Tema 12 : Recogida de la información, Técnicas de muestreo. Errores de los muestreos.

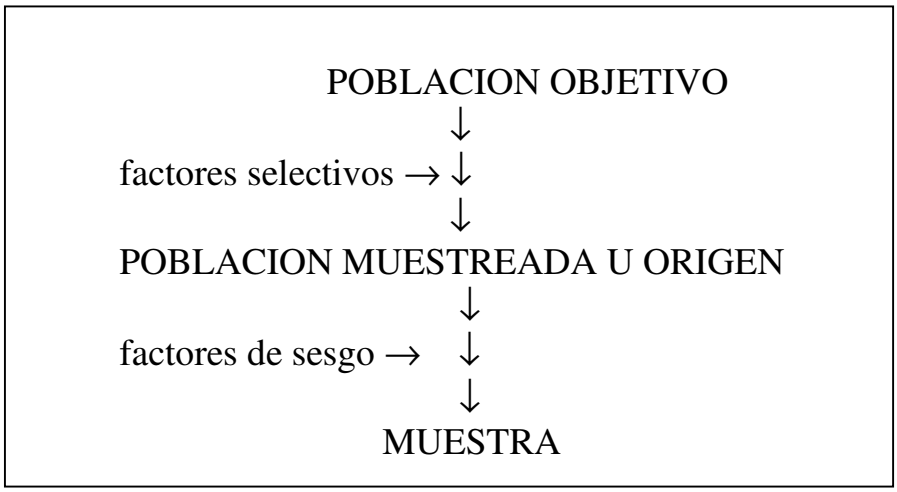
---Una muestra debe ser representativa

Ya vimos en el tema 1 que las muestras deben ser representativas de la población de la que proceden y que la mejor garantía de conseguirlo es un tamaño adecuado de la muestra y la elección al azar de los individuos, es decir, **una muestra aleatoria de tamaño adecuado**. Es un punto crucial.

Esta representatividad puede verse afectada, además de por un tamaño insuficiente, por los llamados **factores de sesgo**, como deficiencias de la aleatoriedad (¿tienen realmente todos los individuos la misma probabilidad de salir elegidos?), errores muestrales extremos y errores personales e instrumentales.

---Origen de la muestra

La población de la que procede la muestra es la **población muestreada o población origen**, que idealmente debe coincidir con **la población objetivo** del estudio, lo que no siempre ocurre por la existencia de **factores selectivos** más o menos intensos. Es posible que el investigador no se de cuenta de esta situación y pueda llegar, honestamente, a conclusiones erróneas.



Siempre hay que comprobar que la población muestreada es realmente la población objetivo

Ejemplo: en los años 50 se realizó en Barcelona un estudio epidemiológico muy importante sobre tuberculosis, que estaba entonces muy extendida. Los datos se obtuvieron de una muestra tomada del Dispensario Antituberculoso. Los resultados se presentaron como reflejo del estado de la tuberculosis en la ciudad de Barcelona. Pronto surgieron críticas al estudio. ¿La muestra era realmente representativa de los tuberculosos catalanes?. ¿O sólo de los pobres?. Los más pudientes y algunos más pobres que hicieron un esfuerzo económico eran atendidos en consultas y clínicas privadas. Y era de sobra sabido la influencia del estado social en la evolución de esta enfermedad. Muy probablemente la muestra estaba contaminada por un factor selectivo: la situación económica.

---Tamaño de la muestra

Depende fundamentalmente de 4 factores: 1) tamaño de la población, 2) dispersión o variabilidad de los individuos de la población, 3) margen de error que estemos dispuestos a admitir y 4) nivel de significación o confianza elegidos.

