

INTRODUCCIÓN

A LA

ESTADÍSTICA

**Conceptos Básicos
Teórico y Práctico**

Escrito por Prof. A. Rodrigo Farinha

Publicado en [Octubre de 2010] en mi sitio

www.arfsoft.com.uy

Queda absolutamente prohibido el uso total o parcial de este material sin dar crédito a su autor. Solamente se puede imprimir y sin modificación alguna.

“Durante el siglo XX, el razonamiento y la metodología estadística se han convertido literalmente en el marco científico para docenas de otros campos incluyendo la educación, agricultura, economía, biología y medicina; y más aún, recientemente con una mayor influencia en las ciencias duras tales como la astronomía, geología y física. Es decir, la estadística ha crecido de un campo desconocido pequeño a un campo desconocido gigante.”

Prof. Bradley Efron

ÍNDICE

Conceptos generales	4
Estadística Descriptiva	6
Tabulación y representación gráfica de los datos	6
Tabla de frecuencia	7
Histograma	8
Polígono de frecuencia	8
Gráfico circular	9
Medición de datos	10
Medidas de Tendencia Central	10
Media o Promedio	10
Moda	10
Mediana	10
Medidas de Dispersión	11
Rango	11
Desviación Estándar o Típica	11
Varianza	11
Coeficiente de Variación	11
Ejercicios	13
Estadística Inferencial	14
Tamaño de una muestra	14
Representatividad de la muestra	14
¿Por qué se usan las muestras para hacer inferencia y no toda la población?	15
Tipos de muestreo	16
Tamaño de la muestra	16
Estudios para determinar parámetros	17
A. Estimar una proporción	17
B. Estimar una media	19
Estudios para contraste o prueba de hipótesis	20
A. Comparación de dos proporciones	20
B. Comparación de dos medias	22
El tamaño muestral ajustado a las pérdidas	23
Intervalo de confianza	24
Intervalo de confianza para una media μ	24
Intervalo de confianza para una proporción p	27
Intervalo de confianza para una desviación estándar σ	28
Fuentes consultadas	30

Conceptos generales

ESTADÍSTICA: Rama de las matemáticas que reúne, organiza, presenta, analiza e interpreta un conjunto de datos, con el objeto de extraer conclusiones de ellos o de tomar decisiones.

Para su estudio, la Estadística se divide en dos ramas de acuerdo con el objetivo que se persigue:

- **Estadística Descriptiva:** Tiene como herramientas la distribución de frecuencias, los gráficos, las medidas de tendencia central y de dispersión, y su objeto es informar acerca de las características que describen los datos. Así, esta rama reúne, organiza y presenta los datos.
- **Estadística Inferencial:** Sus herramientas son el muestreo y la teoría de probabilidad. Su objeto es inferir ó inducir conclusiones acerca de una población, basándose en una muestra de ella. Cómo se selecciona la muestra, cómo se realiza la inferencia, y qué grado de confianza se puede tener en ella son aspectos fundamentales de la estadística inferencial.

Población: conjunto de elementos (no necesariamente personas) de los cuales se estudia cierta característica.

Variable: el objeto de estudio en una población determinada.

Las variables pueden ser:

Cualitativas: indican una característica o cualidad (no son numéricas). Ejemplos: sexo, estado civil

Cuantitativas: se expresan mediante números. Se las puede subclasificar en:

Discretas: son representadas por números enteros o fraccionarios (nota, edad, número de hijos, cantidad de artículos defectuosos fabricados, etc)

Continuas: los datos son números dentro de un rango ó intervalo (duración de una reacción química, altura o peso de personas, porcentaje de mujeres casadas, etc)

Niveles de medición de las variables:

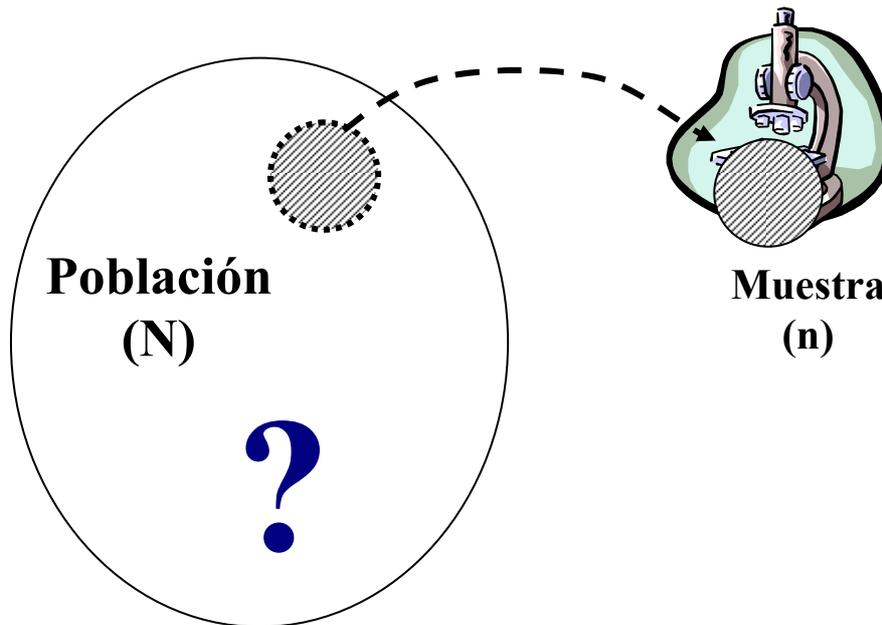
Nivel nominal. Son datos que representan grupos ó categorías, pero en las cuales no existe jerarquía. Por ejemplo, unos datos correspondientes al estado civil, el nombre ó la empresa donde trabaja.

Nivel ordinal. Igual que en el nivel nominal los datos pertenecen a grupos ó categorías, pero la jerarquía es relevante. Por ejemplo, datos correspondientes a la categoría en la cual está federado un deportista, nivel al cual se habla inglés ó categoría en la empresa donde se trabaja.

Nivel de intervalo. En este nivel, los datos representan rangos entre dos valores (uno inferior y uno superior). Por ejemplo, la temperatura en Caracas hoy se ubica entre un mínimo de 24°C y un máximo de 33°C, el número de cheques devueltos por un banco el día de hoy oscila entre 10 y 15, la estatura promedio de un grupo de personas fluctúa entre 1,70 m. y 1,75 m.

Nivel de razón. Aquí los datos son presentados como porcentajes, proporciones, tasas, etc. Por ejemplo, la tasa de mortalidad en un cierto país es de 6 por 1000, el porcentaje de aprobados en una materia es de 70% ó una de cada 3 personas fuman.

