DOCUMENTO

1

CÁLCULO DE MUESTRA

(FÓRMULAS)

EL TAMAÑO DE LA MUESTRA MEDIANTE FÓRMULAS

Para determinar el tamaño de muestra mediante fórmulas es necesario entender los siguientes términos y sus definiciones:

La población o universo, a la que se le suele denominar como *N*, es un conjunto de elementos.

La muestra, a la que se le simboliza como *n*, es un subconjunto de la población *N*.

En una población *N* (previamente delimitada por el planteamiento del problema de investigación), nos interesa establecer valores de las características de los elementos de *N*.

Nos concierne saber valores promedio en la población, lo cual se expresa como:

$\overline{Y}$ = al valor de una variable determinada (Y) que nos interesa conocer, digamos un promedio.

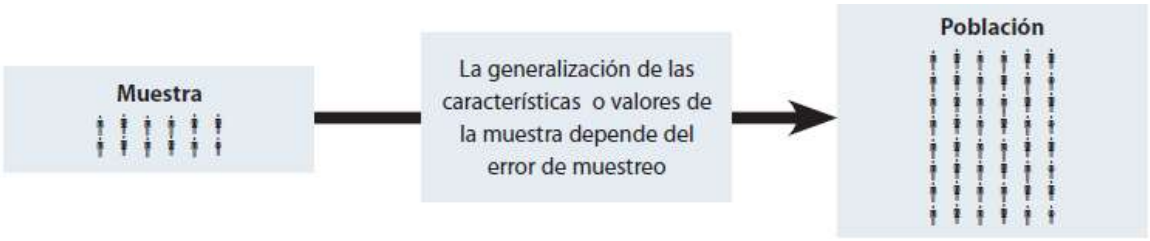
También nos interesa conocer:

V = la varianza de la población con respecto a determinadas variables (la varianza indica la variabilidad).

Como los valores de la población no se determinan directamente, seleccionamos una muestra *n*, además, a través de estimados en la muestra,

inferimos valores de la población ($\bar{y}$ será la estimación del valor de $\bar{Y}$, el cual desconocemos).

Figura 1 Esquema de la generalización de la muestra a la población.



En la muestra, $\bar{y}$ es un estimado promedio que podemos determinar. Sabemos que en nuestra estimación habrá una diferencia ($\bar{Y} - \bar{y} = ?$), es decir, un error, el cual dependerá del número de elementos muestreados. A dicho error se le conoce como error estándar (*se*, por sus siglas en inglés).

se = la desviación estándar de la distribución muestral y representa la fluctuación de $\bar{y}$.

$(se)^2$ = el error estándar al cuadrado, cuya fórmula nos servirá para calcular la varianza (*V*) de la población *(N)*, así como la varianza de la muestra (*n*) será la expresión s^2.

s^2 = varianza de la muestra, la cual podrá determinarse en términos de probabilidad, donde $s^2 = p\,(1 - p)$.

p = porcentaje estimado de la muestra, probabilidad de ocurrencia del fenómeno, la cual se estima sobre marcos de muestreo previos o se define, la certeza total siempre es igual a uno, las posibilidades a partir de esto son "*p*" de que sí ocurra y "*q*" de que no ocurra ($p + q = 1$). De aquí se deriva que $1 - p = q$.

Como se habrá podido observar, cuando hablamos de un término de la muestra se simboliza con una letra minúscula (*n, s, se*). Si se trata de un término de la población, se simboliza con una letra mayúscula (*N, S*).

Lo que se busca es lo mismo que con STATS®: dado que una población es de *N*, ¿cuál es el menor número de unidades muestrales que necesito para conformar una muestra (*n*) que me asegure un determinado nivel de error estándar, digamos menor de 0.01?