Para calcular el tamaño muestral se dispone de fórmulas, que nos orientan sobre el mismo. Siempre se cogen más individuos de los calculados, para compensar posibles fallos. También se dispone de tablas, sobre todo para estimaciones de porcentajes, que no veremos. En la práctica a partir de un tamaño poblacional de 10.000 se pueden usar las fórmulas de "población infinita", que son más sencillas. Dicho de otra forma: a efectos prácticos una población se puede considerar como infinita a partir de un tamaño de 10.000 (hay autores que elevan este tamaño a 60.000).
En las fórmulas aparece c^2. Es el valor de c de la DN tipificada que corresponde al nivel de significación elegido. El nivel de significación, cuyo símbolo es α , expresa el riesgo estadístico de error, el llamado "error tipo 1". Por consenso se consideran significativos los valores de α de 0'05 para abajo. Los programas estadísticos de ordenador calculan este riesgo exactamente. Para cálculos manuales se toman tradicionalmente tres puntos de referencia para α : 0'05 (ó 5%) , 0'01 (ò 1%) y 0'001 (ó 1%º)., que se corresponden con valores de c de 1'96 , 2'53 y 3'30 respectivamente. Si no se exige o desea otro nivel, se toma de oficio el de 0'05 y por tanto c = 1'96.

---Fórmulas

1) para una estimación

	Población finita	Población infinita
media	$N = \frac{c^2 * Np * s^2}{Np * k^2 + c^2 * s^2}$	$N = \left(\frac{c * s}{k}\right)^2$
p ó %	$N = \frac{c^2 * Np * p * q}{(Np-1) * k^2 + c^2 * p * q}$	$N = \frac{c^2 pq}{k^2}$

2) para contraste de variables (N por muestra)
- de medias : $N = 13 * s^2 / d^2$
- de 2 proporciones o porcentajes : $N = 6'5(p_1q_1+p_2q_2)/d^2$

N es el tamaño muestral, N_p el tamaño de la población, k el error máximo admitido, s^2 la varianza de la población, real o estimada a partir de un estudio piloto o incluso de una forma más simple por la fórmula $s^2 \approx (R/4)^2$, siendo R el Recorrido. La "c" es el valor de referencia de la DN tipificada correspondiente al nivel de significación elegido. La "d" es la diferencia mínima que queremos probar entre los porcentajes o medias contrastadas.
En el caso de estimaciones p y q toman su valor real en la población si se conoce; si no, se les da el valor más desfavorable y que conduce a un tamaño mayor: 0'5 a cada una. En el caso de contraste de muestras se procede de la misma forma: dar a cada p y q su valor real, si es conocido y si no, darles el valor de 0'5.
Si los datos son apareados o se trata de una prueba de conformidad, N se divide por 2.

---Recogida de los datos

Los datos se recogen por
1) observación, directa o con aparatos.
2) interrogatorio, que puede ser directo (entrevista) o indirecto (cuestionario). Es típico de encuestas. Presupone preguntas neutrales y por parte del interrogado buena memoria y buena fe.

---Métodos de obtención de muestras al azar

Hay diversos tipos de muestras aleatorias:

1. **Muestras de azar simple o aleatoria elemental**. Presupone lista de todos los individuos, numerados. La unidad muestral es el individuo. Los individuos se eligen por sorteo o utilizando una tabla de números al azar (ver una muy sencilla al final del tema).
2. **Muestras sistemáticas**. Es una variante de la anterior con un procedimiento de elección simplificado. Hay que calcular el coeficiente de elevación (Tamaño de la población dividido por el tamaño de la muestra). Luego se elige al azar un número menor que dicho coeficiente, que será el primer individuo de la muestra. A ese número se la sumando el coeficiente de elevación y así nos va dando los individuos hasta alcanzar el tamaño previsto de la muestra. Por ejemplo: tamaño de la población 1000; tamaño de la muestra 100 ; coeficiente de elevación 1000/100 = 10. Se elige al azar un número menor de 10 y sale el 6. La muestra la compondrán los individuos de la lista cuyos números sean el 6 , 16 , 26 , 36 , 46 , ... hasta el 996.
3. **Muestras estratificadas**. Se hacen estratos de la población, que son grupos homogéneos de individuos, con poca variación intragrupo. Por ejemplo, hombres y mujeres, grupos de edad, grupos raciales, regiones de un país, factores de riesgo. etc. Fijados los estratos se eligen de forma proporcional y al azar los individuos que formarán la muestra. Aquí también la unidad muestral es el individuo y se necesita un listado de la población. son muy utilizadas en investigaciones clínicas.
4. **Muestras de conglomerados.** Los conglomerados son grupos naturales y heterogéneos de individuos. De entrada no se conocen los individuos, sino los conglomerados, que son la unidad muestral. Por ejemplo, tenemos una lista de escuelas o de hospitales (que son los conglomerados) ; se eligen al azar los que hagan falta y una vez en ellos se eligen al azar los individuos necesarios.
5. **Muestras combinadas**. Es una mezcla de estratos y conglomerados.