Muestra: subconjunto de elementos de la población, escogidos para su estudio debido a que no es posible o viable recurrir a todos los integrantes de la población. A partir de los datos extraídos de la muestra, se pretende estimar (conocer de manera aproximada) los parámetros de la población mediante cálculos realizados sobre la muestra. Para que la muestra sea **representativa de la población** a estudiar, debe ser **elegida al azar**. Un tamaño de muestra inadecuado conduce a un inevitable desperdicio y desaprovechamiento de recursos.



ESTADÍSTICA DESCRIPTIVA

Tabulación y representación gráfica de los datos

Tamaño de la muestra: n

Datos: x_i

Ejemplo: Las estaturas en cm de los alumnos de un grupo de 3º año de una escuela son:

138	144	130	146	128	145	133	129	143	136	137	138	129	133
139	145	128	138	140	146	142	148	132	130	143	135	134	136
138	131	141	139	133	130	139	135	138	147	137	135	133	132
137	138	140	142	131	139								

Entonces el tamaño de la muestra y los datos son:

$$n = 48$$

$$x_1 = 138$$

$$x_2 = 144$$

$$x_3 = 130$$

.....

.....

$$x_{47} = 131$$

$$x_{48} = 139$$

Como en este ejemplo hay datos repetidos, es posible facilitar su presentación e interpretación utilizando **tablas y gráficas**:

- Tabla de frecuencia
- Histograma
- Polígono de frecuencia
- Gráfico circular

Nota: Hay otras formas de representar los datos. En este curso se verán solamente estas 4 (con énfasis práctico en las 3 primeras).

Tabla de frecuencia

Consiste en tabular la información basándose en su *frecuencia* (cantidad de veces que cada dato aparece repetido).

La primera columna contiene los **datos** (en el ejemplo, serían las estaturas) ordenados de menor a mayor y sin repetirlos.

La segunda columna contiene la **frecuencia** de cada dato.

La suma total de los números de esta columna debe coincidir con el tamaño de la muestra. $\sum f_i = n$

Opcionalmente, para contestar preguntas del tipo ¿cuántos estudiantes miden menos de 140 cm?, se crea una tercer columna que contiene la **frecuencia acumulada** (es la suma de las frecuencias de todos los datos anteriores hasta el dato correspondiente inclusive).

El último número de esta columna debe coincidir con el tamaño de la muestra.

Estatura (cm)	Frecuencia f_i	Frecuencia acumulada
128	2	2
129	2	4
130	3	7
131	2	9
132	2	11
133	4	15
134	1	16
135	3	19
136	2	21
137	3	24
138	6	30
139	4	34
140	2	36
141	1	37
142	2	39
143	2	41
144	1	42
145	2	44
146	2	46
147	1	47
148	1	48
<i>TOTAL</i>	48	

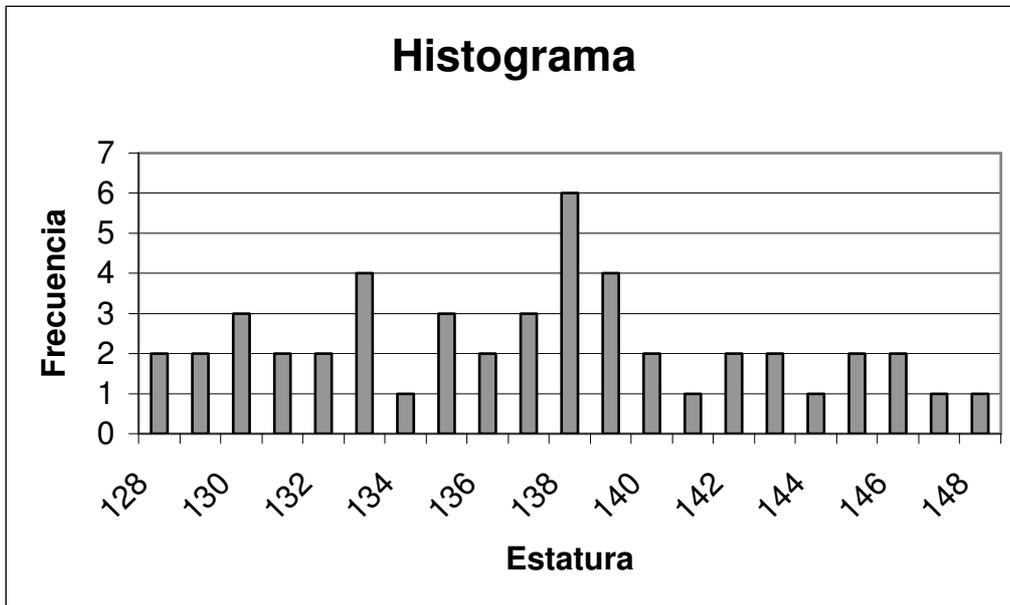
Cómo se lee la tabla:

Ejemplos:

¿Cuántos alumnos tienen 135 cm de estatura?	3
¿Cuántos alumnos tienen 143 cm de estatura?	2
¿Cuántos alumnos miden menos de 140 cm?	34
¿Cuántos alumnos miden hasta 132 cm?	11

Histograma

Se grafica en un sistema de ejes un *gráfico de barras* donde la altura de cada rectángulo es la frecuencia de cada dato.



Polígono de frecuencia

Se grafica en un sistema de ejes los puntos de coordenadas (dato, frecuencia) y luego *se unen mediante segmentos*.

La información aquí utilizada es la misma que la del histograma, pero está presentada de forma diferente.

