Estadística aplicada a las Ciencias Sociales

Tamaño necesario de la muestra: ¿Cuántos sujetos necesitamos?

©Pedro Morales Vallejo
• Universidad Pontificia Comillas • Madrid • Facultad de Humanidades
(Última revisión, 13 de Diciembre, 2012).
Disponible en http://www.upcomillas.es/personal/peter/investigacion/Tama%F1oMuestra.pdf

Índice

1. Tipos de muestras

En muchas investigaciones no se pretende en principio (al menos como objetivo prioritario) extrapolar los resultados. Cuando el interés del investigador tiene como objetivo analizar una muestra concreta (verificar si una innovación didáctica funciona en una clase, estudiar las actitudes de nuestros alumnos, etc.), el tamaño de la muestra es el tamaño del grupo objeto de estudio.

Cuando se trata de *extrapolar* los resultados a la población representada por una muestra hay que tratar de dos temas: el tipo de muestra y el número de sujetos.

Cómo se obtiene la muestra tiene que ver con	la *representatividad* de la muestra
El *tamaño de la muestra* tiene que ver con	los *márgenes de error* al extrapolar de la muestra a la población.

Cuando se trata de recoger unos datos de una muestra (por ejemplo en un *survey* o encuesta sociológica, de opiniones, etc.) para *extrapolar los resultados a la población,* el *número de sujetos* no es el *único* dato que hay que tener en cuenta, y esto hay que tenerlo presente *desde el principio.* La muestra, cualquiera que sea su magnitud, debe ser *representativa* de la población a la que se van a extrapolar los resultados. Debemos recordar que los límites o características de la *población* los determina y define el que investiga (los alumnos de una facultad o de universidad, o de todo el país, etc.). En cualquier caso hay que explicar cómo se hizo este muestreo y describir bien la muestra para poder valorar esta representatividad.

Las muestras (y los métodos para obtenerlas) podemos dividirlas en dos grandes categorías según sean o no sean muestras *probabilísticas* o *aleatorias* (ambos términos, *probabilísticas* o *aleatorias*, expresan lo mismo[1].

1.1. Muestras *probabilísticas* o *aleatorias*

Una *muestra aleatoria* o *probabilística* es aquella en la que todos los sujetos de la población han tenido la misma probabilidad de ser escogidos. Son en principio los tipos de muestra *más profesionales*.

¿Por qué es importante el muestreo aleatorio? Las muestras aleatorias aseguran o garantizan mejor el poder extrapolar los resultados. En una muestra aleatoria tenemos más seguridad de que se encuentran representadas las características importantes de la población en la proporción que les corresponde. Si el 20% de la población tiene la característica A (una determinada edad, una determinada situación económica, etc.) podemos esperar que en la muestra también habrá en torno a un 20% con esa característica.

Si la muestra no es aleatoria (*no probabilística*) puede suceder que esté *sesgada* y que por lo tanto no sea representativa de la población general porque predominan más unos determinados tipos de sujetos que otros. Por ejemplo, si hacemos una pregunta a los conductores que se paran ante un semáforo, prescindimos de los que utilizan otro medio de transporte, y si hacemos la pregunta a la salida de una estación de metro, prescindimos de los que tienen y utilizan coche particular, etc. Si la población es la de una *universidad,* según el día y método seguido, podemos

[1] En Internet StatPac Inc (2003) ofrece una buena información sobre *surveys*, (*Questionnaires & Survey Design*) y en *Sampling Mehods* (menú) una clara explicación de los diversos tipos de muestras. Más información sobre muestras, asequible en Internet, son Stark (2003, chapter 16) y Trochim en *Sampling* (last revised 2006). Los diversos tipos de muestreo aleatorio y cómo llevarlos a cabo pueden verse en muchos textos (como Hernández Sampieri, Fernández Collado y Baptista Lucio, 2008) y en monografías específicas (como Rodríguez Osuna, 1991, 1993). En Internet hay programas sencillos que generan *tablas de números aleatorios* y *calculan el tamaño de la muestra* y los márgenes de error (*intervalos de confianza*) para poblaciones de distinto tamaño; pueden verse algunas direcciones útiles en el anexo.

tener infra-representados a los alumnos de los cursos que asisten menos a clase, o tener representados en exceso a los alumnos que llegan muy pronto por la mañana (y este *llegar antes* o *llegar tarde* puede estar asociado con otras variables), etc. Características aparentemente irrelevantes pueden estar *relacionadas* precisamente con el objeto de la investigación (con la *variable dependiente).*

Sobre el *cómo* hacer un muestreo aleatorio no tratamos aquí de manera específica, hay diversas maneras de hacerlo y por lo general, y para poblaciones grandes, este tema suele tratarse en textos de *sociología*. En principio podemos concretar tres procedimientos para hacer un muestreo aleatorio.

a) Muestreo *aleatorio simple*.

Podemos concebirlo como un sorteo, una lotería. Para poblaciones *pequeñas* (tan pequeñas como los alumnos de una clase o de un curso) hay métodos sencillos (como las tablas de números *aleatorios*, e incluso estos *números* están programados en muchas calculadoras)[2].

b) Muestreo *sistemático*.

Es también sencillo y cómodo para poblaciones pequeñas; se escoge un número al azar que se utiliza como intervalo; si sale por ejemplo el 9, se escoge de una lista *cada noveno sujeto…*

c) Muestreo *estratificado.*

Es un tipo de muestreo muy recomendable, sobre todo para poblaciones grandes: se divide la población en *estratos* o segmentos según algunas características importantes para lo que se desea investigar (como pueden ser: sexo, curso, edad, tipo de vivienda…) y se procura que en la muestra esté representado cada estrato en la proporción que le corresponda. Dentro de cada estrato los sujetos se escogen *aleatoriamente*. Los estratos se establecen en función de características importantes por su interés específico descriptivo y sobre todo porque, si se desea extrapolar a toda la población, pueden tener que ver con la variable dependiente.

1.2. Muestras no *probabilísticas*

Entra las muestras *no probabilísticas* tenemos las siguientes:

a) Muestras *de conveniencia*

Como ya sugiere el mismo término se trata de una muestra disponible. Puede ser útil en estudios preliminares (por ejemplo para depurar un instrumento) pero discutible para extrapolar los datos a la población general; en cualquier caso habría que describirla bien y hacer una extrapolación cautelosa a la población que *pueda estar* representada por esa muestra. Por lo general estas muestras se utilizan para hacer estudios específicos sobre las mismas muestras y en numerosos estudios experimentales con pocos sujetos. Se denominan de *juicio prudencial*, o términos parecidos, cuando se estima y se razona que la muestra es representativa de una determinada población.

b) Muestras *bola de nieve*

Se denominan así cuando la muestra se obtiene yendo de unos sujetos a otros; útil cuando la característica de la población es poco común o de acceso no fácil y unos sujetos informan sobre otros que participan de la misma característica.

---

[2] Pueden verse en el anexo algunos recursos disponibles en Internet. En el SPSS se puede hace una selección aleatoria de casos; en el menú *datos → seleccionar datos → selección aleatoria de casos* y se indica el tanto por ciento de casos aleatorios que se quieren seleccionar.

c) Muestreo *por cuotas*.

Es lo mismo que el muestreo estratificado, pero *sin elección aleatoria* dentro de cada estrato; al menos se garantiza que todas las submuestras están representadas en su debida proporción. En la práctica, y con el objetivo expreso de extrapolar los datos a toda la población, puede ser un buen sistema.

2. Número de sujetos y finalidad pretendida

Cuando nos preguntamos *cuántos sujetos necesitamos*… hay que añadir, necesitamos *¿para qué?* Porque puede haber varias finalidades.

Lo que con más frecuencia se encuentra en los textos es posiblemente las normas y fórmulas para determinar el tamaño de la muestra cuando se quieren extrapolar los resultados a la población (encuestas, sondeos pre-electorales, etc.); éste es el primer punto que vamos a tratar. Pero en la investigación y experimentación en psicología, educación o sociología hay con frecuencia otras finalidades: ¿Cuántos sujetos necesitamos para *construir y analizar un instrumento de medición* (un test, una escala de actitudes), o para llevar a cabo un *análisis factorial* o un *análisis de varianza* en un diseño experimental…?

En los textos de investigación experimental vemos con frecuencia ejemplos de investigaciones con muy pocos sujetos, quizás sobre todo cuando se utilizan diseños experimentales en sentido propio y el análisis de varianza es el método de análisis ¿Qué podemos decir sobre el tamaño de la muestra en estos casos?

Vamos a dividir esta exposición en varios apartados de extensión muy desigual:

Número de sujetos necesario:

a) Para *extrapolar los resultados a una población mayo*r (el caso típico de los sondeos de opinión);
b) Para *construir un instrumento* de medición (test o escala)
c) Para llevar a cabo un *estudio experimental* (contrate de medias, análisis de varianza)
d) Para hacer un *análisis correlacional*

3. Número de sujetos necesarios en la muestra para extrapolar los datos a la población

Aunque podemos acudir a los programas de Internet (en el Anexo) es útil entender las fórmulas y las variables que intervienen para determinar el tamaño de la muestra.