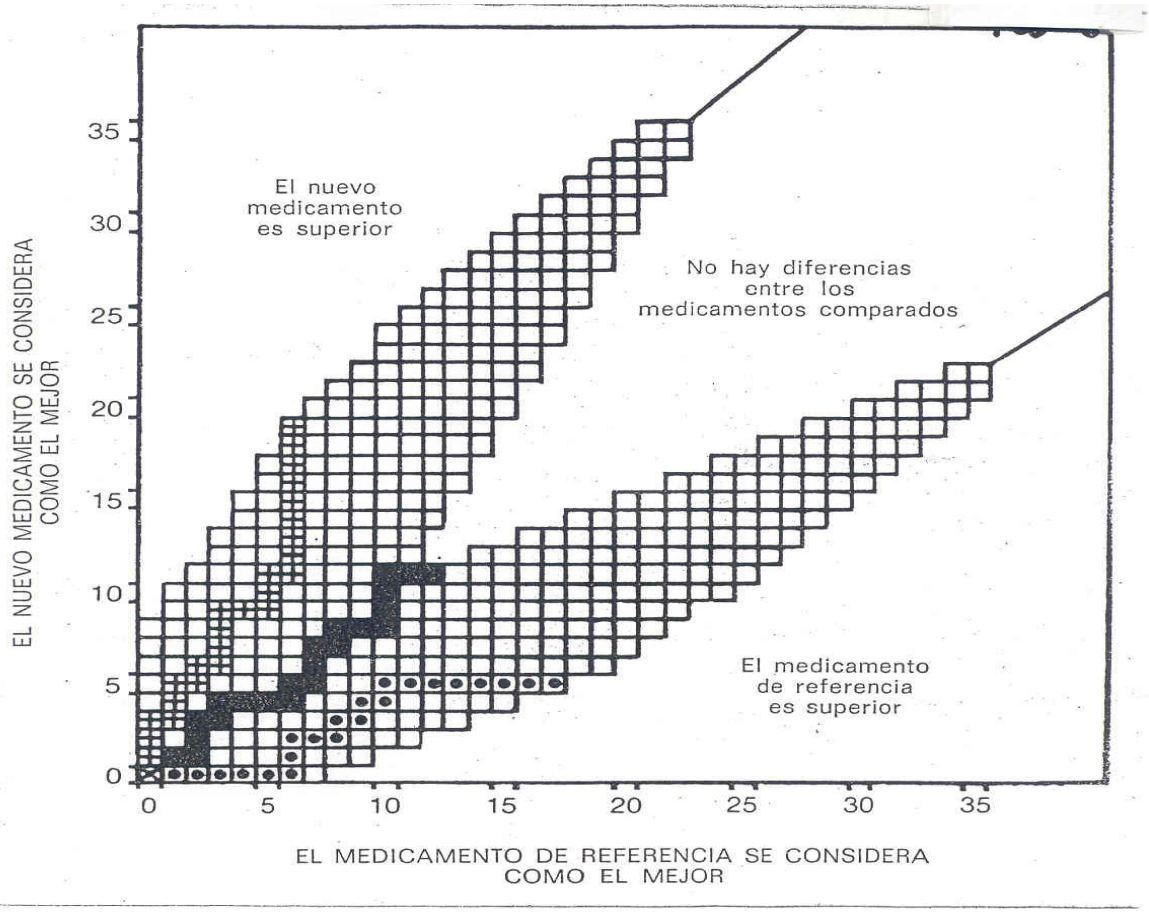
Ejemplos: Deseamos estudiar el nivel de plomo en la sangre de los niños de 3º de ESO en la región R. Sabemos que son 4000 niños, que acuden a 200 escuelas y cada clase tiene 20 alumnos. Tenemos un listado de los 4000 alumnos y un listado de las escuelas. 40 escuelas están en poblaciones grandes, 120 en medianas y 80 en pequeñas- Supongamos que necesitamos una muestra de tamaño 400. ¿Cómo obtenerla?