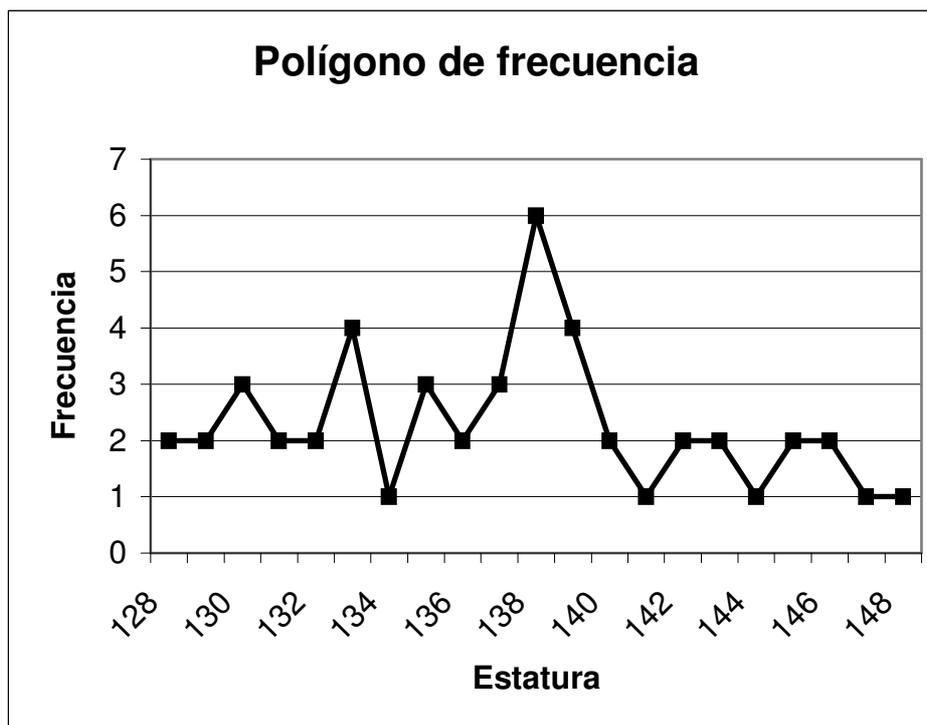
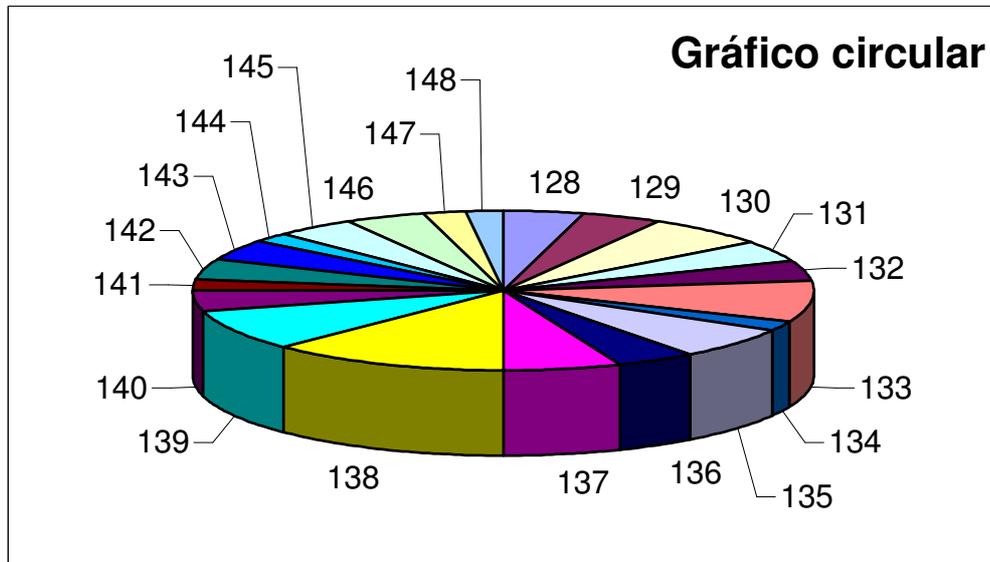


Gráfico circular

La muestra es representada por un círculo y cada dato por un sector proporcional de éste.

El ángulo α que tendrá el sector correspondiente al dato de frecuencia f será: $\alpha = \frac{360f}{n}$



Observando éste gráfico circular del ejemplo, ¿sería conveniente utilizarlo en un informe?

Posiblemente no. Obsérvese que, como en éste ejemplo hay muchas estaturas a considerar, el gráfico queda muy “saturado” y no aporta información visual realmente relevante.

Pero esto no significa que éste tipo de gráfico no sea útil y que jamás deba usarse. Dependerá de cada situación y de cómo se presente.

Medición de datos

Hay 2 tipos de medidas o parámetros: las de *Tendencia Central* y las de *Dispersión*.

Medidas de Tendencia Central: indican la tendencia central de la variable estudiada.

- **Media o Promedio:** cociente entre la suma de todos los datos y el número total de observaciones. La confianza en este valor (su validez) depende de la *desviación estándar*.

$$\text{Si los datos no se repiten: } \bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n}$$

$$\text{Si los datos se repiten: } \bar{x} = \frac{\sum (x_i \cdot f_i)}{n}$$

(f_i es la frecuencia de x_i)

Esta media es la *media muestral*.

La *media poblacional* se simboliza μ y su fórmula es:
(N es el tamaño de la población)

$$\mu = \frac{\sum_{i=1}^{i=N} x_i}{N}$$

- **Moda:** dato que se repite más veces (es decir, es el dato que tiene mayor frecuencia). Puede suceder que exista más de una moda (más de un valor que tienen la mayor frecuencia) o que no haya ninguna (todos los datos tienen la misma frecuencia).

Debe usarse con cuidado. Su objetivo es identificar zonas donde hay aglomeraciones de datos, por lo cual es confiable solo cuando la muestra es relativamente grande.

- **Mediana:** “centro” de los datos (tienen que estar ordenados en forma creciente o decreciente)

Si n es *impar*: es el dato central

Si n es *par*: es el promedio de los dos datos centrales

Una de sus ventajas es que no está influenciada por los valores extremos (datos muy alejados del resto).

Cuando la muestra consta de muchos datos y hay repeticiones, conviene hacer lo siguiente:

1. Hacer la tabla de frecuencia (incluyendo la frecuencia acumulada).

2. Dividir entre 2 la cantidad de datos observados. En el ejemplo: $\frac{48}{2} = 24$

3. Se busca en la tabla si está presente una frecuencia acumulada que tenga ese valor:

Si está presente: la *mediana* será el promedio del dato correspondiente a esa frecuencia acumulada con el correspondiente a la frecuencia acumulada que sigue en la tabla. En el ejemplo: $\frac{137+138}{2} = 137.5$

Si no está presente: la *mediana* será el dato correspondiente a la frecuencia acumulada de la tabla que sea inmediatamente superior al valor calculado.

Medidas de Dispersión: informan acerca de la distribución de los datos.

- **Rango:** diferencia entre el dato mayor y el dato menor. Es muy poco usado.
- **Desviación Estándar o Típica:** mide la dispersión de los datos con respecto a la media. Es la medida de dispersión más utilizada en Estadística, ya que destaca las desviaciones pequeñas y grandes.