3.1. Variables de las que depende el tamaño de la muestra

Suponiendo que la muestra es la adecuada, el *tamaño necesario* de la muestra para poder extrapolar los resultados a la población depende básicamente de tres variables. El por qué estas variables inciden en el tamaño de la muestra es fácil comprenderlo de manera *intuitiva,* al margen de la *traducción* de estas variables a valores estadísticos.

1º El *nivel de confianza* o riesgo que aceptamos de equivocarnos al presentar nuestros resultados: lo que deseamos es que en otras muestras semejantes los resultados sean los mismos o muy parecidos. También podemos denominarlo *grado o nivel de seguridad*. El nivel de confianza habitual es de .05 (α = .05).

El nivel de confianza va a entrar en la fórmula para determinar el número de sujetos con un valor de *zeta,* que en la distribución normal está asociado a una determinada probabilidad de ocurrencia.

2º La *varianza* (o *diversidad de opiniones*…) *estimada en la población*. Esta diversidad en la población es la diversidad *estimada;* si la conociéramos (cuántos van a decir que *sí* y cuántos van a decir que *no*) en primer lugar no necesitaríamos hacer la encuesta.

Si sabemos de antemano que *todos* piensan lo mismo (aunque no sabemos *qué* piensan y por eso lo preguntamos), nos bastará preguntar a un solo sujeto, pero si las opiniones pueden ser muy distintas nos harán falta más sujetos. Por ejemplo, si aterrizamos en un país desconocido donde todos hablan el mismo idioma y no sabemos cuál es…basta con preguntárselo a una sola persona de ese país. Si en una *agrupación* de sujetos que viven juntos, todos se levantan a la misma hora (un cuartel, un convento), y queremos saber a qué hora se levantan por la mañana, basta con preguntárselo a una sola persona. A mayor diversidad esperada, o al menos posible, en las opiniones o posibles respuestas en la población hará falta un mayor número de sujetos en la muestra.

3º El *margen de error* que estamos dispuestos a aceptar.

Si por ejemplo el 20% de la muestra está de acuerdo con una proposición (o dice que va votar a un determinado candidato o que prefiere un determinado producto) eso no significa que el 20% exacto de la población vaya a responder lo mismo, puede ser el 22% o el 18%… necesitaremos muestras mayores si queremos que el *margen de error* o de *oscilación* de muestra a muestra de los resultados sea muy pequeño (el resultado exacto lo tendríamos si respondiera el 100% y la muestra coincidiera con la población).

Esto puede ser más o menos importante según la situación; el margen de error en sondeos pre-electorales es, por ejemplo, muy importante y este margen de error suele ponerse en torno a un 3%.

3.2. Cómo calcular el tamaño de la muestra

Vamos a distinguir entre poblaciones *infinitas* (de tamaño muy grande, indefinido, cuyo tamaño exacto podemos desconocer) y poblaciones *finitas* (tamaño más reducido y que conocemos).

En la práctica disponemos de tablas en numerosos textos y de programas de Internet (ver anexo) en los que podemos ver con toda facilidad el tamaño necesario de la muestra en función del tamaño de la población, del nivel de confianza y del margen de error tolerado; es muy útil sin embargo *entender* las fórmulas (que por otra parte son muy sencillas) porque nos hacen ver cómo se relacionan las variables que condicionan el tamaño de la muestra.

3.2.1. En el caso de poblaciones *infinitas* (tamaño *grande,* indefinido…)

3.2.1.1. Fórmulas para determinar el tamaño de la muestra

Para extrapolar a poblaciones muy grandes utilizamos la fórmula [1] para obtener el tamaño de la muestra:

$$N = \frac{z^2pq}{e^2} \qquad [1]$$

Explicamos los símbolos de la fórmula, que corresponden a las variables indicadas en el apartado anterior.

z = Valor de z correspondiente al *nivel de confianza;*

Un nivel de confianza del 95% (también lo expresamos así: α = .05) corresponde a z = 1.96 *sigmas* o *errores típicos;* z = 2 *(dos sigmas)* corresponde a un 95.5% (aproximadamente, α = .045).

Con z = 2.57 el nivel de confianza sube al 99% (nos equivocaríamos una vez de cada 100), pero como aumenta el numerador aumenta el cociente… que es N, y harán falta más sujetos (y más trabajo y más gastos).

pq = ***Varianza de la población***

Como la varianza de la población la desconocemos, ponemos la *varianza mayor posible* porque a mayor varianza hará falta una muestra mayor.

Recordamos el significado de los símbolos:

p = proporción de respuestas en una categoría (*síes,* respuestas correctas, *unos* en la codificación usual, etc.)

q = proporción de repuestas en la otra categoría (*noes, ceros* en la codificación usual).

La varianza en los ítems dicotómicos (dos respuestas que se excluyen mutuamente) es igual a *pq* y la varianza mayor (*la mayor diversidad de respuestas*) se da cuando p = q = .50 (la mitad de los sujetos responde *sí* y la otra mitad responde *no*) por lo que en esta fórmula [1] pq es siempre igual a (.50)(.50) = .25 (es una constante).

En las *fichas técnica*s de las *encuestas sociológicas* que se publican en la prensa es normal indicar que la muestra ha sido escogida partiendo de la hipótesis de que p = q = .50 (a veces se expresa de otras maneras: P = Q = 50, ó p/q = 50, etc.).

El suponer que p = q quiere decir que para escoger la muestra nos ponemos en la hipótesis de que en la población hay la máxima diversidad posible: un 50% va a decir que *sí* y otro 50% va a decir que *no,* de esta manera, y por lo que respecta a la varianza de la población, no corremos riesgos de quedarnos cortos en el número de sujetos. Este valor de pq (= .25) es válido (válido para calcular el tamaño de la muestra) aun cuando las preguntas no sean dicotómicas.

e = ***Error muestral***

Lo representamos con la letra **e** (no es el único símbolo que se utiliza) que significa *error* o *desviación* posible cuando *extrapolemos* los resultados. Es el *margen de error* que aceptamos.

Si el margen de error es 3.16%, en la fórmula pondremos e = 0.0316. Si *dice que sí* un 64.3% en la *muestra,* entendemos que *dice que sí* en la población *entre* un (64.3 - 3.16)% y un (64.3 + 3.16)%. Cuanto más bajo sea este error probable, que es el denominador, aumenta la precisión pero también subirá obviamente el cociente: harán falta más sujetos (y *sube el precio,* etc.).

Observando la fórmula vemos que, efectivamente, el tamaño de la muestra (cociente o resultado de la fórmula) será mayor según sea mayor el nivel de confianza y la varianza esperada en la población (numerador en la fórmula) y según sea menor el margen de error que estamos dispuestos a admitir (denominador en la fórmula).

Por ejemplo ¿qué muestra necesitaremos con un nivel de confianza del 95% (o α = .05), al que corresponde z = 1.96, y admitiendo un margen de error del 5% o del 2%? Ya sabemos que pq = .25.

error aceptado: 5% y α = .05 (z = 1.96)

$$N = \frac{(1.96^2)(.25)}{.05^2} = 384$$

error aceptado: 2% y α = .05 (z = 1.96)

$$N = \frac{(1.96^2)(.25)}{.02^2} = 2401$$

El nivel de confianza nunca suele ser menor de α = .05 (nos equivocaríamos en nuestra predicción o generalización 5 veces de cada 100), y como pq es igual a .25, el número de sujetos va a depender de hecho del margen de error que estamos dispuestos a tolerar.

Con frecuencia se opera con un *nivel de confianza* del 95.5%; la razón práctica está en que este *nivel de confianza* corresponde a z = 2 *(dos sigmas);* y el numerador de la fórmula queda muy simplificado:

$$(z^2)\,(pq) = (2^2)(.25) = 1, \text{ con lo que} \qquad N = \frac{1}{e^2} \qquad [2]$$

El tamaño de la muestra no varía mucho; en el ejemplo anterior (error tolerado del 5% del 2%) necesitaríamos 400 y 2500 sujetos en vez de 384 y 2401.

Resumiendo lo dicho hasta ahora, el tamaño de la muestra (valor de N) aumentará:

1. Si aumenta nuestro *nivel de confianza* (de seguridad) que requiere un valor mayor de z,
2. Si disminuye el error muestral (*e*).

Es decir, si queremos *mucha seguridad* y *poco margen de error* hará falta un N mayor,

En la tabla 1 podemos ver el tamaño de la muestra (N) para diversos valores de *e (márgenes de error)* y los dos *niveles de confianza* más usuales (z = 1.96 o un 95% de nivel de confianza y z = 2.57 o un 99% de nivel de confianza o probabilidades de no equivocarnos):

		e = .05	e = .04	e = .03	e = .02	e = .01
α = .05	z = 1.96	384	600	1067	2401	9604
α = .01	z = 2.57	660	1032	1835	4128	16512

Tabla 1

En general:

1º Es suficiente un *nivel de confianza* de α = .05 (que equivale a z = 1.96); es la práctica habitual,

2º El *margen de error* no debe ser superior a .05 (5%) para que los resultados sean realmente informativos y útiles.

También podemos ir directamente a los programas de Internet que figuran en el Anexo.

3.2.1.2. *Cuando tenemos los sujetos que tenemos*: estimación del margen de error al extrapolar de la muestra a la población

¿Con qué *margen de error* podemos extrapolar los resultados a la población? Por *margen de error* entendemos *intervalos de confianza* que es otra manera de decir lo mismo.

