas o de variables.

VI.- INFERENCIA ESTADÍSTICA

Cuando se desea investigar con objetivos específicos una determinada población, la Inferencia Estadística entrega los mecanismos que aseguran la elección de una muestra representativa que se ha seleccionado aleatoriamente, con las técnicas descritas anteriormente.

La Inferencia es inductiva y demanda una muestra aleatoria para que se pueda formular una inferencia respecto a alguna característica de la población que se está investigando. La investigación se realiza bajo un objetivo que de una característica medible de la población.

Una característica medible de una población es una variable que se desprende de un objetivo específico, de una desviación o de una excepción de dicha población.

Cuando una investigación se diseña para proporcionar la observación X_1 de la característica medible X y luego se repite bajo las mismas condiciones para proporcionar la observación X_2 y se continúa hasta obtener n las observaciones de la característica X, en tal caso las observaciones se recolectan mediante selecciones independientes.

1.- Muestra Aleatoria

Una muestra aleatoria de tamaño "n", es un conjunto de variables aleatorias seleccionadas de una población X, que tiene una determinada distribución de probabilidad y que cumplen las siguientes condiciones:

- Las variables aleatorias son mutuamente independientes.
- Todas las variables aleatorias tiene la misma distribución de probabilidad de la variable aleatoria de la población.

2.- Estadístico y Estimador

Toda función de la muestra aleatoria, que no contenga parámetros desconocidos, será una **Estadística.** Un **Estimador** es una Estadística que representa al parámetro desconocido de la población**.** No todas las Estadísticas son buenos Estimadores y tanto el uno como el otro son variables aleatorias.

3.- Notación de Estimador

Cuando el parámetro μ desconocido de un modelo de probabilidad, se desea estimar mediante una muestra aleatoria de tamaño "n", se designa $\hat{\mu}$ su estimador.

4.- Estimadores de Interés

Los Estimadores puntuales de los parámetros desconocidos de una distribución normal, son las siguientes estadísticas:

$\hat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ "Media de la Muestra"

$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$ "Varianza de la Muestra"

Estimadores puntuales de parámetros de las distribuciones Bernoulli, Binomial y Poisson, son los siguientes:

Bernoulli: Parámetro $p \rightarrow$ Estimador $\hat{p} = \overline{X}$

Binomial: Parámetro $p \rightarrow$ Estimador $\hat{p} = \frac{\overline{X}}{N}$

Poisson: Parámetro $\theta \rightarrow$ Estimador $\hat{\theta} = \overline{X}$

5.- Distribución de los Estimadores de los Parámetros

a.- De una Población Normal

$$\overline{X} \approx N\left(\mu; \frac{\sigma^2}{n}\right)$$

Las estandarizaciones de la media de la muestra son las siguientes:

$$\boldsymbol{T = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0\,;1)} \quad \text{o bien} \quad \boldsymbol{T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \approx t_{(n-1)}}$$

b.- De una Población Bernoulli

Distribución asintótica del estimador de p de un modelo Bernoulli

$$\boldsymbol{\hat{p} = \overline{X} \approx N\left(p\,; \frac{p(1-p)}{n}\right)} \quad \boldsymbol{T = \frac{\overline{X} - p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} \approx N(0\,;1),\ n \rightarrow \infty}$$

6.- Intervalos de Confianza

En muchas situaciones se recurre a intervalos de confianza en lugar de estimar puntualmente el parámetro desconocido. Para determinar intervalos de confianza es necesario tener una muestra aleatoria, el estimador del parámetro desconocido, la confianza y tener una Estadística que comprometa al parámetro y al estimador tal que tenga una distribución de probabilidad conocida.

Las distribuciones de Estadísticas **T** anteriores, permiten determinar intervalos de confianza para la media μ y σ^2 de una población normal y para el parámetro p de una distribución Bernoulli.

Genéricamente para una confianza de $(1-\alpha)100\%$ de confianza, los intervalos que se obtienen son los siguientes:

$$Para\ \mu: \qquad \left[\overline{X} - z_{\left(1-\frac{\alpha}{2}\right)}\frac{\sigma}{\sqrt{n}}\ ;\ \overline{X} + z_{\left(1-\frac{\alpha}{2}\right)}\frac{\sigma}{\sqrt{n}}\right] \quad cuando\ \sigma\ es\ conocida$$

$$Para\ \mu: \qquad \left[\overline{X} - t_{\left(n;\,1-\frac{\alpha}{2}\right)}\frac{S}{\sqrt{n}}\ ;\ \overline{X} + t_{\left(n;\,1-\frac{\alpha}{2}\right)}\frac{S}{\sqrt{n}}\right] \quad cuando\ \sigma\ es\ desconocida$$

$$Para\ p: \qquad \left[\hat{p} - z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ;\ \hat{p} + z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Los valores de $z_{\left(1-\frac{\alpha}{2}\right)}$ y de $t_{\left(n;\,1-\frac{\alpha}{2}\right)}$ corresponden a puntajes típico de la distribución normal y de la distribución t respectivamente. Para una confianza de 95% se tendrá el siguiente valor de α:

$$(1-\alpha) = 0{,}95 \ \rightarrow\ \alpha = 0{,}05 \ \rightarrow \frac{\alpha}{2} = 0{,}025 \ \rightarrow \left(1-\frac{\alpha}{2}\right) = 0{,}9750 \ \rightarrow z_{0{,}9750} = 1{,}96$$

En los tres intervalos anteriores, las siguientes expresiones tienen nombres usuales:

Error de Estimación: $z_{\left(1-\frac{\alpha}{2}\right)}\frac{\sigma}{\sqrt{n}}$; Error de Muestreo: $\frac{\sigma}{\sqrt{n}}$

Error de Estimación: $z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$; Error de Muestreo: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

7.- Tamaños de Muestras

Cuando las poblaciones tienen un gran número de elementos, se dice que son poblaciones infinitas entonces se puede utilizar el **error de estimación** para determinar el tamaño de la muestra. Este procedimiento no se utiliza frecuentemente pero es una alternativa para determinar el tamaño de la muestra.

Para determinar n, se debe conocer la varianza de la población o estimarla a partir de una muestra piloto y se debe también especificar la confianza para determinar $z_{\left(1-\frac{\alpha}{2}\right)}$ y el valor del error "e" de estimación que se está dispuesto aceptar, así n se obtiene de:

$$n = \left(\frac{z_{\left(1-\frac{\alpha}{2}\right)} \cdot \sigma}{e}\right)^2$$

Para el caso de una proporción, se debe especificar además un valor esperado "p_0" y n se obtiene de:

$$n = \frac{z_{\left(1-\frac{\alpha}{2}\right)}^2 \cdot p_0(1-p_0)}{e^2}$$

8.- Contraste de Hipótesis

Cuando se utilizan muestras aleatorias para estudios estadísticos, los Estadísticos que estiman los parámetros de la población, son aleatorios. Es más, distintas muestras aleatorias pueden generar valores diferentes de estos Estadísticos. Por ejemplo, si tomamos dos muestras al azar de una misma población, la diferencia que puede ocurrir entre las medias, se radica exclusivamente en las fluctuaciones de la muestra y, en tal caso, no interesaría rendirle un tratamiento diferente a ambos grupos, por el contrario, si son distintos, uno de ellos podría convertirse en un segmento objetivo. El contraste de hipótesis se trata de depurar los resultados mediante un estudio más riguroso o más profundo.

Una hipótesis estadística es una afirmación o conjetura con respecto a alguna característica de interés de la población en estudio. Una característica puede ser uno o varios parámetros hasta un modelo en particular. Por ejemplo, se puede formular la

hipótesis de que el monto promedio de los subsidios otorgados a un determinado segmento, de una determinada región o comuna es $ xx, o que la proporción de documentos que no cumplen con los controles es inferior a x%.

Contrastar una hipótesis estadística, es decidir si la afirmación se encuentra apoyada por la información o la evidencia que nos proporciona la muestra. De esta forma, los resultados de la muestra se utilizan para aceptar o rechazar la hipótesis o la afirmación formulada.

Conceptos básicos que se deben tener presente para formular y decidir adecuadamente un contraste de hipótesis, son los siguientes:

9.- Hipótesis Nula

Esta hipótesis es la hipótesis de contraste, que surge examinando los resultados de una encuesta, de estudios similares realizados anteriormente o bien de la propia experiencia de quién está enfrentado el estudio.

10.- Hipótesis Alternativa

Esta hipótesis se conoce como la hipótesis de la investigación o del investigador, es la hipótesis complementaria a la hipótesis nula.

El contraste se puede formular de las siguientes formas, dependiendo el objetivo del contraste:

$$H_0: \mu = \mu_0 \quad v/s \; H_1: \mu \neq \mu_0$$
$$H_0: \mu = \mu_0 \quad v/s \; H_1: \mu > \mu_0$$
$$H_0: \mu = \mu_0 \quad v/s \; H_1: \mu < \mu_0$$

11.- Región Crítica

Cuando se realizan contrastes referentes a parámetros de una población, para poder discrepar con respecto a la hipótesis nula se recurre a una muestra aleatoria. La

muestra aleatoria aporta la Estadística que servirá de base para rechazar o aceptar esta hipótesis. Como esta Estadística es una variable aleatoria, seguramente estará asimilada por una distribución de probabilidad conocida (la misma de la población) entonces dará origen a un estadístico de prueba, que también se conoce como **"Discrepancia"** o simplemente **"Estadístico de Prueba"**.

La discrepancia, que será una cantidad conocida se compara con una región, que es un Intervalo de la recta real, según sea el lado del contraste, que puede ser de dos lados, de lado derecho o de lado izquierdo. Esta región se llama **"Región Crítica"**, y es la región donde se rechaza la hipótesis nula.

Para contrastes de parámetros de modelos de probabilidad más conocidos, las distintas discrepancias se conocen y también sus respectivas regiones críticas, como una forma de ayudar a aquellos profesionales no involucrados con estas teorías pero que necesitan de sus resultados.

