MUESTREO ALEATORIO SIMPLE

Esta clase profundizaremos sobre el muestreo aleatorio simple. La clase anterior señalamos las características generales de este tipo de muestreo, pero en la práctica es algo más complejo y depende del tamaño de la muestra y si esta es finita o infinita. Por ejemplo, el número de alumnos dentro de esta sala es un número finito, mientras que el número de personas que son atendidas en un local de comida rápida es un número infinito. Para cada una de estas situaciones, la pregunta clave es cómo asignarles un número para que todos tengan la misma probabilidad de ser elegidos en la muestra. Eso veremos hoy.

En cualquier caso, siempre necesitamos definir la muestra apropiada para hacer una inferencia. Por ejemplo, supongamos que queremos ver la recepción que ha tenido un perfume para personas y que nuestra muestra contiene personas de 40 a 60 años. Como las personas van cambiando sus gustos a lo largo de su vida, los datos que extraigamos de esa muestra no serían útiles para concluir sobre el gusto de personas de otras edades.

En esta clase primero estudiaremos muestreo aleatorio simple para casos finitos y luego para casos infinitos.

Muestreo de una población finita

Si la población a la que nos enfrentamos es finita, supongamos de tamaño N, entonces un muestreo aleatorio de tamaño n sería definido de la siguiente forma:

- Una muestra aleatoria de tamaño n de una población finita de tamaño N es una muestra seleccionada de manera que cada posible muestra de tamaño n tenga la misma probabilidad de ser seleccionada.

Para entender lo anterior, quizás es conveniente pensar en qué NO ES una muestra aleatoria simple. Nuestro ejemplo de la introducción es un caso no aleatorio. Nuestra población eran todas las personas, ya que queríamos un perfume para ellas. Sin embargo, al elegir una muestra, solo seleccionamos a las personas que estaban entre 40 y 60 años. Supongamos, por ahora, que existen 1000 personas, y que 100 de ellas están en el tramo etario 40-60. Si salgo a la calle, y selecciono a 100 personas al azar, lo que va a pasar es que tendré personas de todos los tramos etarios. Habrá personas entre 40 y 60, pero también existirán personas de 18, de 25, de 30 años, etc. Si repito este ejercicio durante un mes, saliendo a la calle todos los días a reclutar 100 personas, es casi imposible que justo encuentre solo a personas que tengan entre 40 y 60 años. En otras palabras, la probabilidad de que encuentre una muestra con solo personas entre 40 y 60 años, cuando en la población hay personas de todas las edades, es muy baja. En cambio, una muestra con personas de varias edades, tendrán probabilidades más altas de ocurrir.

De lo anterior podemos notar que, para que una muestra tenga la misma probabilidad de ocurrir que otras muestras, tenemos que elegir de tal forma que estas se parezcan a la población. Para hacer esto, la mejor alternativa es elegir al azar a los miembros de la muestra, sin poner ninguna condición previa. De ahí el nombre de muestreo aleatorio.

Para llevar a cabo un muestreo aleatorio cuando ya tenemos datos de la población, lo que hacemos es enumerarlos, y luego seleccionar números al azar. Esto podría hacerse con una tómbola, pero hoy en día los softwares nos permiten generar números aleatorios, que cumplen esta función de forma mejor y más rápida.
Existen dos posibilidades de seleccionar los números. Al ser aleatorios, podría pasar que la observación número 1435 aparezca dos veces. Si permitimos que esto pase, hablamos de un muestreo aleatorio con reemplazo. Sin embargo,

podríamos necesitar prohibir esto, ya que no queremos que, por ejemplo, los gustos de una mujer determinada sean registrados más de una vez. Si activamos esta restricción, hablamos de muestro aleatorio sin reemplazo. Esta última forma es la más común y, a menos que se mencione lo contrario, es a lo que nos referiremos cuando hablemos de muestreo aleatorio.

Número de muestras aleatorias distintas de tamaño n que podemos tomar de una población finita de tamaño N:

$$\frac{N!}{n!\,(N-n)!}$$

Esto significa que el tamaño N de la población factorial, dividido por el tamaño de la muestra n factorial y por la diferencia entre el tamaño de la población N y de la muestra n, factorial.

Ejercicio

Para preparar muestras aleatorias finitas podemos usar una tabla como la siguiente con números aleatorios y organizando con inicios de filas, columnas y números al azar.