El partir de un número de sujetos determinado, sin haber calculado antes el tamaño de la muestra necesario para extrapolar los resultados a la población objeto de estudio, es una situación frecuente. El punto de partida, por lo que respecta al número de sujetos, es con frecuencia uno de estos tres:

a) Partimos de un número de sujetos *aproximado* al que queremos, porque pretendemos extrapolar los resultados a la población (para entendernos: queremos afirmar *qué se piensa aquí en general,* y no *qué piensan estos sujetos concretos);* pero este número no es exacto (nos quedamos cortos o puede que tengamos más de los necesarios).

b) Partimos de los *sujetos disponibles,* que pueden ser menos de los necesarios, pero nuestra intención es extrapolar los resultados a la población, como en el caso anterior;

c) No pensamos extrapolar los resultados, estamos simplemente *describiendo una muestra,* pero siempre cabe el hacernos esta pregunta ¿Entre qué márgenes se encuentra la población que *puede* estar representada por estos sujetos?

La respuesta en estos casos es sencilla; se trata de *verificar al margen de error* cuando extrapolamos a la población a partir de la muestra que *de hecho* tenemos disponible.

En la fórmula [1] conocemos los valores de N, de pq (= .25) y el de z (que puede ser habitualmente 1.96). Despejamos el valor del margen de error (e):

$$\textit{margen de error } e = \sqrt{\frac{z^2 pq}{N}} \qquad [3]$$

Que también se puede expresar así: $e = \dfrac{(z)(.50)}{\sqrt{N}}$ [4]

Donde (.50) es la raíz cuadrada de pq (.25). Al nivel de confianza habitual z = 1.96, tenemos que (1.96)(.50) = .98, por lo que la fórmula [4] nos queda así:

$$e = \frac{.98}{\sqrt{N}} \qquad [5]$$

y si z =2, $$e = \frac{1}{\sqrt{N}} \qquad [6]$$

Por ejemplo: en una encuesta a 200 sujetos responden *sí* a una pregunta el 45%. ¿Entre qué límites se encontrará la población (de tamaño indefinido, desconocido) representada por esta muestra? (no olvidemos que la representatividad puede ser problemática, pero nuestra pregunta es legítima). Nuestro nivel de confianza, para no apartarnos de la práctica convencional y suficientemente segura, es de α = .05 (luego z = 1.96); nuestro margen de error será [5]:

Margen de error: $e = \dfrac{.98}{\sqrt{200}} = .069$, redondeando el 7%

Esto quiere decir que si en nuestra muestra responde *sí* el 45% podemos afirmar que en la población responde *sí* (sumando y restando 7% a nuestro 45%) entre el 38% y el 52%.

Este margen de error *(límites máximo y mínimo probables en la población)* es lo que se denomina coloquialmente *horquilla* en otros contextos (como en las encuestas pre-electorales); un término más apropiado es *intervalos de confianza*). Este margen de error del 7% (lógico con una muestra tan pequeña) podemos considerarlo grande o pequeño según de qué se trate.

De hecho, aun cuando la finalidad sea extrapolar de la muestra a la población, es normal disponer de un número de sujetos *aproximado* al que necesitamos para un determinado nivel de confianza (como α = .05 que corresponde a z = 1.96) y también aproximado para un margen de error dentro de unos límites que juzgamos tolerables. Ese margen de error, pero de manera más exacta, podemos estimarlo *después* con las fórmulas [3] o [4]. Por esto nos encontramos a veces en las *fichas técnicas de las encuestas* con unos márgenes de error que a primera vista parecen arbitrarios o que no responden a un valor que podamos considerar muy lógico o especialmente útil (como 3.17, 4.2, 2.27…); se han calculado a partir del tamaño real de la muestra.

3.2.1.3. Poblaciones y subpoblaciones

Una situación frecuente es la disponer de un número de sujetos suficiente para extrapolar los resultados a la población con un margen de error pequeño, pero si esta población está compuesta de *subpoblaciones* (como sucede en las muestras *estratificadas* o *por cuotas*) de cada una de las cuales se ha tomado una muestra menor, al extrapolar a cada subpoblación el margen de error será mayor (pero es naturalmente calculable y se puede dar la información).

Como ejemplo tomamos estos datos de la *ficha técnica* de una encuesta publicada en la prensa.

Población: País Vasco, N = 1800, con 600 sujetos de cada una de las tres provincias
Nivel de confianza: 95.5% (ó $\alpha = .045$; equivale a $z = 2$).

El margen de error al extrapolar los resultados nos lo da la fórmula [5] (o la fórmula [6] en este caso concreto de $z = 2$)

Todo el País Vasco (N = 1800)

$$e = \frac{1}{\sqrt{1800}} = .024 \text{ (2.4\%)}$$

Cada provincia (N = 600)

$$e = \frac{1}{\sqrt{600}} = .0408 \text{ (4.1\%)}$$

3.2.1.4. *¿De dónde vienen estas fórmulas?*

Podemos utilizar estas fórmulas sin entender *de dónde vienen,* pero es fácil entenderlas si tenemos una idea de lo que es la *distribución muestral de la media,* un modelo teórico de distribución que tiene su desviación típica o, con más propiedad, su *error típico* (y este modelo nos viene del *teorema del límite central).*

Sabemos que el *error típico* (o desviación típica) de la media es igual a

$$\textit{error típico de la media} = \frac{\sigma}{\sqrt{N}} \qquad [7]$$

Por las tablas de la *distribución normal* sabemos que un 66% de los casos (aproximadamente) se encuentran entre la media obtenida *más menos* un error típico, y un 95% de los casos se encuentran entre la media obtenida *más menos* 1.96 *errores típicos* (o desviaciones típicas, el concepto es el mismo).

Este valor de z (o *distancia* en *errores típicos* con respecto a la media) lo decidimos libremente y así establecemos los límites o *intervalos de confianza* de la *verdadera* media *(la media de la población,* no la de la muestra).

La media de la población (que es lo que nos interesa, pues deseamos extrapolar de la muestra a la población) estará entre la media obtenida en la muestra *más menos* $(z)(\sigma/\sqrt{N})$.

Si por ejemplo, y como es usual, escogemos un nivel de confianza de $\alpha = .05$, al que corresponde $z = 1.96$, tenemos un 95% de probabilidades de que la verdadera media se encuentre entre *más menos* $(1.96)(\sigma/\sqrt{N})$ con respecto a la media de la muestra.

Si a este *margen de error* le denominamos *e* (error, desviación), tenemos que:

$$e = \frac{(z)(\sigma)}{\sqrt{N}} \qquad [8]$$

Esta fórmula es exactamente igual a las fórmulas [3] o [4] ($\sqrt{pq}$ es la desviación típica de los ítems dicotómicos cuyas respuestas se codifican como 1 ó 0).

Despejando N en la fórmula [8] tenemos la fórmula [1] para un determinado valor de las otras variables (z, σ y e), y estos valores son $\sigma = .50$ (y σ^2 o pq = .25) y los valores de z *(nivel de confianza)* y e (error tolerado al *extrapolar* de la muestra a la población) los escogemos nosotros.

3.2.2. En el caso de poblaciones *finitas* (tamaño conocido, pequeño)

Las fórmulas anteriores se refieren a poblaciones grandes, de tamaño indefinido que no podemos conocer con exactitud. Más o menos a partir de los 100.000 sujetos ya estamos en ese tipo de poblaciones. En muchas ocasiones trabajamos también con poblaciones mucho más pequeñas; después de todo el que investiga es el que establece los límites de la población. Los alumnos de una universidad o de una carrera, o los profesores de un colegio, etc., pueden ser *nuestra*s poblaciones. A estas poblaciones las denominamos poblaciones *finitas* y su tamaño (de manera más o menos exacta) lo conocemos o podemos conocer.

3.2.2.1. Fórmulas para determinar el tamaño de la muestra

Cuando *conocemos el tamaño de la población,* la muestra necesaria es más pequeña y su tamaño se determina mediante la fórmula [9]:

$$n = \frac{N}{1 + \frac{e^2(N-1)}{z^2pq}} \qquad [9]$$

n = tamaño de la muestra que deseamos conocer,
N = tamaño conocido de la población,
e, z y pq (o σ^2) como antes.

Naturalmente también en estos casos hay que recordar que para poder extrapolar los resultados a la población, la muestra debe ser *representativa,* y estamos de nuevo con el problema del muestreo aleatorio.

Por ejemplo: deseamos hacer un sondeo de opiniones en un centro escolar que tiene 600 alumnos. En este caso N = 600; es el *tamaño de la población* que ya conocemos. Nuestro nivel de confianza va a ser del 95%, por lo tanto z = 1.96. Y como no queremos un error mayor del 3%, tenemos que e = .03. A falta de otros datos y para mayor seguridad suponemos que pq = (.50)(.50) = .25. La muestra necesaria será:

$$n = \frac{600}{1 + \frac{.03^2(600-1)}{(1{,}96^2)(.25)}} = 384$$

; necesitamos por lo tanto una muestra de 384 sujetos.