Si los datos no se repiten:
$$s = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}{n-1}}$$

Si los datos se repiten:
$$s = \sqrt{\frac{\sum [f_i \cdot (x_i - \bar{x})^2]}{n-1}}$$

(f_i es la frecuencia de x_i)

Esta desviación estándar es la *desviación estándar muestral*.

La *desviación estándar poblacional* se simboliza σ y su fórmula es: (N es el tamaño de la población y μ es la *media poblacional*)
Obsérvese que el denominador es N (no N-1).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{i=N} (x_i - \mu)^2}{N}}$$

¿Por qué se calcula s con $n-1$ y no con n ? La razón de $n-1$ es que cuando se divide por $n-1$, la varianza de la muestra (ver punto siguiente) proporciona una varianza estimada mucho más cercana a la varianza de la población, que cuando solo se divide por n . De hecho para tamaños grandes de n (por ejemplo superiores a 30), no existe ninguna diferencia si se divide por n ó por $n-1$. Los resultados son aproximadamente iguales, por lo tanto son aceptables. El factor $n-1$ es lo que se conoce como los "grados de libertad".

Una desviación estándar *pequeña* indica que los datos están concentrados muy cerca de la media, lo cual significa que la media obtenida es *muy confiable*.

Una desviación estándar *grande* indica que los datos están muy dispersos, lo cual significa que la media obtenida es *poco confiable*.

Pero ¿con qué criterio se decide que la desviación estándar calculada es "pequeña" o "grande"? Una forma de decidir esto es mediante el *coeficiente de variación* (ver más adelante).

- **Varianza:** cuadrado de la desviación estándar.

varianza(muestral) = s^2
varianza(poblacional) = σ^2

- **Coefficiente de Variación:** cociente entre la desviación estándar y la media (expresado en %). Se utiliza cuando se quiere comparar la variación entre muestras o entre poblaciones. También es útil para decidir si la desviación estándar está indicando que hay poca o mucha dispersión de datos.

$$CV(\text{muestral}) = \frac{s}{\bar{x}} \cdot 100 \qquad CV(\text{poblacional}) = \frac{\sigma}{\mu} \cdot 100$$

Si CV	{	<i>menor que 10%</i>	⇒ variación muy baja
		<i>entre 10% y 20%</i>	⇒ variación baja
		<i>entre 20% y 30%</i>	⇒ variación media
		<i>mayor que 30%</i>	⇒ variación alta

Hay que tener en cuenta que:

si $\bar{x} < 0$ o $\mu < 0$, se toma su valor absoluto

si $\bar{x} = 0$ o $\mu = 0$, CV no existe

Ejercicios de Estadística Descriptiva

1) Las estaturas en cm de un equipo juvenil de basketball son:

177 176 174 173 171 170 169 168 166 160

Hallar:	Media	$\bar{x} = 170.4 \text{ cm}$
	Moda	<i>no tiene</i>
	Mediana	<i>mediana = 170.5 cm</i>
	Rango	<i>rango = 17 cm</i>
	Desviación estándar	<i>s = 5.06 cm</i>
	Coefficiente de variación	<i>CV = 3%</i>

2) Supóngase que en Salto, en cierta semana de verano, las temperaturas diarias promedio (en °C) fueron:

35 33 30 36 40 37 38

Hallar:	Media	$\bar{x} = 35.57 \text{ °C}$
	Moda	<i>no tiene</i>
	Mediana	<i>mediana = 36 °C</i>
	Rango	<i>rango = 10°C</i>
	Desviación estándar	<i>s = 3.3 °C</i>
	Coefficiente de variación	<i>CV = 9.3%</i>

3) Las estaturas en cm de los alumnos de un grupo de 3° año de una escuela son:

138 144 130 146 128 145 133 129 143 136 137 138 129 133
 139 145 128 138 140 146 142 148 132 130 143 135 134 136
 138 131 141 139 133 130 139 135 138 147 137 135 133 132
 137 138 140 142 131 139

a) Hacer tabla de frecuencia (completa), histograma y polígono de frecuencia.

b) Hallar:	Media	$\bar{x} = 137.08 \text{ cm}$
	Moda	<i>moda = 138 cm</i>
	Mediana	<i>mediana = 137.5 cm</i>
	Rango	<i>rango = 20 cm</i>
	Desviación estándar	<i>s = 5.43 cm</i>
	Coefficiente de variación	<i>CV = 4%</i>

4) En una encuesta realizada a una muestra de 30 personas se les preguntó cuántas personas vivían en su casa. Las respuestas fueron:

6 5 5 2 3 2 3 4 4 6 7 3 2 7 7
 7 3 7 4 5 5 5 7 7 2 5 4 3 4 7

a) Hacer tabla de frecuencia (completa), histograma y polígono de frecuencia.

b) Hallar:	Media	$\bar{x} = 4.7$
	Moda	<i>moda = 7</i>
	Mediana	<i>mediana = 5</i>
	Desviación estándar	<i>s = 1.78</i>
	Coefficiente de variación	<i>CV = 37.9%</i>

c) ¿Qué porcentaje de las personas consultadas tienen 6 integrantes en su hogar? ¿Y menos de 4?

ESTADÍSTICA INFERENCIAL

TAMAÑO DE UNA MUESTRA

Representatividad de la muestra

La muestra debe reproducir las características del universo o población, por lo tanto surgen entonces dos preguntas:

- ✓ la cantidad de elementos que debe incluir la muestra
- ✓ hasta qué punto pueden generalizarse a la población

Ambas preguntas convergen en un problema de exactitud o precisión cuya finalidad es no incurrir en errores a la hora de obtener los resultados. No obstante, los errores son inevitables. Lo importante entonces es minimizarlos.

Existen dos tipos de errores:

- a) los **sistemáticos o distorsiones**, que son causados por factores externos a la muestra y que se pueden producir en cualquier momento de la investigación.
- b) el **error de muestreo**, de azar o de estimación, inevitable, ya que siempre habrá diferencia entre los valores medios de la muestra y los valores medios del universo o población. La magnitud de este error depende del tamaño de la muestra (a mayor tamaño de muestra menor error) y de la dispersión o desviación (a mayor dispersión mayor error). Cuando se toma una muestra, normalmente es porque deseamos hacer inferencia sobre una característica de la población (estatura promedio, porcentaje de fumadores, promedio de ingresos anuales, etc.) llamada parámetro. El error de muestreo es la diferencia entre el valor del parámetro poblacional (medida exacta) y el valor estimado por la muestra (medida aproximada).