El complemento de la región crítica, se llama "**Región de Aceptación**", y en esta región se acepta la hipótesis nula.

12.- Error o Riesgos de un Contraste de Hipótesis

En un contraste de hipótesis se pueden tomar dos decisiones: Rechazar la hipótesis nula o bien no rechazar la hipótesis nula. Al tomar una u otra decisión se pueden cometer errores, que también se llaman riesgos. Existen dos tipos de errores:

a. Error de Tipo I

Se comete este error cuando se rechaza la hipótesis nula siendo esta verdadera. El riesgo de un rechazo incorrecto es conocido como el **Riesgo Alfa** (Riesgo α) y es relativo a la **eficiencia de la auditoría.**

b. Error de Tipo II

Se comete este error cuando se acepta la hipótesis nula siendo esta falsa. El riesgo de una aceptación incorrecta es conocida como el **Riesgo Beta** (Riesgo β) y es relativo a la **efectividad de la auditoría**.

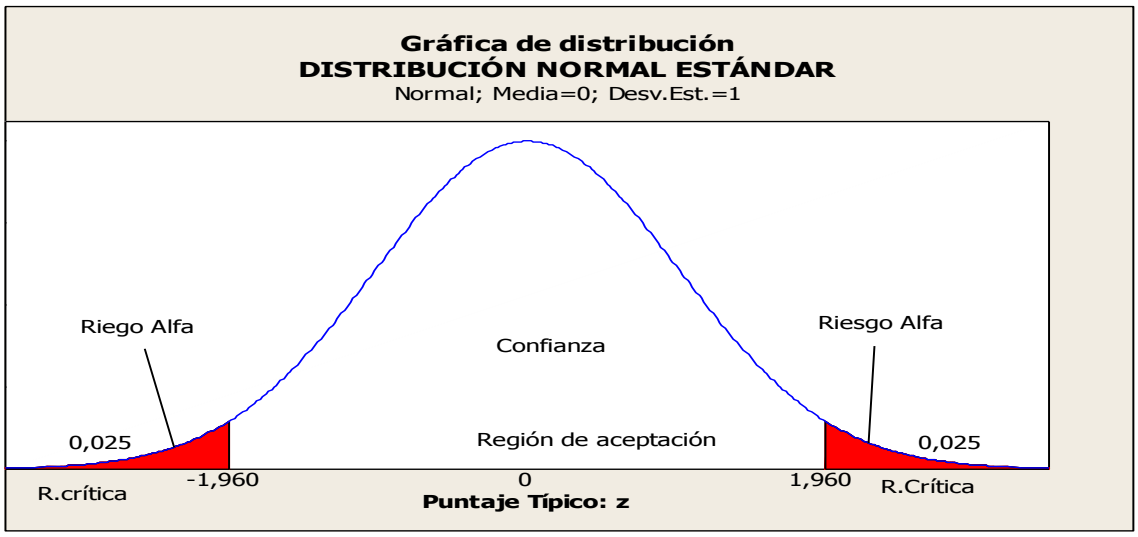
$Estados/Decisión$	Aceptar H_0	Rechazar H_0
H_0 verdadera	Decisión correcta	Error tipo I (Riesgo α)
H_1 verdadera	Error tipo II(Riesgo β)	Decisión correcta

13.- Nivel de Significación

Se anota generalmente con la letra griega α y representa el área bajo la curva sobre la región crítica, cuando la hipótesis nula es verdadera. Para el contraste de dos lados, el nivel de significación α, se puede visualizar en la siguiente figura:

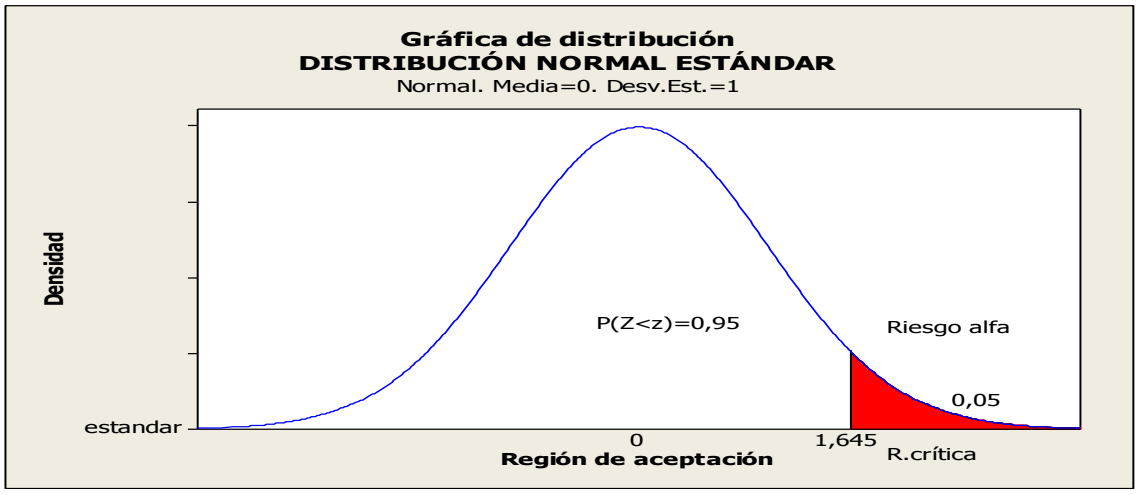
Contrastes:

$H_0: \mu = \mu_0 \quad v/s \; H_1: \mu \neq \mu_0$



Contrastes:

$H_0: \mu = \mu_0 \quad v/s \; H_1: \mu > \mu_0$



14.- Hipótesis Alternativa y la Región Crítica

La probabilidad de cometer error depende del tamaño de la muestra y, generalmente esta probabilidad disminuye en la medida que el tamaño de la muestra aumenta. También la probabilidad de cometer error depende de la región crítica que está regulada por el nivel de significación. En efecto, la probabilidad de cometer un **Error Tipo I**, cuando la hipótesis nula es verdadera, es igual al **Nivel de Significación**. Entonces, elegir un nivel pequeño significa asegurar una baja probabilidad de equivocarse, al tomar una decisión sobre un determinado contraste.

La hipótesis alternativa está ligada con la región crítica, porque indica el "lado" en la cual se sitúa esta región. Por ejemplo, si la hipótesis alternativa se encuentra a la derecha e izquierda de la hipótesis nula, la región crítica se encontrará al lado derecho e izquierdo de la región de aceptación y contraste, la que se llama bilateral. Si está a la izquierda de la hipótesis nula, la región crítica está a la izquierda en la recta real, y si está situada a la derecha la región crítica estará a la derecha. En los dos últimos casos, se dice que el contraste es de lado izquierdo o derecho según sea el caso respectivamente.

Para ilustrar la situación consideremos un contraste para la media de una población normal:

Contrastes	Lado	Discrepancia	Región Crítica
$H_0: \mu = \mu_0 \quad v/s \quad H_1: \mu \neq \mu_0$	Dos lados	$T = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$-\infty; -z_{\left(1-\frac{\alpha}{2}\right)}$ y $z_{\left(1-\frac{\alpha}{2}\right)}; \infty$
$H_0: \mu = \mu_0 \quad v/s \quad H_1: \mu > \mu_0$	Derecho	$T = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$z_{\left(1-\frac{\alpha}{2}\right)}; \infty$
$H_0: \mu = \mu_0 \quad v/s \quad H_1: \mu < \mu_0$	Izquierdo	$T = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$-\infty; -z_{\left(1-\frac{\alpha}{2}\right)}$

Para contrastar una proporción poblacional, se tiene lo que sigue:

Contrastes	Lado	Discrepancia	Región Crítica
$H_0: p = p_0 \quad v/s \quad H_1: p \neq p_0$	Dos lados	$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$-\infty; -z_{\left(1-\frac{\alpha}{2}\right)}$ y $z_{\left(1-\frac{\alpha}{2}\right)}; \infty$
$H_0: p = p_0 \quad v/s \quad H_1: p \neq p_0$	Derecho	$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$z_{\left(1-\frac{\alpha}{2}\right)}; \infty$
$H_0: p = p_0 \quad v/s \quad H_1: p \neq p_0$	Izquierdo	$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$-\infty; -z_{\left(1-\frac{\alpha}{2}\right)}$

15.- Acción o Decisión

Cuando la discrepancia cae en la región crítica, existe evidencia empírica para rechazar la hipótesis nula, de lo contrario no se rechaza la hipótesis nula, pero esto no garantiza que la hipótesis sea cierta, solo que no existe evidencia, apoyada en la muestra, de que sea falsa.

16.- P-valor

En la mayoría de los Contrastes Estadísticos, para tomar una decisión se recurre al "p-valor", que es el nivel mínimo que se debería tomar para rechazar la hipótesis nula. Todos los programas estadísticos traen incorporado este valor, independiente que uno

pueda utilizar un nivel de significación para recurrir a la región crítica y decidir el contraste.

Un **criterio utilizado** para aceptar o rechazar la hipótesis nula, con el **p-valor**, es el siguiente: "Si p-valor es menor o igual que 5%, se rechaza la hipótesis nula, de lo contrario se acepta la hipótesis nula".

17.- Planes de Muestreo

Cuando se formula un contraste simple para el parámetro de la población, se formula:

$$H_0: \mu = \mu_0 \quad v/s \; H_1: \mu = \mu_1$$

Este contraste será de lado derecho si $\mu_1 > \mu_0$ y será de lado izquierdo si $\mu_1 < \mu_0$.

Especificando la región crítica de acuerdo al lado respectivamente $\{\overline{X} > c\}$ o $\{\overline{X} \leq c\}$ y fijando los riesgos Alfa y Beta $(\alpha \; y \; \beta)$, se puede determinar el plan de muestreo $(n\,;c)$, donde "n" será el tamaño de la muestra y "c" será el valor donde comienza o termina la región crítica según ser el contraste.