Cuando la población es grande (más de 30.000 sujetos) esta fórmula no aporta mucho y puede utilizarse la fórmula para poblaciones infinitas [1] que es más sencilla.

También podemos ir directamente a alguno de los programas de Internet, nos basta introducir el nivel de confianza (95%) y el tamaño de la población:

Market Research Surveys Online http://www.macorr.com/ss_calculator.htm
Raosoft sample size calculator**,** http://www.raosoft.com/samplesize.html

Al aumentar el tamaño de la población no aumenta proporcionalmente el tamaño necesario de la muestra, y llega un momento en que las dos fórmulas dan prácticamente los mismos resultados. Podemos verlo en la tabla 2; aplicamos la fórmula para distintos valores de N (tamaño conocido de la población) y cuando las muestras son grandes llegamos a las mismas o parecidas cifras que vimos antes para poblaciones infinitas.

Tamaño de la población	nivel de confianza α = .05 (z = 1.96)	
	para e = .05	para e = .03
N = 100	n = 80	n = 92
N = 150	n = 108	n = 132
N = 200	n = 132	n = 169
N = 250	n = 152	n = 203
N = 500	n = 217	n = 341
N = 1.000	n = 278	n = 516
N = 2.500	n = 333	n = 748
N = 5.000	n = 357	n = 879
N = 10.000	n = 370	n = 964
N = 100.000	n = 383	n = 1056
N = 1.000.000	n = 384	n = 1066
N = 2.000.000	n = 384	n = 1066

Tabla 2

Cuando la población es muy pequeña y el error tolerado muy pequeño, prácticamente hay que tomar a toda o casi toda la población. En la tabla 3 tenemos el *tamaño de muestra* para poblaciones entre 25 y 40 sujetos (40 puede ser el tamaño típico de muchas clases) a partir de la fórmula [9]. El *nivel de confianza* es α = .05.

Tamaño de la población	*error tolerado*		*Tamaño de la población*	*error tolerado*	
N	e = .05	e = .03	N	e = .05	e = .03
40	36	38	32	30	31
39	35	38	31	29	30
38	35	37	30	28	29
37	34	36	29	27	28
36	33	35	28	26	27
35	32	34	27	35	26
34	31	33	26	24	25
33	30	32	25	24	24

Tabla 3

Con un error tolerado del 5% y poblaciones entre 25 y 15 sujetos la muestra debe ser N–1 (podemos prescindir de un sujeto) y con menos de 15 sujetos debemos incluir a toda la población. En determinados casos el número real de respuestas en una clase (ejemplo típico) es muy bajo y podemos preguntarnos en qué medida los resultados (por ejemplo de una escuela, de una clase) son *fiables*.

En ocasiones podemos comprobar *a posteriori,* y con una razonable seguridad, si con una *muestra inadecuada* (en tamaño y falta de aleatoriedad) hemos obtenido unos resultados *válidos*. Un ejemplo típico es cuando los alumnos evalúan a sus profesores y responden muy pocos alumnos; cada clase es una *población* y obviamente la muestra que ha respondido, si es muy pequeña, no representa en principio a la población (a toda la clase).

En un caso semejante podemos calcular, tomando a la clase como *unidad de análisis* (como si cada clase fuera un sujeto), la correlación entre *proporción de respuestas* y *media* del profesor en cualquier ítem (o en la media de todos). Si la correlación es cero o muy baja podemos concluir que el número de respuestas no tiene que ver con cómo es evaluado un profesor[3].

3.2.2.2. Margen de error al extrapolar de la muestra a la población

Si conocemos el tamaño de la población N y el tamaño de la muestra *n,* y establecemos un nivel de confianza determinado (al que corresponde un valor de z en la fórmula), podemos averiguar, como hicimos antes con la fórmula para poblaciones *infinitas,* el margen de error (e) con el que podemos extrapolar los resultados a toda la población. En estas fórmulas nos limitamos a despejar el valor del error; este cálculo es útil porque con frecuencia partimos simplemente de la muestra que hemos podido conseguir.

$$e = \sqrt{\frac{(pqz^2)(N-n)}{n(N-1)}} \qquad \text{o también} \qquad e = \sqrt{\left[\frac{pqz^2}{n}\right]\left[\frac{N-n}{N-1}\right]} \qquad [10]$$

En poblaciones *infinitas* suponemos que el segundo miembro $\left(\frac{N-n}{N-1}\right)$ es igual a 1.

Por ejemplo, de una población N = 360 (pueden ser los alumnos de una Facultad) disponemos de los datos de una muestra n = 151. Si aceptamos un nivel de confianza del 95% (que corresponde a un valor de z = 1.96), ¿con qué margen de error podemos extrapolar los resultados a toda la población?

$$e = \sqrt{\left[\frac{(.25)(1.96^2)}{151}\right]\left[\frac{360-151}{360-1}\right]} = .0608, \text{ o un 6.1\% de margen de error.}$$

Si de una población N = 1000 tenemos datos de 200 sujetos (n = 200), el margen de error es semejante (6.2%). En todas estas fórmulas ya hemos indicado que el valor de pq es constante (.50 x .50 = .25): suponemos la máxima varianza posible en la población. Como la varianza de la población (que en principio nos es desconocida) no tiene por qué ser la máxima posible, todas estas fórmulas tienden de hecho a sobrestimar el tamaño de la muestra.

Una aplicación práctica de la fórmula [10] es conocer el *margen de error* con el que podemos extrapolar los resultados a las *subpoblaciones* que corresponden a las diversas submuestras.

Por ejemplo: en un *campus* de una universidad la población es exactamente de 2962 alumnos. En una encuesta de evaluación de la universidad, y por lo que respecta a los alumnos de todo el *campus,* con un nivel de confianza del 95% (z = 1.96) y un margen de error del 3% necesitaremos este número de alumnos según la fórmula [9]:

$$n = \frac{2962}{1+\frac{.03^2(2962-1)}{(1.96^2)(.25)}} = 784.68 \ (= 785)$$

En este *campus* hay cuatro facultades; en una de ellas el número total de alumnos *(subpoblación)* es N = 884, de los que han respondido al cuestionario una muestra de n = 291 alumnos. ¿Con qué margen de error podemos extrapolar los resultados de esta muestra de 291 alumnos a esa *subpoblación* concreta de 884 alumnos? Aplicamos la fórmula [10]:

---

3 Algunos estudios muestran que en estos casos la correlación es muy baja y frecuentemente negativa: a menor número de alumnos que responden a estos cuestionarios el profesor tiende a ser mejor evaluado. Este resultado tiene su lógica; los alumnos que van siempre a clase no suelen ser los más inclinados a evaluar mal a sus profesores.

$$e = \sqrt{\left[\frac{(.25)(1.96^2)}{291}\right]\left[\frac{884-291}{884-1}\right]} = .047$$

Redondeando tenemos un 4.7% de margen de error al extrapolar a la población.

Es decir que si un 60% por ciento responde que está satisfecho con la calidad de la enseñanza que recibe, habrá que entender que el tanto por ciento de satisfechos está entre el 55.3% (ó 60 - 4.7) y el 64.7% (ó 60 + 4.7).

La fórmula para poblaciones finitas (tamaño conocido) es de especial utilidad en el caso de poblaciones de tamaño moderado. Una reducción en el tamaño de la muestra puede suponer un ahorro importante en costes y trabajo.

3.3. Observaciones sobre el tamaño de la muestra

1º Recordamos lo dicho al principio: no conviene olvidar que un tamaño adecuado de la muestra *no es suficiente* para poder extrapolar los resultados; además es necesario que la muestra sea *representativa* de la población. No debemos preguntarnos solamente *cuántos* sujetos necesitamos, sino *cómo son seleccionados* (*cuántos* y *quiénes,* son las dos cuestiones importantes para extrapolar los resultados).

Cuando no es posible seleccionar una muestra aleatoria (trabajamos, por ejemplo, con grupos hechos) hay que tenerlo en cuenta en la interpretación de los resultados. Siempre podemos preguntarnos *¿A qué población puede representar esta muestra?*

2º Una *muestra grande,* o mayor de lo que realmente necesitamos según las fórmulas adecuadas, *no es mejor necesariamente* ni garantiza por sí sola el poder extrapolar los resultados con un menor margen de error. Una muestra grande puede estar *sesgada*, a veces precisamente por ser una muestra muy grande, con determinados segmentos de la población poco representados o representados en exceso.

3º Si nos encontramos de hecho con una muestra grande, podemos intentar dividirla en submuestras según características importantes y verificar en qué medida estas submuestras están representadas en la proporción que les corresponde (tendríamos en este caso un *muestreo por cuotas*). Podemos también reducir el tamaño de alguna submuestra, o eliminarla y no tenerla en cuenta… en cualquier caso debemos examinar y describir bien la muestra para interpretar los resultados.

4º Ya hemos visto que una muestra puede ser adecuada para extrapolar los resultados a *toda* una población general previamente definida, pero cada submuestra (varones y mujeres, subgrupos de edades, cursos, etc.) puede no tener el tamaño suficiente para extrapolar los resultados a cada subpoblación con el mismo margen de error. Ya hemos indicado la fórmula adecuada [10] para verificar estos márgenes de error.