Se concluye entonces que para que una muestra sea representativa, debe estar dentro de ciertos límites y proporciones establecidas por la estadística.

¿Por qué se usan las muestras para hacer inferencia y no toda la población?

1. *El costo de estudiar a todos los individuos de la población puede ser prohibitivo.*

Por ejemplo, en los Estados Unidos, las encuestadoras de estudios de mercado por lo general entrevistan menos de 2.000 familias de un total aproximado de 50 millones de éstas que existen allí. La encuestadora, para este tamaño de muestra, cobra aproximadamente \$40 mil por el estudio. Utilizando los 50 millones, tendría que cobrar \$1.000 millones.

2. *La confiabilidad que proporcionan las muestras.*

Siguiendo el ejemplo anterior, aún contando con los 50 millones de personas, que proporcionarían una confianza del 100%, probablemente una muestra de 2.000 personas podría tener una confiabilidad cercana al 95%.

3. *Imposibilidad de contactar a toda la población.*

Por ejemplo: imagine que estamos haciendo un estudio para determinar la velocidad promedio de cierto tipo de tigre de África. ¿Cómo podríamos convocarlos a todos?

4. *La naturaleza destructiva de ciertas pruebas.*

Si esta vez nuestro ejemplo consiste en determinar la calidad promedio de la producción de vinos en Chile, es fácil ver la implicación de probar todas las botellas de vino de la producción.

5. *Contactar toda la población supondría mucho tiempo.*

Supongamos que deseamos saber la estatura promedio del venezolano. ¿Cuánto tiempo llevaría hacer un estudio con toda la población, la cual en estos momentos ronda los 25 millones de personas?

Tipos de muestreo

Muestreo probabilístico. Es aquel en el cual los integrantes de la muestra son escogidos al azar.

- **Muestreo aleatorio simple.** Todos los elementos de la población tienen la misma probabilidad de ser seleccionados para la muestra. Por ejemplo, tengo 5 canarios en una jaula y tomo uno de ellos como muestra sin mirar.
- **Muestreo aleatorio sistemático.** Se divide la población en un cierto número de grupos, se selecciona un elemento al azar en el primer grupo y luego se repite el mismo número de elementos en los demás grupos. Ejemplo: tenemos una población de 100 elementos y queremos una muestra de 5 elementos. Dividimos el listado poblacional en 5 grupos de 20 elementos. Si en el primer grupo el seleccionado fue el número 12, en todos los demás será igual. Es decir, la muestra es: el número 12, el 32, el 52, el 72 y el 92.
- **Muestreo aleatorio estratificado.** Se divide a la población en subgrupos ó estratos y de cada uno se selecciona un cierto número de muestras, que depende de la representatividad porcentual de cada estrato. Ejemplo: deseamos conocer la opinión del electorado en torno a un candidato a un cargo político. Si por ejemplo un departamento o provincia de cierto país tiene el 35% del electorado, y la muestra debe ser de 1.000 personas, entonces en ese estado se toman $1.000 \times 35\% = 350$ muestras.
- **Muestreo por conglomerados.** Es similar al muestreo estratificado, pero los conglomerados no son estratos definidos por alguna condición en particular, sino que son grupos tomados sin ningún criterio particular de importancia. Ejemplo: para el ejemplo anterior, relativo a la opinión electoral, podemos dividir el país en 27 grupos tomados por la inicial del primer apellido.

Muestreo no-probabilístico o intencional. Es aquel en el cual los integrantes de la muestra son escogidos por el criterio del investigador.

Tamaño de la muestra

Cuando vamos a la parte práctica de la investigación estadística, necesitamos saber qué tamaño de muestra debemos tomar. **Éste tamaño no debe ser tomado en forma caprichosa, porque entonces podría no garantizarse la confiabilidad necesaria en la muestra. Además, tampoco tendríamos una medida para el error de muestreo.**

A continuación se presentan algunos métodos que se utilizan para determinar el tamaño apropiado de la muestra. Estos dependen de qué tipo de parámetro vamos a estimar (media, proporción) y del tamaño de la población (infinita o finita).

Para determinar el tamaño muestral de un estudio, debemos considerar diferentes situaciones:

- **Estudios para determinar parámetros:** Se pretende hacer inferencias a valores poblacionales (*proporciones, medias*) a partir de una muestra.
- **Estudios para contraste o prueba de hipótesis:** Se pretende *comparar* si las medias o las proporciones de las muestras son diferentes.

Estudios para determinar parámetros

A. Estimar una proporción

Para calcular el tamaño de la muestra, es necesario conocer:

- ✓ El **tamaño de la población** N
- ✓ La **proporción esperada** p (valor aproximado de la proporción que se quiere medir)
Si no se tiene ninguna idea acerca de su valor, puede obtenerse:
 - Revisando la literatura (buscar estudios pilotos previos).
 - Realizar un estudio piloto en una muestra pequeña y arbitraria para tener idea de su valor.
 - En caso de no tener dicha información, se utilizará un valor conservador $p = 0.5$ (50%).

Además, es necesario que el investigador decida:

- ✓ El **riesgo** (α) o el **nivel de confianza o de seguridad** ($1 - \alpha$). Este parámetro determina el valor z (ver tabla)
- ✓ El **error máximo permitido** (e) para nuestro estudio. Es el error máximo que estamos dispuestos a cometer en la estimación. Se expresa en forma de proporción o porcentaje. El máximo margen de error que se suele permitir es de 6 %.

Riesgo	Nivel de confianza o seguridad	z
1%	99%	2,576
2,5%	97,5%	2,24
5%	95%	1,96
10%	90%	1,645
15%	85%	1,44
20%	80%	1,282

$$n = \frac{N z^2 p(1-p)}{(N-1)e^2 + z^2 p(1-p)}$$

Tamaño de la muestra para estimar una proporción de una población finita

Ejemplo:

¿A cuántas personas tendríamos que estudiar de una población de 15000 habitantes para conocer la prevalencia de diabetes, si se fija una seguridad de 95% y un error máximo de 3%?

Se asume que la proporción esperada es próxima al 5% (si no tuviésemos ninguna idea de dicha proporción utilizaríamos el valor $p = 0,5$ (50%) que maximiza el tamaño muestral).

$$N = 15000$$

$$p = 0.05 \text{ (hay que expresar 5\% como 0.05)}$$

$$z = 1.96 \text{ (la seguridad es 95\%)}$$

$$e = 0.03 \text{ (en este caso deseamos un 3\%)}$$

$$n = \frac{(15000)(1.96)^2(0.05)(1-0.05)}{(15000-1)(0.03)^2 + (1.96)^2(0.05)(1-0.05)} = 200.06$$

Se debe tomar una muestra de 201 habitantes.