La respuesta a esta pregunta busca encontrar la probabilidad de ocurrencia de $\bar{Y}$, así como que el estimado de $\bar{y}$ se acerque a $\bar{Y}$, el valor real de la población. Si establecemos el error estándar y lo fijamos en 0.01, sugerimos que esta fluctuación promedio de nuestro estimado $\bar{y}$ con respecto a los valores reales de la población $\bar{Y}$ no sea > 0.01, es decir, que de 100 casos, 99 veces mi predicción sea correcta y que el valor de $\bar{y}$ se sitúe en un intervalo de confianza que comprenda el valor de $\bar{Y}$.

Resumiendo, para una determinada varianza (*V*) de *Y*, ¿qué tan grande debe ser mi muestra? Ello se determina en dos pasos:

1. $n' = \frac{s^2}{V^2}$ = Tamaño provisional de la muestra[1] = varianza de la muestra/varianza de la población.

2. $n = \frac{n'}{1 + n'/N}$

Pongamos el siguiente caso:[2] supongamos que necesitamos entrevistar a directores de recursos humanos de empresas para determinar su ideología respecto a cómo tratan a sus colaboradores. Requerimos extraer una muestra probabilística de un universo o población de 1 176 organizaciones que cuentan con director de recursos humanos (*N*). ¿Cuál es entonces el número de empresas (*n*) que se debe considerar, para tener un error estándar menor de 0.015, y dado que la población total es de 1 176?

N = tamaño de la población de 1 176 empresas.

---

[1] Se corrige con otros datos, ajustándose si se conoce el tamaño de la población N.

[2] Este ejemplo fue tratado en el texto impreso al comentar la muestra probabilística estratificada.

$\bar{y}$ = valor promedio de una variable = 1, un director de recursos humanos por empresa.

se = error estándar = 0.015, determinado por nosotros. Nivel deseado de error.

V^2 = varianza de la población al cuadrado. Su definición se^2: cuadrado del error estándar.

s^2 = varianza de la muestra expresada como la probabilidad de ocurrencia de $\bar{y}$.

p = 0.9

n' = tamaño de la muestra sin ajustar.

n = tamaño de la muestra.

Si lo sustituimos, tenemos que:

$$n' = \frac{s^2}{V^2}$$

$$s^2 = p(1 - p) = 0.9(1 - 0.9) = 0.09$$

$$V^2 = (0.015)^2 = 0.000225$$

$$n' = \frac{0.09}{0.000225} = 400$$

$$n = \frac{n'}{1 + (n'/N)} = \frac{400}{1 + (400/1176)} = 298.5$$

$$n = 298 \text{ casos}$$

Es decir, para nuestra investigación necesitaremos una muestra de 298 directores de recursos humanos.

Se trata del primer procedimiento para obtener la muestra probabilística: determinar su tamaño con base en estimados de la población. El segundo procedimiento estriba en cómo y de dónde seleccionar a esos 298 directores o casos.

¿CÓMO HACER UNA MUESTRA PROBABILÍSTICA ESTRATIFICADA Y POR RACIMOS?

En el capítulo 8 del texto impreso respecto a que en ocasiones se combinan tipos de muestreo, por ejemplo: una muestra probabilística estratificada y por racimos. Ahora lo ejemplificamos.

EJEMPLO

***Problema de investigación*:**

Una estación de radio local necesita saber con precisión cómo utilizan la radio los adultos de una ciudad de 2 500 000 habitantes, con la finalidad de planear sus estrategias. Es decir, qué tanto radio escuchan, a qué horas, qué contenidos prefieren y sus opiniones con respecto a los programas noticiosos.

***Procedimientos*:**

Se diseñará un cuestionario que indague estas áreas sobre el uso de la radio. Los cuestionarios se aplicarán por entrevistadores a una muestra de adultos.

Población:

Todos aquellos sujetos hombres o mujeres de más de 18 años de edad y que vivan en una casa o un departamento propio o rentado de la ciudad.

***Diseño por racimos*:**

Los directivos de la estación de radio desconocen el número total de personas con las características señaladas. Sin embargo, nos piden que diseñemos una muestra que abarque a todos los sujetos adultos de la ciudad, por edad cronológica y por ser jefes de familia, es decir, se excluye a los adultos dependientes.