El margen de error en las submuestras será mayor que en la población, pero también puede ser una información útil y así se ve con frecuencia en los resultados publicados de encuestas sociológicas, con un margen de error al extrapolar a toda la población y otros márgenes de error mayores al extrapolar a las subpoblaciones.

4. Número de sujetos para construir un instrumento de medición

Por lo que respecta al *número de sujetos* cuando se trata de *construir un instrumento de medición* como un *test* o una *escala de actitudes,* los análisis requeridos también imponen un mínimum de sujetos, independientemente de otros criterios.

Para hacer bien el *análisis de ítems* debe de haber al menos unos *cinco sujetos por ítem inicial;* por lo tanto si se parte de 60 ítems, con este criterio (que tomamos de Nunnally, 1978, que lo propone como número mínimo) harán falta 300 sujetos. Como criterio general es preferible tener más sujetos que menos; con más sujetos los resultados de los análisis serán más estables; si son pocos los sujetos es más probable que los ítems discriminen de manera distinta en otras muestras. Si además pensamos llevar a cabo un *análisis factorial* del instrumento tenemos que pensar también en el número de sujetos necesario para hacer este análisis (lo vemos más adelante).

La experiencia muestra por otra parte que se pueden construir instrumentos con muestras muy pequeñas (con muchos menos sujetos de los requeridos, por ejemplo los sujetos disponibles en nuestras clases…), pero utilizando varias muestras se pueden ir *acumulando* los análisis y verificar qué ítems mantienen una discriminación aceptable.

También cabe, por supuesto, combinar los datos de cada ítem (medias y desviaciones) procedentes de diversas muestras hasta conseguir una muestra del tamaño adecuado y calcular el coeficiente de fiabilidad en la muestra total.

Ciertamente, cuando se pretende construir un instrumento, es preferible utilizar muchos sujetos que pocos, pero como siempre en estos casos no hay que pensar solamente en el número sino en el tipo de muestra; debe ser suficientemente representativo de la población con la que se va a utilizar el instrumento (población general, universitarios, etc.).

5. Número de sujetos para el análisis factorial

a) No existe un criterio o norma definitiva sobre el número de sujetos necesario[4]; además no hay que tener en cuenta solamente el número de sujetos en términos absolutos, sino que es importante la proporción de sujetos con respecto al número de variables.

b) En el análisis factorial hacen falta en principio muestras grandes, porque el análisis factorial se basa en *coeficientes de correlación* y el error típico de las correlaciones (su *oscilación probable* de muestra a muestra) disminuye si aumenta el número de sujetos. Con muestras pequeñas se pueden ir acumulando muchos *errores de medición* a lo largo del proceso y aparecen factores puramente casuales, debidos a particularidades de la muestra.

c) Por otra parte con muestras relativamente pequeñas podemos encontrar con más facilidad ítems que definen más de un factor; si esto sucede con muestras grandes una explicación puede ser que el ítem no es apropiado; si esto sucede con muestras pequeñas no sabemos si debe a la formulación del ítem o a peculiaridades de la muestra (Osborne y Costello, 2004).

d) La recomendación habitual es utilizar *una muestra 10 veces mayor que el número de variables* o ítems (N = 10k donde k es el número de ítems o variables; Nunnally, 1978; Thorndike, 1982). Otros autores (Guilford, 1954; Kline, 1986, 1994) estiman suficiente una muestra menor, dos o tres veces el número de variables (N = 2k ó 3k), con tal de que el número de sujetos no sea muy inferior a 200. Muestras más pequeñas pueden ser aceptables si vamos a replicar el análisis en varias muestras (Kline, 1986:188).

e) La recomendación *mínima* que podemos hacer es utilizar muestras de al menos 150 o 200 sujetos aunque las variables (ítems) sean muy pocas (Kline, 1994:74, 79, pone el número

[4] Los diversos criterios propuestos por diversos autores sobre el número de sujetos en el Análisis Factorial pueden verse comentados en Garson (2003), Mundfrom et al. (2003) y Osborne y Costello (2004). El número de sujetos en el Análisis Factorial lo tratamos también en otro lugar (Morales, 2010).

mínimo en unos 100 sujetos); y el número de sujetos debe ser al menos el doble del número de variables.

f) Cuando el número de variables (ítems) es excesivo con relación al tamaño de la muestra (N pequeño), una sugerencia de Marsh y O'Neill (1984) es reducir el número de variables agrupándolas de dos en dos. Según estos autores las respuestas a dos ítems (o incluso más) se suman y constituyen una nueva variable. De esta manera el número de sujetos aumenta en proporción al número de variables. Las variables agrupadas tienen que ser razonablemente homogéneas, y esta homogeneidad debe ser comprobada empíricamente mediante la correlación ítem-total, que no debe ser inferior a .30 según estos autores.

Con este procedimiento las nuevas variables tienen una fiabilidad mayor (un componente específico, no común a las otras variables, menor) y se ven menos afectadas por peculiaridades de la redacción; se pierde en cambio información sobre las variables individuales.

La misma orientación la vemos en Cattell y Cattell (1995) en la preparación de la quinta edición de su test de personalidad (16PF). La razón aducida no es aumentar la proporción de sujetos por ítem, sino que dos o tres ítems agrupados tienen más fiabilidad que separados y la estructura factorial es más clara. Las subescalas de 14 ítems se reducen a 6 sumando los ítems que tienen entre sí sus mayores correlaciones.

c) Si la muestra es suficientemente grande, algunos autores (Amstrong y Soelberg, 1968) sugieren subdividirla y hacer el análisis factorial en las diversas submuestras. El repetir el análisis factorial en muestras distintas es la mejor manera de verificar si los factores se mantienen.

6. Tamaño de la muestra en estudios de carácter empírico o experimental

Ahora tratamos sobre el número de sujetos en los estudios experimentales (o cuasi-experimentales), más relacionados con la *t de Student* y el análisis de varianza, y en los análisis correlacionales en general.

El primer punto va a ser recordar las variables que en estos casos intervienen en la determinación del tamaño de la muestra.

6.1. Variables que intervienen en la determinación del tamaño de la muestra

Aunque en la práctica podemos limitarnos a consultar unas tablas, es muy conveniente conocer con qué criterios están hechas estas tablas. Se trata de las variables de las que depende el tamaño de la muestra.

1. *El nivel de confianza* (que solemos expresar así: $\alpha = .05$, $\alpha = .01$). Si escogemos un nivel de confianza de .05 (como es práctica común) queremos decir que aceptamos un 5% de probabilidades de error al rechazar la Hipótesis Nula (de no diferencia). Se trata de minimizar el denominado error Tipo I (aceptamos pocas probabilidades de equivocarnos cuando afirmamos una diferencia o una relación).

2. La *potencia de la prueba*. Por potencia entendemos la probabilidad de no cometer el error denominado Tipo II: *no rechazar la Hipótesis Nula cuando podríamos haberla rechazado*. La probabilidad de cometer este tipo de error se simboliza como β, y la potencia es por lo tanto $1-\beta$. Podemos definir la *potencia* como *la probabilidad de rechazar una Hipótesis Nula que es falsa.*

De la misma manera que un *nivel de confianza* de $\alpha = .05$ es habitualmente aceptado como razonable, por lo que respecta a la *potencia* ($1-\beta$) se estima que es razonable establecer una potencia de .80, es decir tener un 80% de probabilidades de detectar una diferencia (o relación) de una determinada magnitud[5]. Si deseamos una potencia mayor (.90 o incluso 1) el tamaño requerido de la muestra puede ser ya excesivamente grande.

El error Tipo I (decir *sí* cuando habría que decir que *no* hay diferencia, relación, etc.) es más serio que el error Tipo II (decir *no* cuando podríamos haber dicho que *sí),* de ahí la práctica generalizada de utilizar unos niveles de confianza muy estrictos, como son .05 ó .01: aceptamos muy pocas probabilidades de equivocarnos cuando afirmamos una diferencia. Si establecemos un nivel de significación muy estricto (un valor de a muy bajo) es muy improbable que cometamos el error Tipo I: si rechazamos el azar (o la variabilidad normal) como explicación de una diferencia es muy poco probable que nos equivoquemos.

Lo que sucede es que con un valor muy bajo de α podemos caer en el error Tipo II: puede ser que la Hipótesis Nula sea falsa, pero como somos muy estrictos en nuestro nivel de confianza no llegamos a rechazarla (la Hipótesis Nula se rechaza con mayor facilidad con $\alpha = .05$ que con, por ejemplo, $\alpha = .001$). En la práctica hay que *sopesar* ambos tipos de error. El minimizar el error Tipo I no significa que no tengamos que prestar atención al error Tipo II. Aunque las decisiones sobre el tamaño de la muestra se toman frecuentemente en función de los datos disponibles, o imitando lo que han hecho otros, no es racional, (como señala Cohen, 1988:55), el determinar el tamaño de la muestra sin tener en cuenta el error Tipo II.

3. La *magnitud de la diferencia* (o de la relación, etc.) que deseamos detectar y que solemos denominar *tamaño del efecto*. El término *efecto* no implica *causalidad,* sino simplemente el *grado* en que un fenómeno (diferencia, relación, etc.) está presente.