Continuando con el planteo anterior, veamos de qué tamaño debería ser la muestra si queremos más seguridad (menos riesgo) y si fijamos el error máximo permitido en un valor más pequeño que el anterior. Para esto fijemos una seguridad de 97.5% y un error máximo de 1%.

$$z = 2.24 \text{ (la seguridad es 97.5\%)}$$

$$e = 0.01 \text{ (en este caso deseamos un 1\%)}$$

$$n = \frac{(15000)(2.24)^2(0.05)(1-0.05)}{(15000-1)(0.01)^2 + (2.24)^2(0.05)(1-0.05)} = 2056.7$$

Se debe tomar una muestra de 2057 habitantes.

Entonces, puede observarse que *para obtener más precisión en la inferencia del parámetro poblacional, es necesario tomar una muestra de mayor tamaño y que, por otro lado, un tamaño muestral de poco tamaño (con lo tentador que este pueda llegar a ser en términos de simplicidad y costos) repercute negativamente en la precisión, aumentando la “incertidumbre” inferencial.*

Observación: Si el tamaño de la población es “infinita”, es decir muy grande ($N > 100.000$), entonces la fórmula anterior se puede aproximar mediante la siguiente:

$$n = \frac{z^2 p(1-p)}{e^2} \quad \text{Tamaño de la muestra para estimar una proporción de una población infinita}$$

De hecho si utilizamos esta aproximación en el ejemplo (aunque $N < 100.000$), se obtiene un valor muy parecido al anterior:

$$n = \frac{(1.96)^2(0.05)(1-0.05)}{(0.03)^2} = 202.75 \Rightarrow 203 \text{ habitantes}$$

B. Estimar una media

Para calcular el tamaño de la muestra, es necesario conocer:

- ✓ El **tamaño de la población** N
- ✓ La **varianza esperada** σ^2 (valor aproximado de la varianza de la población que se quiere medir)
Si no se tiene ninguna idea acerca de su valor, puede obtenerse:
 - Revisando la literatura (buscar estudios pilotos previos).
 - Realizar una muestra piloto, de tamaño pequeño y arbitrario, para tener una idea aproximada de dicha varianza.

Además, es necesario que el investigador decida:

- ✓ El **riesgo** (α) o el **nivel de confianza o de seguridad** ($1 - \alpha$). Este parámetro determina el valor z (ver tabla)
- ✓ El **error máximo permitido** (e) para nuestro estudio. Es el error máximo que estamos dispuestos a cometer en la estimación.

$$n = \frac{Nz^2\sigma^2}{(N-1)e^2 + z^2\sigma^2}$$

Tamaño de la muestra para estimar la media de una población finita

Ejemplo:

Se desea conocer la media de la glucemia basal de una población de 10000 habitantes, con una seguridad del 90 % y un error de ± 3 mg/dl y tenemos información por un estudio piloto o revisión bibliográfica que la varianza es de 250 mg/dl

$$N = 10000$$

$$z = 1.645 \text{ (la seguridad es del 90\%)}$$

$$\sigma^2 = 250$$

$e = 3$ (en el cálculo de medias el error no es porcentual, es un valor expresado en la misma unidad que la variable estudiada)

$$n = \frac{(10000)(1.645)^2(250)}{(10000-1)(3)^2 + (1.645)^2(250)} = 74.6 \Rightarrow 75 \text{ habitantes}$$

Observación: Si el tamaño de la población es “infinita”, es decir muy grande ($N > 100.000$) entonces la fórmula anterior se puede aproximar mediante la siguiente:

$$n = \frac{z^2\sigma^2}{e^2}$$

Tamaño de la muestra para estimar la media de una población infinita

Estudios para contraste o prueba de hipótesis

Estos estudios pretenden comparar si las medias o las proporciones de las muestras son diferentes. Por ejemplo en Medicina el investigador pretende comparar dos tratamientos.

A. Comparación de dos proporciones

Para calcular el tamaño de la muestra, es necesario:

- ✓ Tener una idea aproximada de las **proporciones de ambas muestras** (obtenerlas mediante consulta bibliográfica o de estudios previos; o realizar una prueba piloto). p_1 y p_2

Para fijar ideas, veamos un ejemplo de la medicina:

p_1 = Valor de la proporción en el grupo de referencia, placebo, control o tratamiento habitual.

p_2 = Valor de la proporción en el grupo del nuevo tratamiento, intervención o técnica.

- ✓ Hallar la **media de esas dos proporciones**:
$$p = \frac{p_1 + p_2}{2}$$

Además, es necesario que el investigador decida:

- ✓ **Seguridad** del estudio (riesgo de cometer un error α). Este parámetro determina el valor z_α (ver tabla)
- ✓ **Poder estadístico o Potencia estadística** (riesgo de cometer un error β). Este parámetro determina el valor z_β (ver tabla)
- ✓ Si la hipótesis va a ser **unilateral o bilateral**:

Unilateral: Cuando se considera que una de las proporciones debe ser mayor que la otra.

$$p_1 < p_2$$

Bilateral: Cuando se considera que las dos proporciones son distintas. $p_1 \neq p_2$

La bilateral es la hipótesis más conservadora.

Riesgo α	Z_{α}	
	Test unilateral	Test bilateral
0.200	0.842	1.282
0.150	1.036	1.440
0.100	1.282	1.645
0.050	1.645	1.960
0.025	1.960	2.240
0.010	2.326	2.576
Potencia Estadística $1 - \beta$	Z_{β}	
0.99	2.326	
0.95	1.645	
0.90	1.282	
0.85	1.036	
0.80	0.842	
0.75	0.674	
0.70	0.524	
0.65	0.385	
0.60	0.253	
0.55	0.126	
0.50	0.000	

$$n = \left(\frac{z_{\alpha} \sqrt{2p(1-p)} + z_{\beta} \sqrt{p_1(1-p_1) + p_2(1-p_2)}}{p_1 - p_2} \right)^2$$

n será la cantidad de sujetos necesarios en cada una de las muestras

Ejemplo:

Deseamos evaluar si el Tratamiento T_2 es mejor que el tratamiento T_1 para el alivio del dolor para lo que diseñamos un ensayo clínico. Sabemos por datos previos que la eficacia del fármaco habitual está alrededor del 70% y consideramos clínicamente relevante si el nuevo fármaco alivia el dolor en un 90%. Nuestro nivel de riesgo lo fijamos en 0.05 y deseamos un poder estadístico de un 80%.

$$p_1 = 0.7$$

$$p_2 = 0.9$$

$$p = \frac{0.7 + 0.9}{2} = 0.8$$

$$z_\alpha = 1.645 \text{ (riesgo} = 0.05; \text{ test unilateral } p_1 < p_2)$$

$$z_\beta = 0.842 \text{ (potencia} = 80\%)$$

$$n = \left(\frac{1.645\sqrt{2(0.8)(1-0.8)} + 0.842\sqrt{0.7(1-0.7) + 0.9(1-0.9)}}{0.7 - 0.9} \right)^2 = 48.4$$

En cada grupo precisamos 49 pacientes.