Tenemos entonces que $n' = \frac{s^2}{V^2}$ para una muestra probabilística simple.

$$n' = \frac{s^2}{V^2} = \frac{p(1-p)}{(0.015)^2} = \text{error estándar } \frac{0.5(1-0.5) = 0.25}{0.000225}$$

$$n' = 1\,111.11$$

$$n = \frac{n'}{1+n'/N'} = \frac{1111.11}{1+1111.11/5000} = 909.0902$$

$$n = 909$$

Necesitaremos una muestra de 909 cuadras para estimar los valores de la población con una probabilidad de error menor a 0.015.

Sabemos que la población *N* = 5 000 cuadras está dividida por estudios previos de acuerdo con cuatro estratos socioeconómicos, que categorizan esa población según el ingreso mensual promedio de sus habitantes, de manera que se distribuyen como sigue:

ESTRATO	**NÚM. DE CUADRAS**
1	270
2	1 940
3	2 000
4	790
	N = 5 000

¿Cómo distribuiremos los 909 elementos muestrales de *n*, para optimizar la muestra, de acuerdo con la distribución de la población en los cuatro estratos socioeconómicos?

Estratificación de la muestra:

$$\sum fh = \frac{n}{N} = ksh$$

$$fh = \frac{909}{5000} = 0.1818$$

ESTRATO	NÚM. DE CUADRAS	*f h* = 0.1818	*nh**
1	270	(0.1818)	49
2	1 940	(0.1818)	353
3	2 000	(0.1818)	364
4	790	(0.1818)	143
	N = 5 000		*n* =909

* Se redondeó para cuadrar el ejemplo, recordemos que son individuos y no se pueden fragmentar. A veces es a la alta o a la baja, y en ocasiones podemos aumentar la muestra en un caso para que el redondeo sea ideal.

En principio tenemos que de 5 000 cuadras se seleccionarán 49 del estrato uno, 353 del estrato dos, 364 del estrato tres y 143 del estrato 4. Esta selección comprende la elección de los racimos, los cuales se pueden numerar y elegir aleatoriamente hasta completar el número de

cada estrato. En una última etapa, se seleccionan los participantes dentro de cada racimo. Este procedimiento también se hace de manera aleatoria, hasta lograr un número de personas (unidades de análisis) determinados en cada racimo. A continuación presentamos dicho procedimiento.

ESTRATO	*Nh* CUADRAS	*nh*	NÚMERO DE HOGARES PARTICIPANTES EN CADA CUADRA	TOTAL DE HOGARES POR ESTRATO
1	270	49	20	980
2	1 940	353	20	7 060
3	2 000	364	20	7 280
4	790	143	20	2 860
	N = 5 000	*n* = 909		18 180

NÚMEROS RANDOM O NÚMEROS ALEATORIOS

El uso de números *random* no significa la selección azarosa o fortuita, sino la utilización de una tabla de números que implica un mecanismo de probabilidad muy bien diseñado. La tabla más completa de números aleatorios fue producida por la Corporación Rand (CR), y éstos fueron generados con una especie de ruleta electrónica. Existe una tabla de un millón de dígitos publicada por esta organización (pero su costo asciende a varias decenas de dólares, se puede comprar en una página de CR: http://www.rand.org/pubs/monograph_reports/MR1418.html). Algunas partes de tal tabla se encuentran en los apéndices de ciertos libros de estadística (y desde luego, en la presente obra: en apéndice 4, "tablas estadísticas", al final, tabla 5: *Números aleatorios*); o bien, en diversas páginas web. Un fragmento de la tabla original con números *random* se muestra en la tabla 1 (disculpen la repetición de términos).

Tabla 1 Números aleatorios o *random*.