Lo *normal* es buscar resultadas (diferencias, relaciones,) *estadísticamente significativos* y no tanto pensar en qué *magnitud* podríamos estar interesados. Sin embargo la magnitud es un dato obviamente importante y para poder valorar adecuadamente los resultadas obtenidos (sin que esto quiera decir que las única magnitudes de interés sean las *grandes*).

La implicación de la magnitud en el tamaño de la muestra es obvia: cuando las diferencias son grandes, nos bastan pocos sujetos para detectarlas, pero cuando son muy pequeñas necesitamos muchos sujetos; si solamente nos interesan diferencias grandes, necesitaremos muchos menos sujetos. Podemos intuirlo con un ejemplo muy claro. Si estamos interesados en comprobar si difieren en *altura* los escandinavos y los *twa* (pigmeos de Ruanda y Burundi) no necesitaremos muchos sujetos en las muestras; nos bastarán muy pocos sujetos de cada grupo para caer en la cuenta de que se trata de poblaciones muy distintas en altura. En cambio si se trata de encontrar diferencias pequeñas entre las poblaciones, no nos bastará con comparar muestras de tamaño pequeño. Es claro por otra parte que con muestras grandes es fácil encontrar diferencias *estadísticamente significativas* pero pequeñas y con frecuencia irrelevantes.

Al planificar cualquier tipo de experimento o análisis debemos tener en cuenta también en qué tipo de *magnitud* estamos interesados, porque si solamente son de interés magnitudes más bien grandes podemos ahorrar costes y trabajo utilizando muestras relativamente pequeñas.

4. La *varianza de la población:* ya sabemos que si los sujetos son muy iguales dentro de cada grupo, necesitaremos muestras menores para detectar diferencias (si todos son de idéntica altura, o todos piensan lo mismo, etc., nos bastaría un solo sujeto de cada grupo para ver si hay alguna diferencia entre los grupos).

---

[5] La recomendación de establecer una potencia de .80 la propone y justifica Cohen (1988:56; Jacob Cohen es la fuente principal que suele seguirse en este tema). El peligro de *cometer* el error Tipo II queda reducido a .20 (20% de probabilidades) y está en equilibrio con $\alpha = .05$; suponemos que el error Tipo I es cuatro veces más serio que el error Tipo II (.20 es cuatro veces .05). esta recomendación no es tan seguida cono la de establecer un nivel de confianza de .05, porque con frecuencia no se tiene en cuenta el error Tipo II.

Estas cuatro variables se combinan en las fórmulas apropiadas para determinar el tamaño óptimo de las muestras. Aunque en principio son preferibles las muestras grandes, por razones de economía (costos, trabajo) podemos *calibrar* el tamaño de la muestra de acuerdo con nuestras especificaciones en estas cuatro variables.

No necesitamos aplicar las fórmulas para conocer el tamaño de la muestra porque ya disponemos de tablas para las situaciones más frecuentes; las tablas que ponemos aquí están muy reducidas pero pueden ser suficientes como orientación sobre el tamaño de la muestra que debemos buscar (tablas más completas pueden encontrase en los autores que citamos y en otros). Sí es de interés conocer qué variables inciden en el número de sujetos que necesitamos.

No sobra recordar aquí que el *tamaño de la muestra* es importante, pero no es la única característica de la muestra que nos interesa. En diseños experimentales en sentido propio necesitaremos *muestras aleatorias,* y en cualquier caso siempre debemos preguntarnos a qué población pueden estar representando las muestras que utilizamos.

6.2. Tamaño de cada muestra cuando comparamos dos grupos *(t de Student)*

En la tabla 4 tenemos el tamaño de *cada* muestra necesario para comparar *dos* muestras[6]. Tablas semejantes, más o menos extensas o adaptadas, pueden encontrarse en diversos autores; no siempre coinciden exactamente las cifras del número de sujetos debido al distinto redondeo de decimales al aplicar las fórmulas.

Suponemos: varianzas iguales,
muestras de idéntico tamaño,
hipótesis bilaterales
potencia (1-β) de .80

nivel de confianza	d =.20	d = .30	d =.50	d = .70	d =.80	d =.1.0	d = 1.20
.05	392	174	63	32	25	16	12
.01	586	260	93	48	36	23	18

Tabla 4

Estamos suponiendo muestras de *idéntico tamaño,* pero si tenemos ya una muestra con un determinado número de sujetos, podemos calcular el tamaño necesario en la otra muestra.

La fórmula [Cohen, 1988:59] es ésta: $n_{nuevo} = \frac{(n_{disponible})(n_{tablas})}{2n_{disponible} - n_{tablas}}$ [11]

Vamos a suponer, por ejemplo, que tenemos ya un grupo experimental de 40 sujetos que ha tenido una determinada experiencia y deseamos compararlo con otro (grupo de control, o al menos como término de comparación); estamos interesados en detectar al menos una diferencia moderada (d = .50) a un nivel de confianza de $\alpha = .05$: ¿Cuántos sujetos deberemos incluir en el nuevo grupo de control? En las tablas vemos que necesitaríamos 63 sujetos en *cada* grupo; por lo tanto el tamaño del *nuevo* grupo deberá ser:

$$n_{nuevo} = \frac{(40)(63)}{(2x40) - 63} = 148$$

6 Valores seleccionados de la tabla 2.4.1 de Cohen (1988). Tablas semejantes podemos verlas en otros autores, como por ejemplo Light, Singer y Willett, J.B., (1990).

6.3. Tamaño de la muestra en el análisis de varianza
6.3.1. Cuando tenemos más de dos muestras *(análisis de varianza unifactorial)*

En la tabla 5[7] tenemos el número de sujetos necesario en *cada* muestra cuando tenemos más de dos muestras (entre tres y seis muestras).

En esta tabla hemos puesto como orientación los valores correspondientes a $\alpha = .05$ y $1-\beta$ *(potencia)* de .70 y .80; suponemos también un número idéntico de sujetos en cada muestra.

Podemos tomar como referencia de magnitud o el valor de ω^2 o el valor de f, el tamaño del efecto propuesto por Cohen (1988) cuando tenemos más de dos grupos[8]

a) El coeficiente ω^2 nos cuantifica el *grado de asociación* entre la variable independiente (el pertenecer a uno u otro grupo) y la variable dependiente.

b) El *tamaño del efecto f* propuesto por Cohen (1988) cuando tenemos más de dos grupos[9].

Cuando tenemos solamente dos grupos, ya sabemos que el tamaño del efecto es igual a la diferencia entre las dos medias dividida por la desviación típica combinada.

Cuando hay más dos grupos el denominador es el mismo, pero lo que tenemos en el numerador es la dispersión o desviaciones de todas las medias con respecto a la media común (un valor análogo a la desviación típica de las medias). En la práctica el cálculo más sencillo de *f* es a partir de ω^2 (que es habitual calcular como complemento al análisis de varianza); ambos valores están relacionados de esta manera:

$$f = \sqrt{\frac{\omega^2}{1-\omega^2}} \qquad [12]$$

Realmente si hemos calculado ω^2 ya no necesitamos conocer el valor de f, pues no va a aportar una información que nos lleve a una interpretación o a una valoración distinta. Por lo que respecta a tener una orientación sobre el tamaño de la muestra, nos basta consultar las tablas teniendo en cuenta, al menos de manera aproximada, el tipo de magnitud en la que estamos interesados.

Las valoraciones (magnitud *pequeña, moderada y grande)* son de Cohen y constituyen una referencia comúnmente aceptada como guía orientadora. Los tres valores de f = .10, .25 y .40 equivalen a .20, .50 y .80 cuando se trata de comparar dos medias (Cohen 1988:284-288).

Suponemos que las k muestras son de idéntico tamaño; si son de tamaño desigual podemos utilizar el *tamaño medio* de las muestras (N/k)[10].

Los valores tabulados son el número de sujetos en *cada* muestra. En el caso de tres muestras, si estamos interesados en verificar solamente si hay diferencias valoradas como

---

[7] Los valores de referencia seleccionados en la tabla están tomados de Cohen (1988, tablas 8.4.4 y 8.4.5); también pueden verse en Kirk (1995:186 y tabla E.13). Las tablas de Cohen (válidas hasta 25 muestras) son más fáciles de consultar, y utiliza f como criterio de magnitud; otros autores como Kirk utilizan ambos valores f y ω^2. Cohen utiliza el símbolo η^2 en vez de ω^2 (y comenta la falta de unanimidad en los símbolos en p. 282)

[8] Ponemos los dos valores porque podemos encontrar los dos como referencia en otras tablas.

[9] Explicado por Cohen (1988:274ss, 284). Se trata de un tamaño del efecto *global,* teniendo en cuenta todas las diferencias de las medias con respecto a la media total (no se trata de la diferencia entre *dos* medias, como sucede en el tamaño del efecto convencional al comparar dos medias).