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289

Por ejemplo, el centro de estudiantes de la facultad quiere ver el número de estudiantes que está a favor de suspender el área de estadística en la escuela. Existe una lista con los 850 alumnos de la facultad enumerados. Entonces, tomando números aleatorios, avanzando desde la segunda fila, podemos tomar una muestra con 10 alumnos de la escuela. En este caso los números seleccionados serían: 470-695-436-791-150-830-301-41-200-306

88547 09896 95436 79115 08303 01041 20030 63754 08459 28364

El resultado de la muestra de estos 10 estudiantes fue que no aumenten, ni disminuyan los cursos de Estadística, pero que no tengan notas bajo 4.

Muestreo de una población infinita

Anteriormente, habíamos supuesto que existían solo 1000 personas. Si levantamos este supuesto, la realidad es bastante diferente. En Chile existen aproximadamente 7 millones de personas. Es un número tan grande, que nos conviene pensar en el cómo infinito. De este modo, tenemos que definir ahora una muestra para una población infinita, el cual debe satisfacer dos condiciones:

1. Cada uno de los elementos seleccionados proviene de la población.
2. Cada elemento se selecciona independientemente.

Lo primero, en nuestro ejemplo, es fácil de satisfacer, basta con ver que la persona sea, en efecto, una mujer. Lo segundo es más complicado. En nuestro ejemplo anterior, había solo 1000 personas, por lo tanto, en los 30 días que repetimos el ejercicio es probable que les hayamos preguntado a la mayoría. Sin embargo, ahora que levantamos el supuesto, tenemos que tener cuidado de como seleccionamos la muestra. Por ejemplo, si vamos a la salida de un local de La Dehesa, las personas tendrán perfiles similares, y bastante diferentes que si vamos a la salida de un local de Santiago centro. Lo mismo sucede si solo le preguntáramos a nuestras amigas; lo más probable es que nuestras amigas no sean todas iguales, pero si deben tener algunas características en común, lo que las hace dependientes. Por lo tanto, para realizar un muestreo aleatorio de esta población tan grande, tendríamos que poner encuestadores en varios puntos de la ciudad, que entrevistasen a personas variadas, para así eliminar el sesgo de estar solo en un lugar.

El ejemplo anterior es fácil de modificar para entender otros tipos de muestras. Por ejemplo, el perfume puede, y lo más probable es que así sea, ser solamente para un público dirigido, por ejemplo, a las jóvenes. No obstante, es interesante notar que, según como definamos nuestro objetivo y nuestra población, tendremos que usar uno u otro método de muestreo.

TAMAÑO MUESTRAL TEOREMA CENTRAL DEL LÍMITE

Distribución muestral de la media de las muestras

Para poder acercarnos al Teorema Central del Límite, primero tenemos que ver la Distribución Muestral de la Media de las Muestras, si bien puede ser medio capcioso, usemos este simple ejemplo para poder identificar este concepto. Si dividimos el curso en 5, seguramente tendríamos cinco muestras de 9 alumnos aproximadamente. Si calculamos un estadístico cualquiera, por ejemplo la media de la altura de cada muestra, tendríamos 5 nuevos valores $\bar{x}_i, i = 1,..,5$. Donde asociamos estos valores a una nueva variable aleatoria $\bar{X}$. Su distribución la llamaremos distribución muestral.

Esto nos sirve para la siguiente relación, si tenemos una variable aleatoria cualquiera X, de media μ y desviación estándar σ, entonces:

1. Todas las muestras aleatorias posibles de tamaño n, se cumple que $\mu_{\bar{x}} = \mu$ y $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. La media de las medias muestrales se denotan por $\mu_{\bar{x}}$, lo mismo sucede con la desviación estándar de las medias muestrales, denotadas por $\sigma_{\bar{x}}$.
2. Si X sigue una distribución normal, $\bar{X}$ también sigue una distribución normal.

Teorema Central del Límite

Como vimos hace algunas clases el teorema central del limite reúne un conjunto de variables aleatorias que se distribuyen independiente e idénticamente entre sí. Con media μ y varianza σ^2 distinta de cero.

A medida que aumenta el tamaño (n) de las muestras, la distribución de las medias muestrales se aproxima a la de una normal:

$$\bar{X} = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Esto nos dice que esta nueva variable aleatoria $\bar{X}$, se distribuye normal, con media mu y desviación estándar sigma dividido por la raíz del tamaño muestral.

Hay ciertas reglas básicas de uso común sobre este teorema que es muy útil tenerlas en cuenta:

1. Mientras mayor sea el tamaño de las muestras más se acerca a una distribución normal $(n > 30)$. Esto sucede porque a medida que aumenta n es más exacta la aproximación.
2. Si la población está distribuida normal, todas las muestras también se distribuyen normal para cualquier tamaño de n.