Cuando $\mu_1 < \mu_0 \rightarrow RC = \{\overline{X} \leq c\}$. Para este caso se obtiene que:

$$c = z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0 \quad , \quad c = z_{1-\beta} \frac{\sigma}{\sqrt{n}} + \mu_1 \quad , \quad n = \left(\frac{\sigma(z_{1-\beta} - z_\alpha)}{\mu_0 - \mu_1} \right)^2$$

18.- Riesgo Alfa y Riesgo Beta

Riesgo Alfa: Riesgo de Rechazo Incorrecto

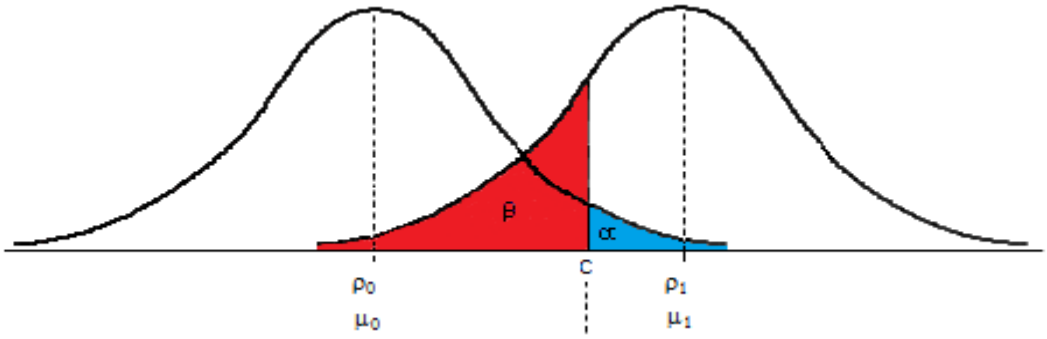
Riesgo Beta: Riesgo de Aceptación Incorrecta

Contrastes:

$$H_{0:}\ \rho = \rho_0 \quad V/S \quad H_1: \rho = \rho_1$$

$$H_{0:}\ \mu = \mu_0 \quad V/S \quad H_1: \mu = \mu_1$$

$H_{0:}$ Verdadero $\qquad$ H_1: Verdadero



Aceptar H_0		Rechazar H_1
$\overline{X} > c$	➔	Rechazar H_0 / H_0 Verdadero
$\overline{X} \leq c$	➔	Aceptar H_0 / H_1 Verdadero

VII.- FORMAS DE SELECCIONAR UNA MUESTRA

1.- Muestreo Aleatorio Simple

Tenemos una población de N elementos homogéneos, por tanto cualquiera de los elementos de la población puede formar parte de la muestra de tamaño n que se desea seleccionar. Una forma de seleccionar al azar las n unidades de la muestra, es

recurrir a una tabla de números aleatorios mediante un software estadístico en análisis de datos y muestra o Excel®.

Ejemplo 1. Seleccionar una muestra al azar de 10 documentos de un archivo, los cuales están numerados del 1 al 500 utilizando números aleatorios.

Forma 1. Seleccionar números de 3 cifras y eliminar los mayores de 500.

Forma 2. Seleccionar números de 3 cifras y a los mayores de 500 restar 500.

Ejemplo 2. Seleccionar una muestra de 10 documentos numerados desde 2500 hasta 8800 con números aleatorios. Elegimos números de 4 cifras y se descartan los que no estén entre 2500 y 8800.

Ejemplo 3. Seleccionar una muestra de 4 documentos que están numerados desde 5000 a 12000. Al seleccionar números de 5 dígitos se descartará una gran cantidad de números. Para tal efecto se elige una constante 4 dígitos, por ejemplo 3000, y se eligen números entre 2000 y 9000.

Luego a los números seleccionados se suma 3000 y los documentos con esos números formarán la muestra.

Seleccionar con Excel® el procedimiento es mucho más sencillo, pues ese software internamente realiza las operaciones anteriores.

2.- Muestreo Sistemático

Suponga que tiene una población de N documentos numerados correlativamente y se desea seleccionar n documentos (n<N). Se calcula previamente el factor de elevación N/n y se determina de esta forma el intervalo inicial $\left[x_{(1)} + \frac{N}{n}\right]$, donde $x_{(1)}$ es el primer elemento del listado de documentos. En este intervalo se elige un número al azar (entero), que será el arranque aleatorio $x_{(a)}$. El primer elemento de la muestra es

este arranque, el segundo elemento de la muestra es $x_{(a)} + \frac{N}{n}$, el tercer elemento de la muestra es $x_{(a)} + 2 * \frac{N}{n}$ y así sucesivamente hasta completar las n unidades de la muestra.

Ejemplo 1. Suponga que N=1000 y n=100. El factor de elevación es 10, entonces el primer intervalo comprende los números entre 1 y 10 inclusive.

Forma1. Seleccionar un número aleatorio entre 1 y 10 y a ese número seleccionado sumar el factor de elevación hasta encontrar la muestra de tamaño 100.

Supongamos que el número elegido al azar entre 1 y 10 es el 7. Entonces el primer número seleccionado es el 7, el segundo 17, el tercero 27, y así sucesivamente hasta seleccionar los 100 números.

Forma 2. Seleccionar más de un número aleatorio de partida. El número mínimo es 5 números de partida. Como el factor de elevación es 10, elegimos 5 números entre 1 y 50 ($^{5N}/_{n}$). Suponga que los números elegidos, a partir de la tabla de números aleatorios, son los siguientes: 22, 8, 48, 12, 34.

Los primeros 5 números seleccionados son :	8,	12,	22,	34,	48
Los 5 siguientes (c/u +50):	58,	62,	72,	84,	98

Y así sucesivamente hasta completar los 100 números deseados. Este procedimiento se utiliza para corregir en parte el sesgo que arroja el muestreo sistemático.

3.- Selección Aleatoria Sistemática

Es una combinación de una selección aleatoria y una selección sistemática. Generalmente involucra más números aleatorios en la selección. En este caso se utilizan intervalos variables de la misma amplitud $^{N}/_{n}$. Para determinar los intervalos

variables, se eligen números aleatorios entre el primer y segundo intervalo. A partir de estos números se selecciona el resto de los números hasta completar la muestra.

Ejemplo 1. Suponga que N=1000, n=100, N/n=10.

Elegimos números aleatorios entre 1 y 20. Supongamos que los primeros números elegidos de la muestra son 8, 2, 14, 15, 3 y 8, entonces los siguientes números se determinan de la siguiente forma:

(8+2)=10, (10+14)=24, (24+15)=39, (39+3)=42, (42+8)=50. Luego se seleccionan nuevamente número al azar entre 1 y 20: 13, 5, 15, 4, 1 y 18. Con estos números se obtienen los siguientes seleccionados: (50+13)=63, (63+5)=68, (68+15)=83, (83+4)=87, (87+1)=88, (88+18)=106. El proceso continúa de la misma forma hasta enterar 100 elementos que conformarán la muestra. Se propone como ejercicio completar la lista de números.

4.- Muestreo Estratificado

Esta selección se utiliza cuando la población de N elementos es heterogénea y se realiza una partición, llamada estratos (grupos homogéneos), para luego seleccionar una muestra aleatoria simple de tamaño n_h en cada uno de los estratos, de la forma que $n = \sum_{h=1}^{k} n_h$.

5.- Muestreo por Conglomerados

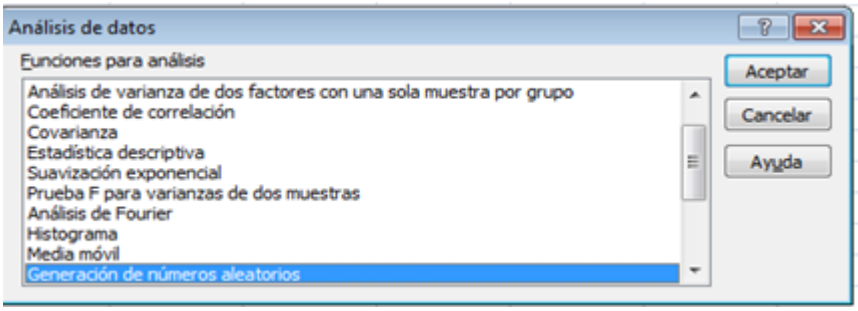
Cuando corresponde utilizar este método, en cada una de sus etapas se puede recurrir a las técnicas explicadas anteriormente. Esta forma de muestreo es adecuada aplicar cuando los grupos son homogéneos entre ellos, pero dentro de ellos sus elementos son heterogéneos.

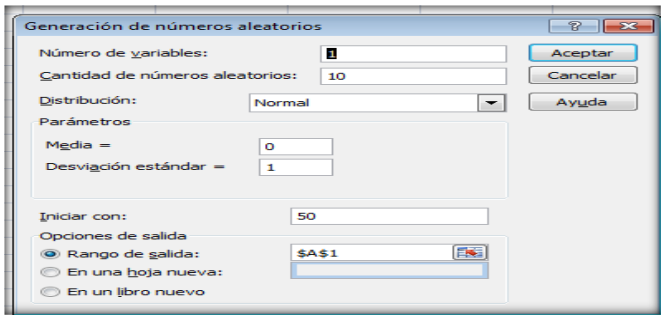
Ejemplo 1. Suponga que se desea investigar acerca de la proporción de vehículos que circulan por las carreteras de la región metropolitana con un único ocupante. Se requiere aplicar un método de muestreo que permita estimar la proporción de dichos vehículos.

Podemos utilizar conglomerados, los cuales pueden ser las comunas de la región. Luego dentro de cada comuna se puede hacer una estratificación dividiendo los conductores en frecuentes y poco frecuentes. Se puede dar una ponderación a cada estrato definido y luego tomar una muestra aleatoria simple en cada estrato registrando el número de pasajeros por vehículo.

6.- Selección Aleatoria con Excel®

Excel® permite obtener números aleatorios independientes de una distribución de probabilidad: Normal, Uniforme, Bernoulli, Binomial, Poisson. Para obtener los números se recurre al cuadro de diálogo **Análisis de Datos** y luego se selecciona Generación de números aleatorios y se obtiene el siguiente cuadro de diálogo:





Distribución Normal: Requiere que se especifique los parámetros de la distribución, media y desviación estándar. Por defecto utiliza media cero y desviación estándar 1.