1. *Muestra al azar*. De la lista de los 4000 niños se sacan al azar (sorteo o por la tabla de números al azar) los 400 que se necesitan.
2. *Muestra sistemática*. Necesitamos también la lista de los 4000 alumnos. Coeficiente de elevación : 4000/400=10. Se elige al azar un número <10 y sale el 3. Por tanto saldrán elegidos para formar parte de la muestra los alumnos con los números 3, 13, 23, 33, 43,.....y así hasta el 3993.
3. *Muestra estratificada*. Hay indicios de que el tamaño de las ciudades y pueblos puede ser de importancia en el estudio. Elegimos 3 estratos representativos y les asignamos un porcentaje (fruto del estudio de la situación): ciudades o pueblos grandes, de los que sacaremos el 20% de la muestra; medianos con el 60% y pequeños con el 20%. Esto equivale a tomar 80 alumnos del estrato grande, 240 del mediano y 80 del pequeño. Su elección se hace por el método 1 ó el 2.
4. *Muestra de conglomerados*. Aquí no hay lista de alumnos, sólo de escuelas. Se eligen al azar 20 escuelas y se toman los 20 alumnos de cada una de ellas.
5. *Muestra combinada*. Une 3 y 4. Agrupamos las escuelas (que son los conglomerados) por estratos de tamaño poblacional (40, 120, 40) y se eligen el 10% de cada estrato, o sea 20 , 12 y 4 escuelas respectivamente. tomando los 20 alumnos de cada una de estas escuelas tenemos los 400 necesarios.

---Otras formas de obtener muestras

En investigaciones clínicas se utiliza con frecuencia la llamada **asignación al azar**, que evita elecciones subjetivas. Por ejemplo, en estudios en que cada paciente nuevo debe ser asignado a un grupo de tratamiento distinto; se dispone de una serie de sobres cerrados en los que está el tratamiento a recibir y cuando llega el paciente se coge un sobre y se le aplica el tratamiento que indica.

En el **análisis secuencial** no es necesario siquiera conocer previamente el tamaño muestral. Los datos se comparan por parejas, uno del grupo que podemos llamar A y otro del grupo B. Hay 3 resultados posibles: A es mejor, B es mejor y ninguno es mejor (0). Se utiliza una gráfica en V, como la que sigue, que sirve para $\alpha = 0{,}05$. Se van rellenado casillas con los datos que vamos obteniendo. Se empieza por el vértice de la V. Si A es mejor se rellena la casilla superior, si es mejor B la casilla de la derecha y si no hay diferencias no se rellena ninguna casilla. Llega un momento en que nos salimos del gráfico por algún sitio. Por arriba si A es mejor, por abajo si B es mejor y por el centro si no hay diferencias.



Supongamos que queremos ver si un nuevo medicamento (A) es superior al que actualmente se utiliza (B) en el tratamiento de la migraña. Cada paciente recibe en un orden prefijado al azar un medicamento, en una ocasión A y B en otra. Luego informa de cual ha sido más eficaz. Se obtiene lo siguiente:

paciente: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 ..
mejor: A A B B A A 0 0 A A A B A A 0 B A A A A B A B A 0 0..

paciente: ... 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
mejor: .... B A A A B 0 0 B A A 0 B A A A 0 A 0 A

En el paciente 45 nos salimos de la V por arriba. Por tanto A es mejor.

---Errores de los muestreos

I. PROPIOS DE LA MUESTRA
   i. muestra no representativa
   ii. *ERROR MUESTRAL*, que es inevitable y se debe a la variabilidad natural. Se puede medir hasta donde puede llegar. Lo veremos enseguida.