Ejercicio

Supongamos que el promedio de los curso de Inferencia Estadística del semestre de primavera 2015 es de 6,3. También suponemos que la desviación estándar de la población es de 0,5. Si escogemos aleatoriamente una muestra de los 200 alumnos que están cursando el curso, con $n = 40$. ¿Cuál sería la desviación estándar de la muestra?

Como sabemos:

- $\mu_{\bar{x}} = \mu$
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

MUESTREO ALEAOTRIO USANDO STATA

Usando el programa estadístico podemos tomar muestras aleatorias de maneras bien simples, utilizando algunos comandos. Vamos a partir utilizando una de las bases de datos que nos proporciona el programa, esta es la base de autos que hemos visto anteriormente.

```
. use auto
(1978 Automobile Data)
```

Luego generamos una variable, denominada “random”, con números seudo-aleatorios con distribución uniforme entre 0 y 1. Después se ordena de menor a mayor la variable, esto ordena a todas las observaciones dependientes de la variable creada anteriormente llamada “random”. Para finalizar se genera otra variable con la característica que

asigna valor 1 a la primera mitad de la base y el valor 2 a la segunda. A su vez podemos darnos cuenta de que el tamaño maestral es de 34 observaciones cada uno, siendo mayores a 30.

```
. generate random= runiform()

. sort random

. generate group = ceil(2 * _n/_N)
```

	weight	length	turn	displacement	gear_ratio	foreign	random	group
25	3,430	197	43	250	2.56	Domestic	.2648021	1
26	2,830	189	37	131	3.20	Foreign	.2655343	1
27	3,600	206	46	318	2.47	Domestic	.2769154	1
28	2,750	184	38	146	3.55	Foreign	.290384	1
29	3,310	198	42	231	2.93	Domestic	.3533874	1
30	3,330	201	44	225	3.23	Domestic	.3713805	1
31	3,700	214	42	231	2.73	Domestic	.3795409	1
32	1,930	155	35	89	3.78	Foreign	.3840067	1
33	3,740	220	46	225	2.94	Domestic	.4079702	1
34	4,720	230	48	400	2.47	Domestic	.4216726	1
35	4,290	204	45	350	2.24	Domestic	.4241557	1
36	2,070	174	36	97	3.70	Foreign	.4528397	1
37	4,840	233	51	400	2.47	Domestic	.4611429	1
38	1,760	149	34	91	3.30	Foreign	.4854506	2
39	3,690	212	43	250	2.56	Domestic	.5219247	2
40	2,160	172	36	97	3.74	Foreign	.541567	2
41	3,880	207	43	231	2.93	Domestic	.5552388	2
42	2,640	168	35	121	3.08	Domestic	.5578017	2
43	2,750	179	40	151	2.73	Domestic	.5644092	2
44	3,370	200	43	231	3.08	Domestic	.5773884	2
45	3,210	201	45	231	2.93	Domestic	.5823284	2
46	4,060	220	43	350	2.41	Domestic	.5948404	2
47	3,250	196	40	196	2.93	Domestic	.6047949	2
48	2,230	170	34	304	2.87	Domestic	.6184582	2
49	2,200	165	35	97	3.21	Foreign	.6267227	2
50	2,130	161	36	105	3.37	Foreign	.628377	2
51	3,350	173	40	258	2.53	Domestic	.6432207	2
52	4,060	221	48	302	2.75	Domestic	.6759487	2

El número de observaciones de la base es de 74 automóviles, y podemos darnos cuenta de que se dividieron en dos muestras de igual tamaño.

Para la siguiente parte veremos como la media poblacional se puede acercar mucho a la media de las medias muestrales. En este caso tomaremos la media de los precios de ambas muestras generadas.

```
. sum price if group==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |        37    6469.054    3182.542       3291      15906

. sum price if group==2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |        37    5861.459    2705.996       3299      13466
```

Podemos ver que la media de ambas muestras difiere en más 600 dólares, y que las desviaciones estándar también difieren. Y si recordamos, $\mu_{\bar{x}} = \mu$ la media de las medias muestrales es igual a la media poblacional. Comprobémoslo.

```
. sum price

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |        74    6165.257    2949.496       3291      15906
```

La media poblacional es de 6165,257 dolares, y si obtenemos la media de la media de ambas muestras, podemos darnos cuenta de que son iguales. Comprobando el teorema.

```
. display (6469.054+5861.459)/2
6165.2565
```

El comando display sirve para hacer operaciones matemáticas simple en STATA.