Distribución Uniforme: Requiere que se especifique el rango de la variable y las variables se extraen con la misma probabilidad dentro del rango. Por defecto utiliza rango 0 a 1.

Distribución Bernoulli: Requiere que se especifique la probabilidad de éxito.
Distribución Binomial: Requiere que se especifique la probabilidad de éxito y el número de pruebas independientes de la distribución.

Distribución de Poisson: requiere se especifique el valor del parámetro de la distribución (media) y se ingresa el valor inverso (1/media).

Discreta: Requiere de dos columnas, la primera con los valores y la segunda con las probabilidades de ocurrencia.

7.- Muestras

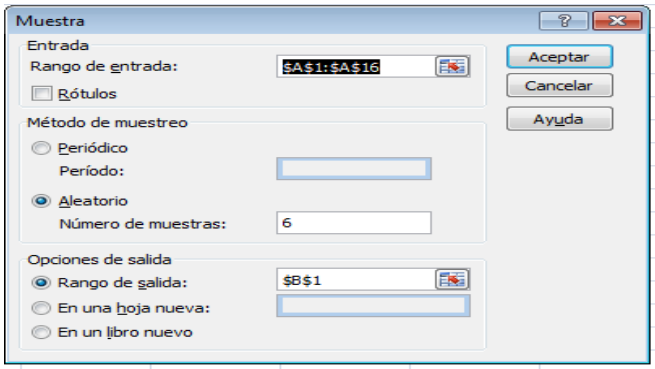
Para obtener muestras con Excel®, se tiene que introducir el rango de datos de la población de valores de los cuales se extraerá la muestra. Si hay más de una columna, Excel® extraerá una muestra de la primera columna, luego de la segunda y así sucesivamente.

Rótulos; se debe activar si las columnas del rango de entrada contienen rótulos. En caso contrario no activar, Excel® se encargará de poner los rótulos.

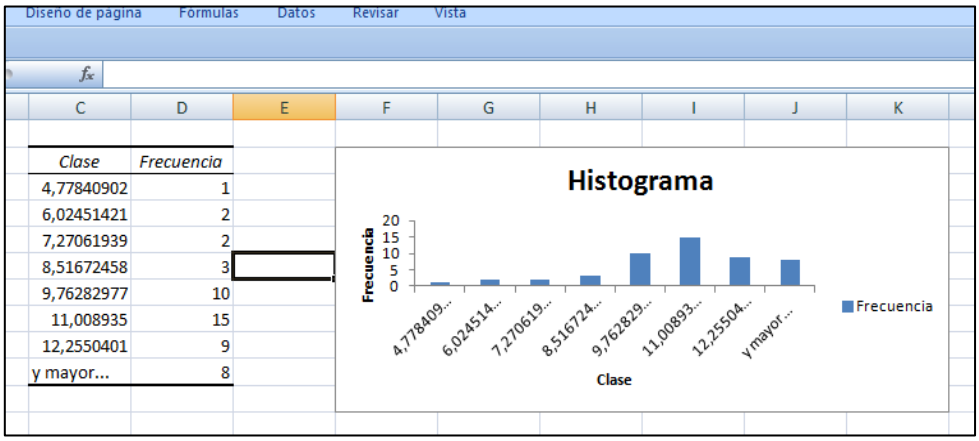
Método de muestreo; puede ser periódico o aleatorio para indicar el intervalo de muestreo que desee.

Periodo; se debe indicar el periodo en el que desee realizar la muestra. El número de valores seleccionados será igual al número de valores del rango de entrada dividido por la tasa de muestreo.

Número de muestras; Excel® selecciona la muestra con reposición. Si selecciona "Aleatorio", el número de valores seleccionados será igual al tamaño de muestra indicado.



Para representar el histograma de frecuencias de cada muestra, en **Análisis de Datos** se selecciona la opción **Histograma**.

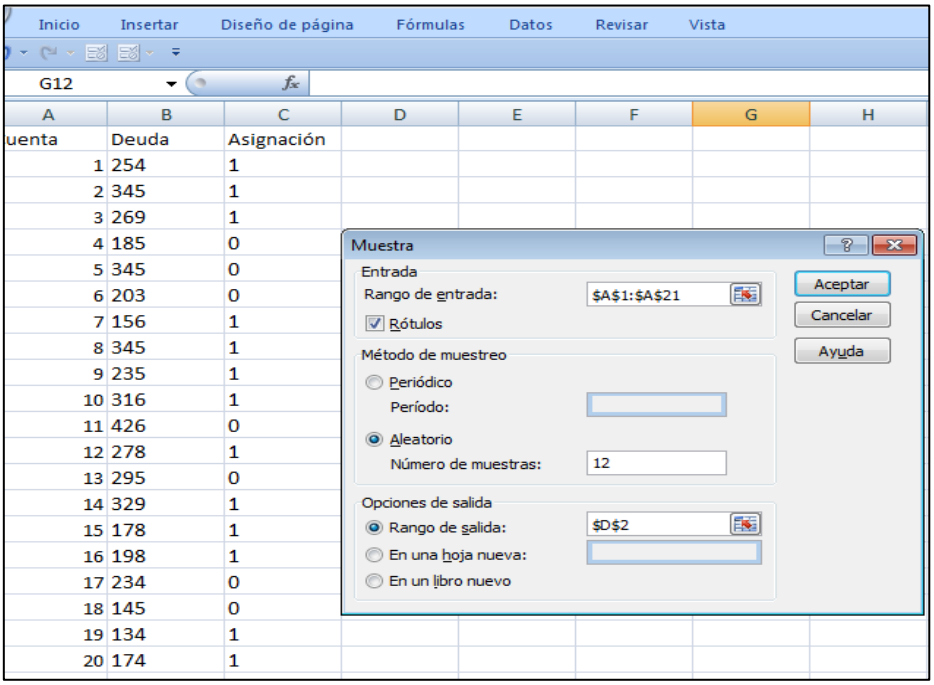


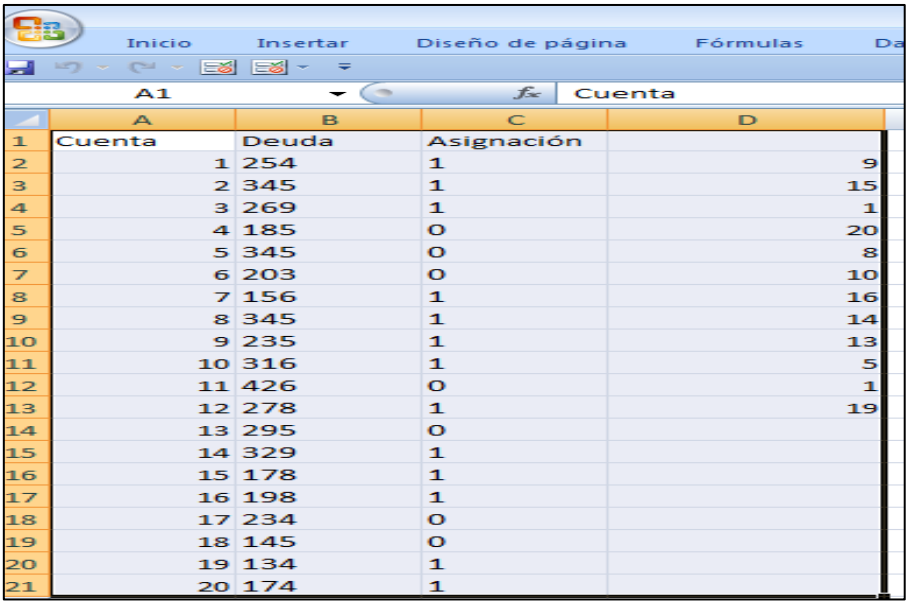
Para seleccionar muestras con rótulos, se procede de la siguiente manera, para ello se tiene la siguiente situación hipotética: Se tienen 20 cuentas impagas de una empresa, con la cantidad adeudada (miles de pesos). Se ha asignado 1 si los documentos cumplen con los procedimientos y 0 si no cumplen. Los registros son los siguientes:

Cta.	Deuda	Asig.	Cta.	Deuda	Asig.	Cta.	Deuda	Asig.
1	254	1	8	345	1	15	178	1
2	345	1	9	235	1	16	198	1
3	269	1	10	316	1	17	234	0

4	185	0	11	426	0	18	145	0
5	345	0	12	278	1	19	134	1
6	203	0	13	295	0	20	174	1
7	156	1	14	329	1			

De esta población de cuentas impagas, tomamos una muestra de 12 cuentas utilizando Excel®:





	A	B	C	D
1	Cuenta	Deuda	Asignación	
2	1	254	1	9
3	2	345	1	15
4	3	269	1	1
5	4	185	0	20
6	5	345	0	8
7	6	203	0	10
8	7	156	1	16
9	8	345	1	14
10	9	235	1	13
11	10	316	1	5
12	11	426	0	1
13	12	278	1	19
14	13	295	0	
15	14	329	1	
16	15	178	1	
17	16	198	1	
18	17	234	0	
19	18	145	0	
20	19	134	1	
21	20	174	1	

VIII.- ELEMENTOS DE MUESTREO ESTADÍSTICO

La adquisición de la información que se necesita para realizar un trabajo determinado, es muchas veces ingrato, difícil y costoso. La recogida de datos es el evento más importante en todo estudio estadístico, porque la metodología estadística de análisis de datos no es apropiado cuando se aplica a datos falsos o tomados en malas condiciones. Las técnicas de muestreo tienen por objeto suministrar la metodología que guíe los problemas de recogida de la información.

En la práctica, los problemas que se presentan, nunca son exactamente iguales a los modelos estadísticos conocidos, modelos que son indispensables para guiar las elecciones y establecer las condiciones bajo las cuales unas estrategias son preferibles a otras y también para evitar errores no siempre evidentes.