B. Comparación de dos medias

Para calcular el tamaño de la muestra, es necesario:

- ✓ Tener una idea aproximada de la **varianza** de la variable cuantitativa que tiene el grupo control o de referencia (σ^2)

Además, es necesario que el investigador decida:

- ✓ **Seguridad** del estudio (riesgo de cometer un error α). Este parámetro determina el valor z_α (ver tabla)
- ✓ **Poder estadístico** o **Potencia estadística** (riesgo de cometer un error β). Este parámetro determina el valor z_β (ver tabla)
- ✓ Si la hipótesis va a ser **unilateral** o **bilateral**:

Unilateral: Cuando se considera que una de las medias debe ser mayor que la otra. $\bar{x}_1 < \bar{x}_2$

Bilateral: Cuando se considera que las dos medias son distintas. $\bar{x}_1 \neq \bar{x}_2$

La bilateral es la hipótesis más conservadora.

- ✓ El **valor mínimo de la diferencia que se desea detectar** d (válido si los datos son cuantitativos)

$$n = \frac{2(z_\alpha + z_\beta)^2 \sigma^2}{d^2}$$

n será la cantidad de sujetos necesarios en cada una de las muestras

Ejemplo:

Deseamos utilizar un nuevo fármaco antidiabético y consideramos que sería clínicamente eficaz si lograrse un descenso de 15 mg/dl respecto al tratamiento habitual con el antidiabético estándar. Por estudios previos sabemos que la desviación típica de la glucemia en pacientes que reciben el tratamiento habitual es de 16 mg/dl. Aceptamos un riesgo de 0.05 y deseamos un poder estadístico de 90% para detectar diferencias si es que existen.

$$z_{\alpha} = 1.645 \text{ (riesgo} = 0.05; \text{ test unilateral } \bar{x}_1 < \bar{x}_2)$$

$$z_{\beta} = 1.282 \text{ (potencia} = 90\%)$$

$$\sigma = 16$$

$$d = 15$$

$$n = \frac{2(1.645 + 1.282)^2 (16)^2}{15^2} = 19.5$$

Precisamos 20 pacientes en cada grupo.

El tamaño muestral ajustado a las pérdidas

En todos los estudios de índole médica es preciso estimar las posibles pérdidas de pacientes por razones diversas (pérdida de información, abandono, no respuesta, ...) por lo que se debe incrementar el tamaño muestral para tener en cuenta dichas pérdidas.

El tamaño muestral ajustado a las pérdidas se puede calcular mediante la siguiente fórmula:

$$\text{Muestra ajustada a las pérdidas} = \frac{n}{1 - R}$$

n = número de sujetos sin pérdidas

R = proporción esperada de pérdidas

Así por ejemplo si en el estudio anterior esperamos tener un 15% de pérdidas, el tamaño muestral necesario sería:

$$\frac{20}{1 - 0.15} = 23.5$$

Es decir que se necesitarán 24 pacientes en cada grupo.

INTERVALO DE CONFIANZA

Intervalo de confianza para una media μ

Nivel de confianza (100c)	$z_{c/2}$
90%	1,645
95%	1,96
98%	2,33
99%	2,58

Muestra grande ($n \geq 30$)

No importa si la población no es Normal.

Población infinita o finita con reemplazo

$$\bar{X} \pm z_{c/2} \frac{S}{\sqrt{n}}$$

Población finita sin reemplazo

$$\bar{X} \pm z_{c/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

\bar{X}	media de la muestra
S	desviación estándar de la muestra
c	nivel de confianza (0,90 a 0,99)
z	distribución Normal
n	tamaño de la muestra
N	tamaño de la población

Ejemplos:

- 1) $n = 49$ $\bar{X} = 72$ $S = 9$
 99% $68.68 \leq \mu \leq 75.32$
- 2) $n = 36$ $\bar{X} = 57$ $S = 6$
 90% $55.355 \leq \mu \leq 58.645$
- 3) $n = 40$ $\bar{X} = 2400$ $S = 150$ $N = 500$
 95% $2355 \leq \mu \leq 2445$
 99% $2341 \leq \mu \leq 2459$
- 4) $n = 60$ $\bar{X} = 11.09$ $S = 0.73$
 95% $10.91 \leq \mu \leq 11.27$
 99% $10.85 \leq \mu \leq 11.33$
- 5) $n = 250$ $\bar{X} = 0.72642$ $S = 0.00058$
 90% $0.72636 \leq \mu \leq 0.72648$
 95% $0.726348 \leq \mu \leq 0.726492$
 98% $0.726335 \leq \mu \leq 0.726505$
 99% $0.726325 \leq \mu \leq 0.726515$
- 6) $n = 100$ $\bar{X} = 67.45$ $S = 2.93$
 95% $66.88 \leq \mu \leq 68.02$ ($N = 1546$) $66.89 \leq \mu \leq 68.01$
 99% $66.69 \leq \mu \leq 68.21$ ($N = 1546$) $66.72 \leq \mu \leq 68.18$
- 7) $n = 200$ $\bar{X} = 0.824$ $S = 0.042$
 90% $0.8191 \leq \mu \leq 0.8289$
 95% $0.8182 \leq \mu \leq 0.8298$
 98% $0.8171 \leq \mu \leq 0.8309$
 99% $0.8163 \leq \mu \leq 0.8317$
- 8) $n = 50$ $\bar{X} = 75$ $S = 10$ $N = 200$
 95% $72.6 \leq \mu \leq 77.4$
- 9) $n = 65$ $\bar{X} = 72$ $S = 9$
 90% $70.15 \leq \mu \leq 73.85$

Muestra pequeña ($n < 30$)

La población debe ser Normal.