[10] Las implicaciones del tamaño desigual pueden vese comentadas en Cohen (1988:360ss). Si las muestras mayores tienen también las mayores medias, el tamaño del efecto será mayor que si las muestras fueran de idéntico tamaño y también será mayor la potencia (y a la inversa también es verdad).

grandes y con una probabilidad de encontrarlas (si las hay) del 80%, necesitaremos una *muestra total* de (21)(3) = 63 sujetos; si en cambio consideramos un buen resultado el encontrar diferencias *pequeñas* (pero significativas), necesitaremos una *muestra total* de (322)(3) = 966 sujetos.

número de grupos	***potencia***	***pequeña*** ω^2 **= .01** **f = .10**	***magnitud moderada*** ω^2 **= .06** **f = .25**	***grande*** ω^2 **= .14** **f = .40**
3	**.70**	**258**	**42**	**17**
	.80	**322**	**52**	**21**
4	**.70**	**221**	**36**	**15**
	.80	**274**	**45**	**18**
5	**.70**	**195**	**32**	**13**
	.80	**240**	**39**	**16**
6	**.70**	**175**	**29**	**12**
	.80	**215**	**35**	**14**

Tabla 5

Comparando los valores correspondientes a una potencia de .70 y 80 podemos apreciar cómo al disminuir el número de sujetos disminuyen también las probabilidades de rechazar la Hipótesis Nula.

6.3.2. Tamaño de la muestra en los diseños factoriales *(análisis de varianza, dos factores)*

En la tabla 6 tenemos el número necesario de sujetos en *cada celda* cuando tenemos *dos criterios de clasificación* (o *factores)* divididos en entre dos y cuatro niveles. Suponemos que en cada clasificación hay un idéntico número de sujetos, como es usual en estos planteamientos.

Suponemos también un *nivel de confianza* de α = .05 y una *potencia* (1-β) de .70 o de .80 En estas tablas los niveles (o subclasificaciones) de cada factor pueden ser 2, 3 ó 4.

Para valorar la *magnitud* utilizamos los mismos criterios de la tabla 5[11]

El *número total de sujetos* será igual al número de sujetos que aparece en la tabla multiplicado por el número de subclasificaciones o celdas. En una tabla 2x3 tenemos 6 celdas; si estamos interesados en detectar diferencias *moderadas* y *potencia* de .70, el número total de sujetos será 6x21 = 126, y con *potencia* de .80 6x26 = 156.

El número de sujetos especificado en la tabla es suficiente para detectar si uno de los dos factores (o los dos) es estadísticamente significativo (si hay diferencias entre los niveles de cada factor, o, lo que es lo mismo, entre las medias de cada columna o de cada fila), pero en el caso de la interacción con estos mismos números la potencia es menor porque intervienen menos sujetos (no los totales de cada fila o columna sino los que hay en cada clasificación).

[11] Los valores del tamaño de la muestra (en *cada* clasificación), puestos como referencia orientadora, están seleccionados de las extensas tablas de Kirk (1995:401 y tabla E.13).

Tamaño de la tabla	*potencia*	*pequeña* $\omega^2 = .01$ f = .10	*magnitud* *moderada* $\omega^2 = .06$ f = .25	*grande* $\omega^2 = .14$ f = .40
2x2	.70	152	25	11
	.80	193	32	13
2x3	.70	127	21	9
	.80	158	26	11
2x4	.70	109	18	8
	.80	134	22	9
3x3	.70	85	14	6
	.80	106	18	7
3x4	.70	73	12	5
	.80	90	15	6
4x4	.70	55	9	4
	.80	67	12	5

Tabla 6

6.4. El tamaño de la muestra en estudios correlacionales

Como en otros casos semejantes, si estamos interesados en coeficientes grandes (.70 o más) los detectaremos con facilidad en muestras pequeñas; si estamos interesados en coeficientes pequeños, necesitaremos muestras mucho mayores.

El determinar cuándo un coeficiente es suficientemente grande depende del uso que vayamos a hacer de ese coeficiente. Si nuestro interés está en *predecir* resultados futuros (y con mayor razón si la medición va a tener consecuencias importantes, por ejemplo en pruebas de admisión o selección) los coeficientes bajos serán de poca utilidad; coeficientes mayores y con muestras grandes (para que el margen de error sea pequeño) nos darán una información más útil.

En términos generales y orientadores Cohen (1988:79-81) da estas valoraciones: r = .10 relación *baja,* r = .30 relación *moderada* y r = .50 relación *grande.*

Estas valoraciones pueden parecer muy modestas; después de todo una correlación de .10 quiere decir un 1% de varianza compartida por las dos variables, y una correlación de .30 no llega a un 10% de varianza compartida.

La razón de estas valoraciones (juzgar como moderado o medio lo que es claro que en términos absolutos es pequeño) está en que las *ciencias sociales* (de la conducta) son *ciencias blandas* y las verdaderas relaciones están muy atenuadas o empequeñecidas por todo tipo de *ruidos,* como son la baja *fiabilidad* de muchos instrumentos[12] o, quizás en un mayor grado, por la poca o discutible *fidelidad* del instrumento al constructo teórico.

El que la correlación sea pequeña puede significar no que sea *realmente* tan pequeña sino que medimos mal las variables, o que esta correlación está *contaminada* por otras variables que no tenemos en cuenta; casi nunca medimos variables *puras* (así la inteligencia, tal como la medimos, puede estar contaminada por niveles de educación, etc.). Las relaciones *observadas (medidas)* entre constructos pueden ser mucho más bajas de lo que serían si fuera posible medir

[12] Es posible aplicar las fórmulas de *corrección por atenuación* que dan una *estimación* de la correlación que podríamos obtener *si la fiabilidad fuera perfecta.*

rasgos puros, sin contaminaciones de todo tipo[13]. Por este tipo de razones podemos estar interesados en detectar correlaciones (significativas, no aleatorias) tan bajas como .10. Coeficientes de correlación muy pequeños, *si son significativos* (es decir, que probablemente *no son cero* en la población), pueden estar indicando alguna *ley psicológica*[14].

Algunos autores[15] señalan que una correlación de .30 (también aparentemente baja) viene a indicar el tipo de relación que un observador puede *observar casualmente;* es una relación detectable a *simple vista;* por ejemplo, cuando un profesor cae en la cuenta, al cabo de los años, de que entre los alumnos que se sientan en las últimas filas y junto a una ventana hay una proporción mayor de suspensos… esa relación *observable* podría ser del orden de r = .30 y ciertamente puede ser relevante.

Debemos tener en cuenta que a veces intentamos medir sentimientos profundos, recuerdos del pasado, valoraciones difíciles de hacer, etc., con preguntas sencillas que con frecuencia se responden muy rápidamente y sin mayor cuidado[16]; quizás no tenemos otra manera mejor de hacerlo en un momento dado, pero en cuanto instrumentos de medición resultan muy pobres (aunque pueden ser muy útiles).

En la tabla 7 seleccionamos el tamaño de la muestra necesario para detectar

a) una serie de valores de coeficientes de correlación (entre. 10 y .70) como referencia útil;
b) a un *nivel de confianza* de $\alpha = .05$
c) y para una *potencia* de .70, .80 y .90[17].

potencia	*coeficientes de correlación*				
	.10	.20	.30	.50	.70
.70	616	153	67	23	10
.80	783	194	85	28	12
.90	1047	259	113	37	16

Tabla 7

Si estamos interesados en una correlación tan baja como .10, con una muestra de 1047 sujetos tenemos un 90% de probabilidades de encontrarla (si se da)[18]. Si nuestro interés está en coeficientes muy altos, nos bastarán pocos sujetos para detectarlos.

Conviene recordar que si tenemos una correlación de una determinada magnitud y estadísticamente significativa obtenida en una muestra, podemos extrapolar a la población *el hecho* de la relación *(distinta de cero),* pero no su magnitud. Para determinar la magnitud de la

13 Cohen (1988:79) aduce (sin dar la cita) el comentario Thurstone: *en psicología medimos a los hombres por sus sombras.*

14 Guilford y Fruchter (1973:92). Correlaciones estadísticamente significativas pero *muy* bajas (de ninguna utilidad práctica) entre variables medidas de manera muy modesta (como simples preguntas para captar situaciones complejas o lejanas en el tiempo) pueden constituir una buena indicación para intentar medir las mismas variables con instrumentos más perfeccionados y verificar de nuevo esas relaciones.

15 Por ejemplo Cohen P. (1981) y Cohen J. (1988:80), y también otros autores hacen la misma observación. Cohen J. (1988:80) cita coeficientes de correlación *importantes* que son de este tipo de magnitud (.30).

16 Se podría decir, en relación a las preguntas de muchos cuestionarios y a cómo se responden, que lo que hacemos con frecuencia es intentar *atrapar sentimientos con un cazamariposas*. A veces podemos sospechar que una correlación muy pequeña, detectada con instrumentos muy pobres, es simplemente la *punta del iceberg*. Para Cohen (1988:79) muchas de las correlaciones que podemos buscar en las *ciencias blandas* de la conducta son del orden de .10 ya que en las variables, tal como las *operacionalizamos,* hay muchos *ruidos* (falta de fiabilidad o de *fidelidad* al constructo teórico, etc).

17 Los valores seleccionados están tomados de la tabla 3.4.1 de Cohen (1988)

18 Dicho con más propiedad, tenemos un 90% de probabilidades de rechazar (propiamente de *no aceptar)* la Hipótesis Nula de *no relación* en el caso de que sea falsa.

correlación en la población debemos calcular los intervalos de confianza. Aquí entra el error típico de la correlación, que depende del número de sujetos (es menor si aumenta el tamaño de la muestra). Si deseamos una estimación muy precisa de la correlación en la población necesitaremos siempre un número grande de sujetos.