Diseñar una muestra tiene dos aspectos fundamentales: La forma de selección y el número de elementos que se seleccionarán del universo en estudio. También incluye un proceso de estimación para calcular ciertas características de la muestra.

1.- Estrategias para Recopilar Datos de Muestra

La información que interesa obtener de un universo (población) particular, suele servir para estimar el número total de unidades, el valor medio de ciertas características, la proporción de unidades que poseen cierta característica o atributo, entre otras.

a) Examinar todas las unidades de la población, es decir, realizar un censo.
b) Examinar, según una planificación previamente establecida, determinadas unidades de la población (muestra), y suponer que los resultados obtenidos son representativos de la población. Esta estrategia asume un riesgo (riesgo de muestreo) frente a la seguridad que supone el censo.

En muchas ocasiones la mejor solución es tomar una muestra, debido a que:

a) La población es tan grande que excede las posibilidades de la investigación.
b) La población es totalmente homogénea.
c) El proceso de medida o investigación es destructiva (prueba de calidad).
d) Se reducen los costos, se tiene mayor rapidez y/o exactitud en la respuesta.

La decisión de tomar una muestra en lugar de un censo consiste en minimizar la pérdida total en la que se incluyen, los recursos empleados (tiempo, dinero, esfuerzos y recursos) y por otro lado, cuantificar el error y probabilidad de cometerlo.
Para que un muestreo sea óptimo, se requiere de una teoría amplia y delicada y de una preparación detallada. La preparación debe estar presente tanto en los diseñadores como en los entrevistadores, inspectores y supervisores.

Muchas veces las unidades estudiadas son de tamaños desiguales y si la característica en estudio depende del tamaño de las unidades, se puede realizar un censo para las unidades grandes y una muestra entre las unidades de tamaño pequeño. De esta forma se ahorra tiempo, dinero y esfuerzos.

2.- Conceptos Básicos

Denominamos **POBLACIÓN** a un conjunto de unidades del que se desea obtener cierta información. Las unidades pueden ser personas, viviendas, cuentas, etc., y la información puede ser el consumo medio por familia, superficie de la vivienda, saldos de cuentas, etc. En la población es posible contar o medir una o varias características de las unidades en estudio. Con estos resultados será posible llegar a conocer valores como la media, el total de la clase y la proporción, los cuales se denominan parámetros o características de la población.

La población objetivo puede considerarse como un modelo cuya contrapartida es el mundo real. Esta población considera información en la cual existen omisiones,

duplicaciones y unidades extrañas que muchas veces no es fácil obtenerla por diferentes motivos, como la inaccesibilidad, la negativa a colaborar o las ausencias. Todos estos inconvenientes hacen que el conjunto que realmente investigamos, que llamamos población investigada, difiera de la población objetivo.

Llamaremos **MUESTRA**, al conjunto de unidades de la población de las cuales se obtiene información, si no se puede obtener de todas las unidades que la componen. Para elegir las unidades de muestra, es necesario disponer de un conjunto real de unidades que se ajusten idealmente al conjunto de datos de la población objetivo.

Llamaremos **MARCO**, al conjunto de unidades a partir de las cuales se toma la muestra. Determinar este conjunto de unidades, es una de las operaciones más delicadas del muestreo.

Llamaremos **VALORES VERDADEROS** a los valores que toman las variables que deseamos estudiar. En la práctica, debido a imperfecciones o limitaciones en la forma de obtener estos valores y a los errores que se cometen al elegirlos como en las operaciones que se ven sometidos posteriormente, consideramos valores que no coinciden con los verdaderos, a los cuales llamaremos **VALORES OBSERVADOS**.

Cuando elegimos la muestra, los datos obtenidos nos permiten hacer inferencia sobre valores aproximados de la población. A estos valores aproximados se les llama **ESTIMACIONES**, las cuales, debido a muestreo, están afectadas por un error que llamaremos **ERROR DEBIDO A MUESTREO**. Cuanto menor sea este error, diremos que mayor es la **PRECISIÓN** de las estimaciones.

3.- Estimaciones

El primer problema que se presenta es saber qué estimadores elegir de forma que la precisión de estos estimadores sea la mayor posible.

Supongamos que la población depende del parámetro θ, entonces su estimador será $\hat{\theta}$. Una propiedad deseable para el estimador, es que sea un insesgado del parámetro desconocido, es decir, $E(\hat{\theta}) = \theta$. El hecho que el estimador del parámetro poblacional sea insesgado, no significa que sea "bueno" ni siquiera razonable.

Sin embargo, si el estimador es insesgado y tiene una varianza pequeña, resulta que la distribución del estimador estará centrada alrededor de su media y existirá una alta probabilidad de que el estimador esté cercano al parámetro desconocido θ. Por tanto es necesario medir la "bondad" o precisión del estimador.

Mediremos la **PRECISION** del estimador por su varianza, siempre que sea insesgado. Para tal efecto recurrimos al error cuadrático medio (ECM):

$$ECM(\hat{\theta}) = E\left(\hat{\theta} - \theta\right)^2 = V(\hat{\theta}) \Leftrightarrow \hat{\theta} \text{ es insesgado}$$

Se llama **ERROR DE MUESTREO**, a la raíz cuadrada el error cuadrático medio y se llama **ERROR RELATIVO DE MUESTREO**, al cociente entre el error cuadrático medio y la esperanza del estimador:

$$EMR = \frac{\sqrt{ECM(\hat{\theta})}}{E(\hat{\theta})} = \frac{\sqrt{V(\hat{\theta})}}{\theta}$$

En muchas ocasiones, aparte de la estimación puntual del parámetro, es complementario entregar un intervalo de confianza donde se encontraría el parámetro desconocido de la población. Existen dos formas de presentar intervalos de confianza:

a) Partiendo de la desigualdad de Chebyschef, se tiene el siguiente intervalo:

$$\left(\hat{\theta} - k\sqrt{ECM(\hat{\theta})}\,;\, \hat{\theta} + k\sqrt{ECM(\hat{\theta})}\right)$$

Que tiene una confianza de $\left(1 - \frac{1}{k^2}\right)100\%$, es decir, el intervalo dado cubre al parámetro desconocido con probabilidad $\left(1 - \frac{1}{k^2}\right)$.

b) Si el tamaño de la muestra es suficientemente grande para que la distribución en el muestreo del estimador, del parámetro de la población, sea la distribución normal, en virtud del Teorema Central del Límite, el intervalo de (1-α)100%, de confianza para el parámetro desconocido es:

$$\left(\hat{\theta} - z_{\frac{\alpha}{2}} \sqrt{ECM(\hat{\theta})} \ ; \ \hat{\theta} + z_{\frac{\alpha}{2}} \sqrt{ECM(\hat{\theta})} \right)$$

donde $z_{\alpha/2}$ es coeficiente de confianza o valor crítico de una distribución normal estándar, que deja a la derecha el área $\alpha/2$.

4.- Error de Muestreo

Pareciera que los errores de muestreo son una limitante para la utilización del muestreo. Esta impresión es totalmente injustificada, ya que el error de muestreo es solo una parte del error total, y se disponen de métodos para controlarlo y medirlo.

Los métodos estadísticos permiten establecer a priori el error de muestreo que estamos dispuestos a aceptar o tolerar y así, elegir la muestra de forma tal que podamos afirmar con toda certeza, que las conclusiones que obtengamos de la población, basada en la muestra, no estarán afectadas por error superiores a los límites de error previamente establecidos.

Existen otros errores ajenos al muestreo, que se llaman ERRORES DE NO MUESTREO, que se dividen fundamentalmente de la siguiente forma:

$$\textit{Errores no muestrales} \begin{cases} \text{De observación} \begin{cases} \text{Errores de sobrecobertura} \\ \text{Errores de medida} \\ \text{Errores de procedimientos} \end{cases} \\ \text{De no observación} \begin{cases} \text{Errores de cobertura} \\ \text{Falta de respuesta} \end{cases} \end{cases}$$

Los errores de observación se deben a la recogida, registro o procesamiento incorrecto de los datos.

Los errores de no observación se deben a que no es posible obtener información deseada de ciertos elementos de la población.

Los errores de sobrecobertura, ocurren cuando el marco contiene elementos que no pertenecen a la población en estudio.

Los errores de medida, son la diferencia entre el valor final observado y el valor original o verdadero.

Errores de procesamiento, son aquellos que se cometen en procesos de entrada, edición, tabulación y análisis de datos.

Los errores de cobertura, ocurren cuando hay cierta parte de la población investigada que no está presente en el marco y por tanto no puede ser seleccionada.

La falta de respuesta, ocurre cuando para ciertos elementos de la muestra no puede obtenerse toda o parte de la información deseada o bien la información obtenida no puede ser utilizada.

5.- Notación Usual

Usaremos N para indicar el tamaño de la población, cuando la población sea finita, y utilizaremos n, para indicar el número de unidades de la muestra.

La simbología para las unidades de la población y muestra serán las siguientes:
Cada unidad de la población la denotaremos $u_i, i = 1,2,3,\ldots,N$. A cada unidad de la población le haremos corresponder una variable cuantitativa x_i que es el resultado de medir una característica de la unidad de la población. Cuando la variable es de

atributos, le haremos corresponder una variable cualitativa A_i que tomará valores los 0 (cero) o 1 (uno) de acuerdo a si no tiene o tiene la cualidad o atributo.

Llamaremos espacio muestral S, al conjunto de todas las muestras posibles extraídas con un procedimiento de muestreo dado o conocido: $s = \{x_1, x_2, x_3, ..., x_n\}, s \in S.$

6.- Muestreo Probabilístico

Se dice que un muestreo es probabilístico, cuando se puede calcular de antemano cuál es la probabilidad de obtener cada una de las muestras que es posible seleccionar. En este caso se define la función de probabilidad: $\sum_{s \in S} P(s) = 1$.

7.- Muestreo Intencional

En este caso es la persona que selecciona la muestra la que procura que dicha muestra sea "representativa". Aquí la representatividad depende de la intención u opinión que tenga la persona, haciendo prevalecer sus preferencias o tendencias individuales.