26804	29273	79811	45610	22879	72538	70157	17683	67942	52846
90720	96215	48537	94756	18124	89051	27999	88513	35943	67290
85027	59207	76180	41416	48521	15720	90258	95598	10822	93074
09362	49674	65953	96702	20772	12069	49901	08913	12510	64899
64590	04104	16770	79237	82158	04553	93000	18585	72279	01916
06432	08525	66864	20507	92817	39800	98820	18120	81860	68065
02101	60119	95836	88949	89312	82716	34705	12795	58424	69700
19337	96983	60321	62194	08574	81896	00390	75024	66220	16494
75277	47880	07952	35832	41655	27155	95189	00400	06649	53040
59535	75885	31648	88202	63899	40911	78138	26376	06641	97291
76310	79385	84639	27804	48889	80070	64689	99310	04232	84008
12805	65754	96887	67060	88413	31883	79233	99603	68989	80233
32242	73807	48321	67123	40637	14102	55550	89992	80593	64642
16212	84706	69274	13252	78974	10781	43629	36223	36042	75492
75362	83633	25620	24828	59345	40653	85639	42613	40242	43160
34703	93445	82051	53437	53717	48719	71858	11230	26076	44018
01556	58563	36828	85053	39025	16688	69524	81885	31911	13098
22211	86468	76295	16663	39489	18400	53155	92087	63942	99827
01534	70128	14111	77065	99358	28443	68135	61696	55241	61867
09647	32348	56909	40951	00440	10305	58160	62235	89455	73095
97021	23763	18491	65056	95283	92232	86695	78699	79666	88574
25469	63708	78718	35014	40387	15921	58080	03936	15953	59658
40337	48522	11418	00090	41779	54499	08623	49092	65431	11390
33491	98685	92536	51626	85787	47841	95787	70139	42383	44187
44764	14986	16642	19429	01960	22833	80055	39851	47350	70337

Fuente: Rand Corporation.

Si continuamos con el ejemplo anterior, determinaremos una muestra de 909 manzanas o cuadras, y a partir de este número se determina una submuestra para cada estrato. Véase que para el estrato uno, la población es de 270 manzanas o cuadras. Numeramos en nuestro listado o mapa las 270 y seleccionamos (a partir de la tabla de números *random* o aleatorios) los 49 casos que constituirán nuestra muestra.

Se eligen aquellos casos que se dictaminen en la tabla de números *random*, hasta completar el tamaño de la muestra. Los números pueden recorrerse hacia arriba, hacia abajo o de manera horizontal. Al final siempre se logra que cada elemento muestral tenga la misma probabilidad de ser escogido. Se seleccionan aquellos números que contenga el listado. Si en nuestro ejemplo la población es de 270, se escogen los tres últimos dígitos y se procede a seleccionar los casos, hasta completar el número de elementos muestrales necesarios: 49 manzanas (ver tabla 2).

Como puede verse, en la tabla 2 se eligen sólo las primeras ocho manzanas (de las 49 requeridas) para no prolongar el ejemplo (las ocho están numeradas entre paréntesis). Una vez seleccionadas las 49 manzanas se ubican en un mapa o

directorio y acudimos a los hogares (veinte en cada una de las 49 manzanas) y entrevistamos a los adultos, jefes de familia (en el ejemplo, 980).

Tabla 2 Selección muestral basada en la tabla de números aleatorios.

78 986	45 961	28 281	82 933	24 786	55 586
83 830	59 025	40 379	99 989	63 822	99 974
(1)30 226	19 863	(5)95 039	08 909	(7)48 197	(8)23 270
(2)02 073	(4)59 042	26 440	(6)16 161	14 496	24 786
(3)05 250	47 552	95 659	92 356	13 334	23 471

Pero este procedimiento de usar los números *random* o aleatorios para obtener las unidades o casos de la muestra es mucho más complejo que hacerlo mediante STATS®. Era muy típico hace un par de décadas o antes, hoy se usan programas como STATS®. Sin embargo, algunos profesores prefieren los cálculos manuales y mecánicos.