7. Cuando N = 1

Tratando del número de sujetos, podemos preguntarnos ¿Es posible hacer estudios *experimentales,* con *tratamiento estadístico,* cuando N = 1? La respuesta es sí. El estudio de casos individuales puede ser de gran interés para investigar resultados de terapias individuales; en psicología clínica y en educación especial. Estos estudios con N = 1 pueden tener una clara *significación clínica y práctica.* El método requiere obtener del sujeto varias medidas *antes* y *después* de un tratamiento. Por otra parte varios estudios con N =1 pueden integrarse después, también con los análisis estadísticos adecuados[19].

8. Referencias bibliográficas

AMSTRONG, J. C. and SOELBERG, P., (1968), On the Interpretation of Factor Analysis, *Psychological Bulletin,* 70, 361-364.

BUSSE, R. T.; KRATOCHWILL, THOMAS R. and ELLIOT, STEPHEN N., (1995), Meta-Analysis for Single-Case Consultation Outcomes: Applications to Research and Practice, *Journal of School Psychology,* 33 (4), 269-285.

CATTELL, RAYMOND B. and CATTELL, HEATHER E.P., (1995), Personality Structure and The New Fifth Edition of the 16PF, *Educational and Psychological Measurement,* 55, 6, 926-937.

COHEN JCOB (1988). *Statistical Power Analysis for the Behavioral Sciences,* second edition. Hillsdale, N.J.: Lawrence Erlbaum.

COHEN, PETER A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. *Review of Educational Research,* 51, 281-309

GARSON, G. DAVID (2003). *PA 765 Statnotes: An Online Textbook* http://www2.chass.ncsu.edu/garson/pa765/statnote.htm, Factor Analysis http://www2.chass.ncsu.edu/garson/pa765/factor.htm (última revisión 27, 08, 2003)

GUILFORD, J. P. and FRUCHTER, B., *Fundamental Statistics in Psychology and Education,* 1973, New York, McGraw-Hill.

GUILFORD, J.P., (1954), *Psychometric Methods,* New York, McGraw-Hill.

HERNÁNDEZ SAMPIERI, ROBERTO; FERNÁNDEZ COLLADO, CARLOS y BAPTISTA LUCIO, PILAR (2008). *Metodología de la Investigación*. Cuarta Edición. México: McGraw-Hill

HERSEN, MICHEL (1990), Single-Case Experimental Designs, en BELLACK, ALAN S., HERSEN, MICHEL and KAZDIN, ALAN E., (eds.) *International Handboock of Behavior Modification and Therapy* (2nd edit.), New York and London, Plenum Press, 175-212 (438/701).

KENNEDY, CRAIG H. (2004). Recent Innovations in Single-Case Designs . *Journal of Behavioral Education*, Dec2004, Vol. 13 Issue 4, 209-211

KIRK, ROGER E., (1995). *Experimental Design, Procedures for the Behavioral Sciences.* Boston: Brooks/Cole.

KLINE, PAUL, (1986), *A Handbook of Test Construction,* New York, Methuen.

KLINE, PAUL (1994). *An Easy Guide to Factor Analysis.* Newbury Park: Sage

[19] El sujeto único puede ser también una clase o una agrupación (Lundervold y Belwood, 2000). No es éste el lugar para explicar estos procedimientos; se puede sugerir una bibliografía indicativa; por ejemplo Hersen (1990); Sharpley (1990), McGuigan (1993) (dedica un capítulo al análisis de *diseños* de *caso único,* lo mismo que otros autores), Kratochwill, Mott, y Dodson, (1984), Buse, Kratochwill y Elliot (1995, que presentan un *meta-análisis* con 44 estudios con N entre 1 y 4); Lundervold y Belwood (2000); Olive y Smith (2005); Kennedy (2004) tiene la introducción de un número monográfico del *Journal of Behavioral Education* dedicado al caso único (*single case*).

KRATOCHWILL, THOMAS R., MOTT, STACEY y DODSON, CECILIA L. (1984). Estudio de caso e investigación de caso único. En BELLACK, ALAN S. y HERSEN, MICHAEL (Eds.), *Métodos de Investigación en Psicología Clínica* (pp. 67-114). Bilbao: Desclée de Brouwer

LIGHT, R.J., SINGER, J. D. and WILLETT, J.B., (1990) *By Design, Planning Research on Higher Education,* Cambridge, Mas., Harvard University Press.

LUNDERVOLD, DUANE A. and BELWOOD, MARILYN F. (2000). The Best Kept Secret in Counseling: Single-Case (N = 1) Experimental Designs. *Journal of Counseling & Development*, Winter 2000, Volume 78, 92-102.

MARSH, H.W. and O'NEILL, R., (1984), Self-Description Questionnaire III: the Construct Validity of Multidimensional Self-Concept Ratings by Late Adolescents, *Journal of Educational Measurement,* 21, 153-174.

MCGUIGAN, F.J., (1993), *Experimental Psychology, Methods of Research,* Sixth edit., Englewood Cliffs, N.J., Prentice-Hall.

MORALES VALLEJO, PEDRO (2010). *El análisis factorial en la construcción e interpretación de tests, escalas y cuestionarios* http://www.upcomillas.es/personal/peter/investigacion/AnalisisFactorial.pdf (última revisión, 15 de Mayo de 2010)

MUNDFROM, DANIEL J.; SHAW, DALE G.; LU KE, TIAN, KNOFCZINSKI, GREG and MCFANN, KIM (2003). *Minimum Sample Size for Exploratory Factor Analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, April 21, 2003, Chicago, IL.

NUNNALLY, J.C., (1978), *Psychometric Theory,* New York, McGraw-Hill.

OLIVE, MELISSA L. and SMITH, BENJAMIN W. (2005). Effect Size Calculations and Single Subject Designs. *Educational Psychology* Vol. 25, Nos. 2–3, pp. 313–324.

OSBORNE, JASON W. & COSTELLO, ANNA B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation*, 9(11). http://PAREonline.net/getvn.asp?v=9&n=11 (consultado 17 Dic. 2004)

RODRÍGUEZ OSUNA, JACINTO (1991). *Métodos de muestreo.* Cuadernos metodológicos. Madrid: Centro de Investigaciones Sociológicas (CIS).

RODRÍGUEZ OSUNA, JACINTO (1993). *Métodos de muestreo. Casos prácticos*. Cuadernos metodológicos. Madrid: Centro de Investigaciones Sociológicas (CIS).

SHARPLEY, C.F., (1990), Single-Subject Research, in REEVES, JOHN P., (Ed.). *Educational Research, Methodology, and Measurement, An International Handbook,* Oxford, Pergamon Press, 580-586

STARK, PHILIP B. (Department of Statistics, University of California, Berkeley). *SticiGui, Statistics Tools for Internet and Classroom Instruction with a Graphical User* http://www.stat.berkeley.edu/users/stark/SticiGui/index.htm (*chapter 18*) (Content last modified Mon 28 May 2007). (revisado 03, 09, 2008)

STATPAC INC (2003) *Questionnaires & Survey Design* en http://www.statpac.com/surveys/index.htm#toc (en *Sampling Methods,* o directamente en http://www.statpac.com/surveys/sampling.htm) (consultado 5 Agosto, 2010).

THORNDIKE, R.L., (1982), *Applied Psychometrics,* Boston, Houghton-Mifflin.

TROCHIM, WILLAM M. (Cornell University), *Research Methods Knowledge Base* http://www.socialresearchmethods.net/kb/ (en Contents: *Sampling,* Last Revised: 10/20/2006 (consultado 2 de Sept. 2007)

Anexo. Tamaño de la muestra y números aleatorios en Internet

Los recursos disponibles en Internet son muy numerosos; aquí nos limitamos a una breve selección.

1. *Tamaño de la muestra*

Son varios los programas de Internet que nos calculan el tamaño necesario de la muestra y los *intervalos de confianza* (o *márgenes de error*) dado el tamaño de la población y también cuál es el margen de error para un determinado tamaño de la muestra[20].

Algunas direcciones útiles:

Creative Research Systems. The Survey System Sample Size Calculador
http://www.surveysystem.com/sscalc.htm

Custominsight.com: Home: http://www.custominsight.com/index.asp) En menú: *insights* Survey Random Sample Calculator http://www.custominsight.com/articles/random-sample-calculator.asp

Dimension Research, Inc. Confident Intervals for Proportion Calculator
http://www.dimensionresearch.com/index.html (tamaño de la muestra e intervalos de confianza)

Grapentine Company, Inc. Sample Size Calculators
http://www.grapentine.com/calculator.htm

Graphpad Software Random number generator
http://www.graphpad.com/quickcalcs/randomN1.cfm

Market Research Surveys Online http://www.macorr.com/ss_calculator.htm

Raosoft sample size calculator**,** http://www.raosoft.com/samplesize.html

2. *Números aleatorios*

Random.org Web Interface to the True Random Numbers (del *Department of Computer Science, University of Dublin, Trinity College*)http://www.random.org/nform.html

Surf Stat Australia, http://www.anu.edu.au/nceph/surfstat/surfstat-home/ (en *tables*)

[20] Se pueden buscar en Google con *samples size calculators*