8.- Muestreo sin Norma

En este caso, el individuo toma la muestra de cualquier manera, por comodidad o capricho, y de esta forma obtiene las unidades de la población. Esta muestra puede ser representativa, solo en el caso de que la población es homogénea.

El muestreo intencional y sin norma, pueden dar buenos resultados, bajo determinadas condiciones, pero ambos carecen de una base teórica satisfactoria que les ampare.

El muestro probabilístico es el único que tiene respaldo teórico y por tanto puede predecir la validez de las estimaciones muestrales.

9.- Métodos de muestreo probabilísticos más usuales

Este estudio está restringido a tres tipos de parámetros: El total poblacional, la media de la población y proporción de la población:

$$X = \sum_{i=1}^{N} x_i, \qquad \overline{X} = \frac{\sum_{i=1}^{N} x_i}{N}, \quad P = \frac{\sum_{i=1}^{N} A_i}{N}$$

10.- Muestreo con Probabilidades Iguales

El muestreo con probabilidades iguales es el método base de los muestreos probabilísticos y sirve como base cuando se intenta evaluar la calidad de un método cualquiera de muestreo.

La hipótesis de partida que tiene este muestreo, consiste en suponer la existencia de un marco perfecto, es decir, que no tenga omisiones ni duplicaciones. Dos tipos de métodos se utilizan regularmente:

- Muestreo con probabilidades iguales sin reemplazo.
- Muestreo con probabilidades iguales con reemplazo.

11.- Planteamiento del Método de Muestreo

a.- Muestreo Aleatorio Simple

En el muestreo de poblaciones finitas, cuando la muestra se obtiene unidad a unidad, sin reponer, se dice que el muestreo es aleatorio simple o irrestrictamente aleatorio. En este caso:

- Sea una población de N unidades de las cuales tomaremos una muestra de tamaño n.
- Tomamos sucesiva e independientemente las unidades con probabilidades iguales.
- Las muestras constan de las mismas unidades que se consideran idénticas.

Bajo estas condiciones, la probabilidad de seleccionar cualquier muestra $(u_1, u_2, ..., u_n)$ es igual a: $\binom{N}{n}^{-1}$ y la probabilidad de que cualquier u_i pertenezca a la muestra es $\frac{n}{N}$.

b.- Muestreo Con Reemplazo

Cuando la muestra se obtiene unidad a unidad con reposición de estas unidades a la población, después de cada selección, estamos ante un muestreo con reposición y probabilidades iguales. En este caso:

- Sea una población de N unidades de las cuales tomaremos una muestra de tamaño n.
- Seleccionamos sucesiva e independientemente las unidades con igual probabilidad de extracción: 1/N.
- Cada unidad seleccionada se repone a la población.

Bajo estas condiciones la probabilidad de seleccionar cualquier muestra $(u_1, u_2, ..., u_n)$ es igual a:

N^{-n} y la probabilidad de que cualquier u_i pertenezca a la muestra es $1 - \left(1 - \frac{1}{N}\right)^n$.

c.- Estimadores y sus Varianzas

c.1.- Muestreo Sin Reemplazo:

Tabla 1:

	Total	Media	Proporción
Estimadores	$\hat{X} = N\bar{x}$	$\hat{\bar{X}} = \bar{x}$	$\hat{P} = \sum_{i=1}^{n} \frac{A_i}{n}$
Varianza	$V(\hat{X}) = N^2(1-f)\frac{S^2}{n}$	$V(\hat{\bar{X}}) = (1-f)\frac{S^2}{n}$	$V(\hat{P}) = \frac{(N-n)PQ}{(N-1)n}$

Donde,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad f = \frac{n}{N}, \quad S^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{X})^2}{N-1}, \quad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}, \quad Q = 1 - P$$

Aquí, f se llama "fracción de muestreo" y S^2 e s^2 la cuasi-varianza poblacional y muestral, respectivamente. Por lo tanto los estimadores de las varianzas son:

Tabla 2:

	Total	Media	Proporción
Varianza estimada	$\hat{V}(\hat{X}) = N^2(1-f)\frac{s^2}{n}$	$\hat{V}(\hat{\bar{X}}) = (1-f)\frac{s^2}{n}$	$\hat{V}(\hat{P}) = \frac{(1-f)\hat{P}\hat{Q}}{(n-1)}$

c.2.- Muestreo con Reemplazo:

Tabla 3:

	Total	Media	Proporción
Estimadores	$\hat{X} = N\bar{x}$	$\hat{\bar{X}} = \bar{x}$	$\hat{P} = \sum_{i=1}^{n} \frac{A_i}{n}$
Varianza	$V(\hat{X}) = N^2 \frac{S^2(N-1)}{Nn}$	$V(\hat{\bar{X}}) = (N-1)\frac{S^2}{Nn}$	$V(\hat{P}) = \frac{PQ}{n}$

Las Varianzas Estimadas son:

Tabla 4:

	Total	Media	Proporción
Varianza estimada	$\hat{V}(\hat{X}) = N^2 \frac{s^2}{n}$	$\hat{V}(\hat{\bar{X}}) = \frac{s^2}{n}$	$\hat{V}(\hat{P}) = \frac{\hat{P}\hat{Q}}{(n-1)}$

Los intervalos de confianza para los parámetros de la población: Total, media y proporción, son los siguientes:

Tabla 5:

Muestreo sin Reemplazo		Muestreo con Reemplazo	
Parámetro	**Intervalo de confianza**	**Parámetro**	**Intervalo de confianza**
Total	$\left[N\bar{x} \pm k\frac{Ns}{\sqrt{n}}\sqrt{1-f}\right]$	Total	$\left[N\bar{x} \pm k\frac{Ns}{\sqrt{n}}\right]$
Media	$\left[\bar{x} \pm k\frac{s}{\sqrt{n}}\sqrt{1-f}\right]$	Media	$\left[\bar{x} \pm k\frac{s}{\sqrt{n}}\right]$
Proporción	$\left[\hat{P} \pm k\sqrt{\frac{\hat{P}\hat{Q}}{n-1}(1-f)}\right]$	Proporción	$\left[\hat{P} \pm k\sqrt{\frac{\hat{P}\hat{Q}}{n-1}}\right]$

d.- Determinación del Tamaño de la Muestra

El procedimiento a seguir, en el caso más sencillo de muestreo con probabilidades iguales, distinguiendo en cada caso el parámetro que desea estimar, se basa en el máximo error admisible entre parámetro y parámetro estimado, en valor absoluto, y un coeficiente de confianza, k, que fue aludido en intervalos de confianza, bajo supuesto de normalidad o según Chebyschef.

Los tamaños de muestra cuando se quiere estimar el total, la media o la proporción de la población, para muestreo con y sin reemplazo, cuando el tamaño de la población tiende a infinito, se resumen a continuación en la tabla N° 6:

Tabla 6:

Parámetros	Estimador	Tamaño Inicial	Tamaño Muestra s/r	Tamaño Muestra c/r
Total	$N\bar{x}$	$n_0 = \frac{N^2k^2S^2}{e^2}$	$n = \frac{n_0}{1+\frac{n_0}{N}}$	$n = n_0$
	$N\hat{P}$	$n_0 = \frac{k^2PQN^2}{e^2}$	id	$n = n_0$

Media	$\bar{x}$	$n_0 = \frac{k^2 S^2}{e^2}$	$n = \frac{n_0}{1+\frac{n_0}{N}}$	$n = n_0$
Proporción	$\hat{P}$	$n_0 = \frac{k^2 PQ}{e^2}$	$n = \frac{n_0}{1+\frac{n_0}{N}}$	$n = n_0$

Los valores de e y k se fijan con arreglo a nuestro objetivos y limitaciones, N se supone conocido. Con respecto a S^2 y P son valores desconocidos y para determinarlos se puede recurrir a estudios anteriores realizado sobre la misma población o bien tomando una muestra piloto previamente seleccionada. Para el caso del parámetro P, se sabe que el máximo valor que puede tomar, de forma tal que minimice la varianza de la población, es 0,5.

Para S^2, también es usual utilizar aproximación de la cuasi-desviación típica (S), de la variable, en función del rango de ésta, pues es fácil determinar los valores máximos y mínimos de variable determinada: $S = \frac{rango}{4}$

Los tamaños de muestras, cuando solo se fija el error de muestreo, son las siguientes:

Tabla 7:

Parámetros	Estimador	Tamaño de muestra s/r	Tamaño de muestra c/r
Total	$N\bar{x}$	$n = \frac{N^2 S^2}{e^2 + NS^2}$	$n = \frac{N^2 \sigma^2}{e^2}$
	$\hat{A}$	$n = \frac{N^3 PQ}{e^2(N-1) + N^2 PQ}$	$n = \frac{N^2 PQ}{e^2}$
Media	$\bar{x}$	$n = \frac{NS^2}{Ne^2 + S^2}$	$n = \frac{\sigma^2}{e^2}$
Proporción	$\hat{P}$	$n = \frac{NPQ}{e^2(N-1) + PQ}$	$n = \frac{PQ}{e^2}$

Cuando se fija el **Error Relativo de Muestreo**, E= $EMR = \frac{\sqrt{ECM(\hat{\theta})}}{E(\hat{\theta})}$, los tamaños de muestras para los parámetros de la población son los siguientes:

Tabla 8:

Parámetros	Estimador	Tamaño de muestra s/r	Tamaño de muestra c/r
Total	$N\bar{x}$ N $\hat{P}$	$n = \frac{(1-f)}{E^2}\left(\frac{S}{\bar{X}}\right)^2$ $n = \frac{NQ}{PE^2(N-1)+Q}$	$n = \frac{1}{E^2}\left(\frac{\sigma}{\bar{X}}\right)^2$ $n = \frac{Q}{PE^2}$
Media	$\bar{x}$	$n = \frac{(1-f)}{E^2}\left(\frac{S}{\bar{X}}\right)^2$	$n = \frac{1}{E^2}\left(\frac{\sigma}{\bar{X}}\right)^2$
Proporción	$\hat{P}$	$n = \frac{NQ}{PE^2(N-1)+Q}$	$n = \frac{Q}{PE^2}$