$$\bar{X} \pm t_{c,n-1} \frac{S}{\sqrt{n-1}}$$

\bar{X}	media de la muestra
S	desviación estándar de la muestra
c	nivel de confianza (0,90 a 0,99)
t	distribución t de Student
n	tamaño de la muestra

Ejemplos:

- 1) $n = 16$ $\bar{X} = 2311$ $S = 294$
 90% $2178 \leq \mu \leq 2444$
- 2) $n = 10$ $\bar{X} = 4.38$ $S = 0.06$
 95% $4.335 \leq \mu \leq 4.425$
 99% $4.315 \leq \mu \leq 4.445$
- 3) $n = 12$ $\bar{X} = 7.38$ $S = 1.24$
 95% $6.56 \leq \mu \leq 8.2$
 99% $6.22 \leq \mu \leq 8.54$
- 4) $n = 5$ $\bar{X} = 0.298$ $S = 0.0214$
 95% $0.268 \leq \mu \leq 0.328$
 99% $0.249 \leq \mu \leq 0.347$
- 5) $n = 25$ $\bar{X} = 191$ $S = 20$
 95% $182.59 \leq \mu \leq 199.41$
- 6) $n = 20$ $\bar{X} = 46.5$ $S = 4.985$
 95% $44.11 \leq \mu \leq 48.89$
- 7) $n = 7$ $\bar{X} = 25$ $S = 9$
 99% $11.37 \leq \mu \leq 38.63$
- 8) $n = 9$ $\bar{X} = 253$ $S = 30$
 99% $217.36 \leq \mu \leq 288.64$
- 9) $n = 17$ $\bar{X} = 72$ $S = 9$
 90% $68.06 \leq \mu \leq 75.94$

Intervalo de confianza para una proporción p

Población infinita o finita con reemplazo

$$\hat{p} \pm z_{c/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Población finita sin reemplazo

$$\hat{p} \pm z_{c/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

- \hat{p} proporción de la muestra ($0 < \hat{p} < 1$)
 c nivel de confianza (0,90 a 0,99)
 z distribución Normal
 n tamaño de la muestra
 N tamaño de la población

Ejemplos:

1) $n = 100$ $\hat{p} = 0.55$
 95% $0.45 \leq p \leq 0.65$
 99% $0.42 \leq p \leq 0.68$

2) $n = 40$ $\hat{p} = 0.6$
 95% $0.45 \leq p \leq 0.75$

3) $n = 240$ $\hat{p} = 0.2$
 95% $0.15 \leq p \leq 0.25$

4) $n = 50$ $\hat{p} = 0.15$
 95% $0.05 \leq p \leq 0.25$

5) $n = 60$ (con reemplazo) $\hat{p} = 0.70$
 95% $0.58 \leq p \leq 0.82$
 99% $0.55 \leq p \leq 0.85$

Intervalo de confianza para una desviación estándar σ

Muestra pequeña ($n < 100$)

$$S \cdot \sqrt{\frac{n}{\chi^2_{\frac{1+c}{2}, n-1}}} \leq \sigma \leq S \cdot \sqrt{\frac{n}{\chi^2_{\frac{1-c}{2}, n-1}}}$$

c	(1+c) / 2	(1-c) / 2
0,90	0,95	0,05
0,95	0,975	0,025
0,98	0,99	0,01
0,99	0,995	0,005

si $0 < n < 30 \rightarrow$ hallar χ^2 mediante: Tabla

$30 \leq n < 100 \rightarrow$ hallar χ^2 mediante: $\chi^2_{p,v} \cong \frac{(z_{p-0.5} + \sqrt{2v-1})^2}{2}$

$$\Rightarrow \begin{cases} \chi^2_{\frac{1+c}{2}, n-1} \cong \frac{(z_{c/2} + \sqrt{2n-3})^2}{2} \\ \chi^2_{\frac{1-c}{2}, n-1} \cong \frac{(-z_{c/2} + \sqrt{2n-3})^2}{2} \end{cases}$$

Muestra grande ($n \geq 100$)

$$S \pm z_{c/2} \frac{S}{\sqrt{2n}}$$

- S desviación estándar de la muestra
- c nivel de confianza (0,90 a 0,99)
- z distribución Normal
- χ^2 distribución Chi Cuadrado
- n tamaño de la muestra

Ejemplos:

- 1) $n = 200$ $S = 100$
95% $90.2 \leq \sigma \leq 109.8$
99% $87.1 \leq \sigma \leq 112.9$

- 2) $n = 16$ $S = 2.40$
95% $1.83 \leq \sigma \leq 3.84$
99% $1.68 \leq \sigma \leq 4.49$

- 3) $n = 100$ $S = 1800$
95% $1551 \leq \sigma \leq 2049$
99% $1472 \leq \sigma \leq 2128$

- 4) $n = 10$ $S = 120$
95% $87.0 \leq \sigma \leq 230.9$
99% $78.1 \leq \sigma \leq 288.5$

- 5) $n = 25$ $S = 120$
95% $95.6 \leq \sigma \leq 170.4$
99% $88.9 \leq \sigma \leq 190.8$

- 6) $n = 100$ $S = 120$
95% $103.4 \leq \sigma \leq 136.6$
99% $98.1 \leq \sigma \leq 141.9$

- 7) $n = 23$ $S = 7$
99% $5.13 \leq \sigma \leq 11.42$

- 8) $n = 7$ $S = 7.874$
95% $5.49 \leq \sigma \leq 18.71$

- 9) $n = 7$ $S = 3$
99% $1.845 \leq \sigma \leq 9.654$

- 10) $n = 23$ $S = 10.758$
95% $8.505 \leq \sigma \leq 15.556$

- 11) $n = 19$ $S = 11.68$
99% $8.35 \leq \sigma \leq 20.35$

..:| Fuentes Consultadas |:..

Bonilla; “Elementos de Estadística”

Canavos, G. C.; “Probabilidad y Estadística”, Primera edición, Editorial McGraw-Hill

Duffour, G.; “Matemática de Cuarto”, Tercera edición, Ediciones Matemática 2000

Guerra Bustillo, C. y otros; “Estadística”, edición 1987, Editorial Pueblo y Educación

Lipschutz, S.; Lipson, M.; “Probabilidad”, Segunda edición, Editorial McGraw-Hill

Pita Fernández, S.; “Determinación del tamaño muestral”, (Internet)

Pons, J.C.; “Estadística para metodología”, (Internet)

Spiegel, M. y otros; “Probabilidad y Estadística”, Segunda edición, Editorial McGraw-Hill

(Aquí están reproducidos textualmente algunos párrafos y frases de esos trabajos)