II. EXTRAÑOS A LA MUESTRA
   i. personales (del observador), que dependen de su preparación, estado psicofísico, ambiente, etc. Hay variaciones intraobservador e interobservador.
   ii. sistemáticos (del método de medida). Dependen de su sensibilidad, precisión y exactitud.

Sesgos de recuerdo ("recall bias")

Los pacientes son reiteradamente preguntados por la existencia de factores de riesgo y los suelen recordar muy bien. Cosa que no ocurre con los controles en un estudio caso-control.

---Disminución de los errores

--los del observador, mediante una buena preparación, condiciones adecuadas de trabajo y utilización de controles de calidad.
--los del método, mediante aparatos de calidad, buen mantenimiento, controles de calidad, buenos cuestionarios.

---ERROR MUESTRAL (E)

Si sacamos de una población diversas muestras y calculamos uno o más parámetros, veremos que no obtenemos exactamente los mismos resultados. Esto se debe a la existencia de un error, el error muestral, que es inevitable, pero que puede ser valorado, ya que los parámetros obtenidos de muestras repetidas de una misma población (>30) siguen la ley normal aunque la población de origen no sea normal. Y por tanto tienen su margen de variación, cuyo máximo puede ser medido. Es el error muestral.

E = c*e ó t*e , siendo e el llamado error estándar. Si la muestra es <30 se utiliza t, la t de Student, y si es grande (≥30) la c de la DN.

---ERROR ESTANDAR (e)

Es la desviación estándar de la distribución de los parámetros estadísticos muestrales (media, %, etc.) cuando se extraen repetidas muestras. No se debe confundir con la desviación estándar de una muestra (s). Se han encontrado fórmulas con las que a partir de una sola muestra se puede calcular ya el error estándar:

$$\textbf{para una media: } e = \frac{s}{\sqrt{N}}$$

$$\textbf{para un porcentaje: } e = \sqrt{\frac{pq}{N}}$$

TABLA VII

TABLA DE NUMEROS ALEATORIOS

10	09	73	25	33	76	52	01	35	86	34	67	35	48	76	80	95	90	91	17	39	29	27	49	45
37	54	20	48	05	64	89	47	42	96	24	80	52	40	37	20	63	61	04	02	00	82	29	16	65
08	42	26	89	53	19	64	50	93	03	23	20	90	25	60	15	95	33	47	64	35	08	03	36	06
99	01	90	25	29	09	37	67	07	15	38	31	13	11	65	88	67	67	43	97	04	43	62	76	59
12	80	79	99	70	80	15	73	61	47	64	03	23	66	53	98	95	11	68	77	12	17	17	68	33
66	06	57	47	17	34	07	27	68	50	36	69	73	61	70	65	81	33	98	85	11	19	92	91	70
31	06	01	08	05	45	57	18	24	06	35	30	34	26	14	86	79	90	74	39	23	40	30	97	32
85	26	97	76	02	02	05	16	56	92	68	66	57	48	18	73	05	38	52	47	18	62	38	85	79
63	57	33	21	35	05	32	54	70	48	90	55	35	75	48	28	46	82	87	09	83	49	12	56	24
73	79	64	57	53	03	52	96	47	78	35	80	83	42	82	60	93	52	03	44	35	27	38	84	35
98	52	01	77	67	14	90	56	86	07	22	10	94	05	58	60	97	09	34	33	50	50	07	39	98
11	80	50	54	31	39	80	82	77	32	50	72	56	82	48	29	40	52	42	01	52	77	56	78	51
83	45	29	96	34	06	28	89	80	83	13	74	67	00	78	18	47	54	06	10	68	71	17	78	17
88	68	54	02	00	86	50	75	84	01	36	76	66	79	51	90	36	47	64	93	29	60	91	10	62
99	59	46	73	48	87	51	76	49	69	91	82	60	89	28	93	78	56	13	68	23	47	83	41	13
65	48	11	76	74	17	46	85	09	50	58	04	77	69	74	73	03	95	71	86	40	21	81	65	44
80	12	43	56	35	17	72	70	80	15	45	31	82	23	74	21	11	57	82	53	14	38	55	37	63
74	35	09	98	17	77	40	27	72	14	43	23	60	02	10	45	52	16	42	37	96	28	60	26	55
69	91	62	68	03	66	25	22	91	48	36	93	68	72	03	76	62	11	39	90	94	40	05	64	18
09	89	32	05	05	14	22	56	85	14	46	42	75	67	88	96	29	77	88	22	54	38	21	45	98
91	49	91	45	23	68	47	92	76	86	46	16	28	35	54	94	75	08	99	23	37	08	92	00	48
80	33	69	45	98	26	94	03	68	58	70	29	73	41	35	53	14	03	33	40	42	05	08	23	41
44	10	48	19	49	85	15	74	79	54	32	97	92	65	75	57	60	04	08	81	22	22	20	64	13
12	55	07	37	42	11	10	00	20	40	12	86	07	46	97	96	64	48	94	39	28	70	72	58	15
63	60	64	93	29	16	50	53	44	84	40	21	95	25	63	43	65	17	70	82	07	20	73	17	90
61	19	69	04	46	26	45	74	77	74	51	92	43	37	29	65	39	45	95	93	42	58	26	05	27
15	47	44	52	66	95	27	07	99	53	59	36	78	38	48	82	39	61	01	18	33	21	15	94	66
94	55	72	85	73	67	89	75	43	87	54	62	24	44	31	91	19	04	25	92	92	92	74	59	73
42	48	11	62	13	97	34	40	87	21	16	86	84	87	67	03	07	11	20	59	25	70	14	66	70
23	52	37	83	17	73	20	88	98	37	68	93	59	14	16	26	25	22	96	63	05	52	28	25	62
04	49	35	24	94	75	24	63	38	24	45	86	25	10	25	61	96	27	93	35	65	33	71	24	72
00	54	99	76	54	64	05	18	81	59	96	11	96	38	96	54	69	28	23	91	23	28	72	95	29
35	96	31	53	07	26	89	80	93	54	33	35	13	54	62	77	97	45	00	24	90	10	33	93	33
59	80	80	83	91	45	42	72	68	42	83	60	94	97	00	13	02	12	48	92	78	56	52	01	06
46	05	88	52	36	01	39	09	22	86	77	28	14	40	77	93	91	08	36	47	70	61	74	29	41
32	17	90	05	97	87	37	92	52	41	05	56	70	70	07	86	74	31	71	57	85	39	41	18	38
69	23	46	14	06	20	11	74	52	04	15	95	66	00	00	18	74	39	24	23	97	11	89	63	38
19	56	54	14	30	01	75	87	53	79	40	41	92	15	85	66	67	43	68	06	84	96	28	52	07
45	15	51	49	38	19	47	60	72	46	43	66	79	45	43	59	04	79	00	33	20	82	66	95	41
94	86	43	19	94	36	16	81	08	51	34	88	88	15	53	01	54	03	54	56	05	01	45	11	76
98	08	62	48	26	45	24	02	84	04	44	99	90	88	96	39	09	47	34	07	35	44	13	18	80
33	18	51	62	32	41	94	15	09	49	89	43	54	85	81	88	69	54	19	94	37	54	87	30	43
80	95	10	04	06	96	38	27	07	74	20	15	12	33	87	25	01	62	52	98	94	62	46	11	71
79	75	24	91	40	71	96	12	82	96	69	86	10	25	91	74	85	22	05	39	00	38	75	95	79
18	63	33	25	37	98	14	50	65	71	31	01	02	46	74	05	45	56	14	27	77	93	89	19	36
74	02	94	39	02	77	55	73	22	70	97	79	01	71	19	52	52	75	80	21	80	81	45	17	48
54	17	84	56	11	80	99	33	71	43	05	33	51	29	69	56	12	71	92	55	36	04	09	03	24
11	66	44	98	83	52	07	98	48	27	59	38	17	15	39	09	97	33	34	40	88	46	12	33	56
48	32	47	79	28	31	24	96	47	10	02	29	53	68	70	32	30	75	75	46	15	02	00	99	94
69	07	49	41	38	87	63	79	19	76	35	58	40	44	01	10	51	82	16	15	01	84	87	69	38