Tamaños de muestra para un Error de Muestreo y Coeficiente de Confianza:

Tabla 9:

Parámetros	Estimador	Tamaño de muestra s/r	Tamaño de muestra c/r
Total	$N\bar{x}$ N $\hat{P}$	$n = \frac{k^2S^2N^2}{e^2+k^2S^2N}$ $n = \frac{k^2N^2PQ}{e^2(N-1)+k^2N^2PQ}$	$n = \frac{k^2N^2\sigma^2}{e^2}$ $n = \frac{k^2N^2PQ}{e^2}$
Media	$\bar{x}$	$n = \frac{k^2NS^2}{Ne^2+k^2S^2}$	$n = \frac{k^2\sigma^2}{e^2}$
Proporción	$\hat{P}$	$n = \frac{k^2NPQ}{e^2(N-1)+k^2PQ}$	$n = \frac{k^2PQ}{e^2}$

Tamaños de muestra para un Error Relativo y Coeficiente de Confianza:

Tabla 10:

Parámetros	Estimador	Tamaño de muestra s/r	Tamaño de muestra c/r
Total $cv = \frac{S}{\overline{X}}$	$N\overline{x}$ $\hat{A}$	$n = \frac{k^2 N (cv)^2}{NE^2 + k^2 (cv)^2}$ $n = \frac{k^2 NQ}{E^2 P(N-1) + k^2 Q}$	$n = \frac{k^2}{E^2}\left(\frac{\sigma}{\overline{X}}\right)^2$ $n = \frac{k^2 Q}{E^2 P}$
Media	$\overline{x}$	$n = \frac{Nk^2 (cv)^2}{NE^2 + k^2 (cv)^2}$	$n = \frac{k^2}{E^2}\left(\frac{\sigma}{\overline{X}}\right)^2$
Proporción	$\hat{p}$	$n = \frac{k^2 NQ}{E^2 P(N-1) + k^2 Q}$	$n = \frac{k^2 Q}{PE^2}$

12.- Comparaciones entre muestreo aleatorio simple con y sin reposición.

Total de la Población:

$$\frac{V_{s/r}(\hat{X})}{V_{c/r}(\hat{X})} = \frac{\frac{N^2 (N-n)\sigma^2}{n(N-1)}}{N^2 \frac{\sigma^2}{n}} = \frac{N-n}{N-1} < 1 \Rightarrow V_{s/r}(\hat{X}) < V_{c/r}(\hat{X})$$

Media y Proporción de la población

$$\frac{V_{s/r}(\hat{\overline{X}})}{V_{c/r}(\hat{\overline{X}})} = \frac{V_{s/r}(\hat{A})}{V_{c/r}(\hat{A})} = \frac{V_{s/r}(\hat{P})}{V_{c/r}(\hat{P})} = \frac{N-n}{N-1} \Rightarrow V_{s/r} < V_{c/r}$$

En las cuatro situaciones, la varianza de los estimadores con muestreo sin reposición es menor que las varianzas con muestreo con reposición, por tanto el muestreo sin reposición es más preciso que el muestreo con reposición.

Si se comparan los tamaños de muestras, se encontrará que los tamaños para muestreo sin reposición son menores que los tamaños de muestra para muestreo con

reposición, avalando nuevamente la precisión del muestreo sin reponer, pues necesita un tamaño inferior para cometer el mismo error haciéndolo más eficiente.

Ejemplos de Cálculo de Tamaño de Muestra

1.- Proporciones

Se desea estimar la proporción total de piezas correctas producidas por un proceso industrial que fabrica 6000 unidades. Una muestra piloto indica que 1/3 de las piezas son defectuosas.

a) Encontrar el tamaño de muestra para estimar la proporción de piezas correctas, con un error de muestreo de 0,1 y también con un error relativo de 0,2.

b) Hallar el tamaño de muestra necesario para que el error de muestreo sea de 600 unidades con una confianza de 99,7%, asumiendo MAS con reposición. Encontrar n, en las condiciones anteriores, con un error relativo de muestreo de 10%.

Solución

a) Se tiene N=6000, P=2/3, error de muestreo (e)=0,1. Como se fija solo el error 10%, aplicamos tabla 7:

$$n = \frac{NPQ}{e^2(N-1)+PQ} = \frac{6000*2/3*1/3}{0{,}1^2*5999+2/3*1/3} = 22{,}14 \approx 23\,\text{piezas}$$

Cuando tenemos el error relativo, utilizamos la tabla 8:

$$n = \frac{NQ}{PE^2(N-1)+Q} = \frac{6000*1/3}{2/3*0{,}2^2*5999+1/3} = 12{,}47 \approx 13\,\text{piezas}$$

b) Se tiene e=600 piezas, α=0,003; $z_{\alpha/2}$=2,997. Utilizamos tabla 6:

$$n = \frac{k^2 PQN^2}{e^2} = \frac{2{,}997^2 * 2/3 * 1/3 * 6000^2}{600^2} = 199{,}9 \approx 200 \text{ piezas}$$

Cuando tenemos error relativo y coeficiente de confianza, tabla 10:

$$n = \frac{k^2\ Q}{E^2 P} = \frac{2{,}997^2 * 1/3}{0{,}1^2 * 2/3} = 449{,}1 \approx 450\, piezas$$

2.- De una población de 33 millones de individuos se desea tomar una muestra para estimar la tasa de actividad con un error absoluto e=0,02 y una confianza de 95%. Del último censo poblacional se sabe que hay un 39% de activos. En las mismas condiciones cuál es el tamaño de muestra para cometer un error relativo de 5%.

Solución: Para e=0,02 y confianza 95%, se tiene el coeficiente z=1,96. De la tabla 9, se tiene:

$$n = \frac{k^2 NPQ}{e^2(N-1) + k^2 PQ} = \frac{1{,}96^2 * 33000000 * 0{,}39 * 0{,}61}{0{,}02^2 * 32999999 + 1{,}96^2 * 0{,}39 * 0{,}61} = 2284$$

Para un error relativo de 5%, utilizamos la tabla 10:

$$n = \frac{k^2 NQ}{E^2 P(N-1) + k^2 Q} = \frac{1{,}96^2 * 33000000 * 0{,}61}{0{,}05^2 * 0{,}39 * 32999999 + 1{,}96^2 * 0{,}61} = 2403$$

3.- Una muestra aleatoria de tamaño 600, tomada de una población de 15000 unidades, entrega los siguientes datos: $\sum_{i=1}^{600} X_i = 2946;\ \sum_{i=1}^{600} X_i^2 = 18694$

a) Encuentre intervalos de 95% de confianza para la media y el total poblacional, suponiendo normalidad para la distribución de los estimadores.

b) Tomando la muestra anterior como muestra piloto, encuentre el tamaño de muestra para un error absoluto de muestreo de 1000 unidades y para un error relativo de 15%.

Solución: a)

$$\hat{X} = N\bar{x} = 15000 * \frac{2946}{600} = 73650 \text{ unidades; } \hat{S}^2 = \frac{\left[\sum_{i=1}^{600} X_i^2 - n(\bar{X})^2\right]}{n-1} = 7{,}06$$

a) $\hat{V}(\hat{X}) = N^2(1-f)\frac{\hat{S}^2}{n} \Rightarrow \sqrt{\hat{V}(\hat{X})} = 1594{,}239; \ [73650 \pm 1{,}96 * 1594{,}239]$

$$\bar{x} = 4{,}91; \ \sqrt{\hat{V}(\bar{x})} = \sqrt{(1-f)\frac{\hat{S}^2}{n}} = 0{,}106282; \ \ [4{,}91 \pm 1{,}96 * 0{,}106282]$$

b) e=1000, de la tabla 7: $n = \frac{N^2 S^2}{e^2 + NS^2} = 1437$

para, E=0,15, de la tabla 8 se tiene: $n = \frac{(1-f)}{E^2}\left(\frac{S}{\bar{X}}\right)^2$ =13

IX.- GLOSARIO DE PRINCIPALES TÉRMINOS UTILIZADOS EN MUESTREO ESTADÍSTICO

Muestreo de Auditoría

Aplicación de un procedimiento de auditoría a menos del 100% de los elementos de una población con el objetivo de sacar conclusiones acerca de toda la población.

Riesgo del Muestreo

Riesgo de que la conclusión del auditor interno basada en pruebas sobre muestras tal vez sea diferente de la conclusión a la que se arribaría si el procedimiento de auditoría se hubiese aplicado a todos los elementos de la población.

Riesgo Residual

La proporción de riesgo inherente que queda remanente luego de que la dirección implementa sus respuestas a los riesgos (a veces se lo denomina riesgo neto).

Riesgo de Control

Posibilidad de que las actividades de control fallen y no reduzcan el riesgo controlable a un nivel aceptable.

Riesgo Controlable

La parte del riesgo inherente que la dirección puede reducir a través de las actividades de gestión y las operaciones cotidianas.

Riesgo No Derivado del Muestreo

Es el riesgo asociado a cuestiones como realizar procedimientos de auditoría inapropiados, aplicar deficientemente un procedimiento apropiado o malinterpretar los resultados del muestreo.

Muestreo de Atributos

Método de muestreo estadístico que permite arribar a una conclusión acerca de una población en términos de una tasa de ocurrencia.

Riesgo de Evaluar el Riesgo de Control Demasiado Bajo

Riesgo de que el auditor interno concluya erróneamente que la actividad de control especificada es más eficaz de lo que realmente es.

Tasa de Desviación Tolerable

Tasa máxima de desviaciones que el auditor interno está dispuesto a aceptar y, aun así, concluir que la actividad de control es aceptablemente eficaz.

Tasa Esperada de Desviación de Población

La mejor estimación que tiene el auditor interno de la tasa de desviación real en la población de elementos bajo análisis.

Selección de Muestra Aleatoria

Garantiza que cada artículo de la población definida tenga las mismas posibilidades de ser seleccionado.

Muestreo Indiscriminado

Técnica de selección no aleatoria que utilizan los auditores internos para seleccionar una muestra que se espera que sea representativa de la población.

Muestreo de Probabilidad Proporcional al Tamaño (PPS, en inglés)

Forma modificada de muestreo de atributos que utilizan los auditores internos para establecer una conclusión en montos en pesos, en lugar de tasas de ocurrencia.

Muestreo por Descubrimiento

Caso especial de muestreo por atributo que se utiliza para determinar una probabilidad especificada de encontrar al menos un ejemplo de una ocurrencia (atributo) en una población. También se conoce como muestreo exploratorio.

Muestreo Secuencial

Tipo de muestreo por atributos, que permite frenar el muestreo para un cierto número de sucesos observados. Las unidades de muestreo se examinan en grupos hasta que la evidencia acumulada es suficiente para lograr una precisión y fiabilidad definida. También se conoce como muestreo de stop-or-go.

Pruebas de Control

Pruebas que se utilizan para determinar la efectividad del diseño y operación de los controles.

Muestreo de Variable

Plan de estadística que se utiliza para proyectar un carácter cuantitativo, como por ejemplo una cantidad de dinero. Muestreo de variable incluye media por unidad no estratificada, media por unidad estratificada, y la estimación de la diferencia.

X.- BIBLIOGRAFÍA

- Auditoría, Alvin Arens, James Loebbecke. Pearson Prentice Hall.
- Auditoría Interna: Servicios de Aseguramiento y Consultoría. Kurt F. Reding. Fundación de Investigaciones del IIA.
- Audit Sampling. Dan Guy. John Wiley.
- Consejo para la Práctica 2320-3: Muestreo de Auditoría del Instituto de Auditores Internos (IIA).
- Estadística para Administradores. W. Mendenhall.
- Guía de Muestreo Estadístico en Auditoría Gubernamental. OFS Estado de Guanajuato. México.
- Introducción de la Estadística. Pedro Marín, apuntes de clases. Usach.
- Introducción a la Estadística para las Ciencias Sociales. Daniel Peña.
- Material técnico preparado para el Consejo de Auditoría Interna General de Gobierno, por don Pedro Marín Álvarez, Profesor de Estado en Matemática y Estadística, Licenciado en Matemática, Magister en Estadística, Diplomado en Investigación de Operaciones y Doctor en Estadística©.
- Muestreo Estadístico, César Pérez. Pearson Prentice Hall.
- Muestreo: Diseño y Análisis, Sharon L. Lohr. Thomson.

XI.- CASO PRÁCTICO

Este estudio, tiene relación con la integridad de la información contenida en las fichas clínicas y la obtención de los datos de forma directa de primera fuente mediante los registros del hospital.

Para esta investigación se resolvió definir la población como todas las fichas nuevas creadas durante el año 2004 en las distintas unidades del hospital que pueden crear este documento, a partir de esta población fueron estudiados los atributos de los datos que constituyen esta población, concluyendo que son al parecer homogéneos dentro de los grupos pero heterogéneos en conjunto, por eso la estratificación) homogéneos, situación que permitió estratificar la población considerando como estrato cada una de las unidades relacionadas con la generación de fichas, para lograr posteriormente una muestra más representativa, tabulando esta población de la siguiente manera:

CUADRO Nº 1

MES	ARCHIVO	CONVENIOS	HOSPITALIZACIÓN	POLICLÍNICO	TOTAL
ENERO	99	81	137	48	**365**
FEBRERO	81	56	117	48	**302**
MARZO	102	86	120	22	**330**
ABRIL	216	76	160	16	**468**
MAYO	179	77	140	27	**423**
JUNIO	208	60	139	56	**463**
JULIO	206	68	156	40	**470**
AGOSTO	217	79	146	32	**474**
SEPTIEMBRE	177	75	136	38	**426**
OCTUBRE	171	74	139	60	**444**
NOVIEMBRE	183	79	163	32	**457**
DICIEMBRE	136	63	116	47	**362**
TOTAL	**1975**	**874**	**1669**	**466**	**4984**
%	**39.63%**	**17.54%**	**33.49%**	**9.35%**	**100%**

Los datos del cuadro anterior presentan una población total de 4.984 fichas médicas generadas en el año 2004, tabulados por unidades y meses calculando sus porcentajes respecto a la población total, con los datos emanados se procedió al calculó del tamaño de la muestra.

MUESTRA DE LA POBLACIÓN:

Para determinar el tamaño de la muestra de la población se realizó por medio de una afijación proporcional a partir de los estratos antes señalados y que arrojó como resultados los siguientes valores:

CUADRO Nº 2				
MES	**ARCHIVO**	**CONVENIOS**	**HOSPITALIZACIÓN**	**POLICLÍNICO**
ENERO	5%	9%	8%	10%
FEBRERO	4%	6%	7%	10%
MARZO	5%	10%	7%	5%
ABRIL	11%	9%	10%	3%
MAYO	9%	9%	8%	6%
JUNIO	11%	7%	8%	12%
JULIO	10%	8%	9%	9%
AGOSTO	11%	9%	9%	7%
SEPTIEMBRE	9%	9%	8%	8%
OCTUBRE	9%	8%	8%	13%
NOVIEMBRE	9%	9%	10%	7%
DICIEMBRE	7%	7%	7%	10%
TOTAL	**100%**	**100%**	**100%**	**100%**

Los datos anteriores corresponden a una muestra total de 257 fichas médicas, indicando por unidades y meses el número de fichas a revisar en cada uno de los estratos y que en su totalidad constituyen la muestra, permitiendo realizar posteriormente por cada estrato una selección aleatoria simple.

SELECCIÓN DE MUESTRA ALEATORIA SIMPLE:

CUADRO Nº 3					
MES	**ARCHIVO**	**CONVENIOS**	**HOSPITALIZACIÓN**	**POLICLÍNICO**	**TOTAL**
ENERO	5	4	7	2	**18**
FEBRERO	4	3	6	3	**16**
MARZO	5	4	6	1	**16**
ABRIL	11	4	8	1	**24**
MAYO	9	4	7	1	**21**
JUNIO	11	3	7	3	**24**
JULIO	11	4	8	2	**25**
AGOSTO	11	4	8	2	**25**
SEPTIEMBRE	9	4	7	2	**22**
OCTUBRE	9	4	7	3	**23**
NOVIEMBRE	10	4	9	2	**25**
DICIEMBRE	7	3	6	2	**18**
TOTAL	**102**	**45**	**86**	**24**	**257**

Para la elección de los elementos que contemplan la muestra se consideró igual probabilidad de ser elegidos y componer el conjunto de la muestra, la cual se tomó sin reposición y con un nivel de confianza de 90%. Para la elección de estos elementos se utilizó la técnica de números aleatorios, mediante la utilización del programa computacional Excel®, tomando la base de datos contenida en una hoja de cálculo de este programa, contemplando el total de la población ya estratificada y los siguientes pasos:

1. De la Barra de menú de la hoja de cálculo, se desplegó el menú INSERTAR.
2. Posteriormente se pinchó el comando FUNCIÓN, en categoría de la función pinchando TODAS, luego se buscó en el nombre de la función ALEATORIO y se aceptó.
3. En la barra de fórmulas, se aplicaron las expresiones por estratos, debiendo posicionarse previamente en la celda donde se registrarían todos los números aleatorios elegidos por el programa, las fórmulas aplicadas fueron las siguientes:

=ALEATORIO()*B10:B1984*(1975-1)+1
=ALEATORIO()*B1985:B2858*(874-1)+1
=ALEATORIO()*B2859:B4527*(1669-1)+1
=ALEATORIO()*B4528:B4993*(466-1)+1

Con los números aleatorios obtenidos, se escogió las fichas médicas posicionada en el número elegido.

Para la revisión de las fichas médicas se consideró el Máximo error estándar para X_i="Ficha i-ésima", es decir, se propone que el 50% de las fichas médicas contienen error, entendiéndose que si los contiene, no cumple con la estructura propuesta.

$P[X_i = 1] = 0.5 = P$ (Éxito de encontrar una ficha con error de especificación).

$$X_i = \begin{cases} 0 & \text{si la ficha i cumple con estructura propuesta} \\ 1 & \text{si la ficha i no cumple con estructura propuesta} \end{cases}$$

RESUMEN DE DATOS ESTADÍSTICOS Y FÓRMULAS UTILIZADAS:

Población: Todas las fichas nuevas creadas durante al año 2004.

N: 4984 fichas

Período: Año 2004.

Unidad de Muestreo : Ficha clínica de los pacientes

Estratos: Unidad Archivos, Servicio Hospitalización, Unidad Convenios, Unidad Policlínico

Tamaño de Estratos: 102 fichas de unidad Archivos

45 fichas de servicio Hospitalización

86 fichas de unidad Convenios

24 fichas de unidad Policlínico

JUSTIFICACIÓN DE LOS RESULTADOS

Nivel de Confianza: 90%

Coeficiente de confianza k = 1,645

Máxima desviación estándar:

$$P = 0{,}5 \qquad Q = 1 - P = 0{,}5$$

$$SE = \sqrt{P(1-P)} = \sqrt{0.5 \times 0.5} = 0.5$$

Determinación del tamaño de muestra:

Error de Muestreo: e = $|p - \hat{p}| = 0{,}05$

e = 0.05, K =1,645, N =4984, P =0,5 =Q

$$n = \frac{z^2 \cdot N \cdot P \cdot Q}{e^2 \cdot (N-1) + k^2 \cdot P \cdot Q} = n = \frac{1{,}645^2 \cdot 4984 \cdot 0{,}5 \cdot 0{,}5}{0{,}05^2 \cdot 4983 + 1{,}645^2 \cdot 0{,}5 \cdot 0{,}5}$$

$$n = \frac{3372}{13{,}134} = 256{,}738 \approx 257$$