INTRODUCCIÓN AL ESTUDIO DE LAS VARIABLES ALEATORIAS. CONCEPTOS ESENCIALES

Prof Dr Mario Parisi

CAPITULO I

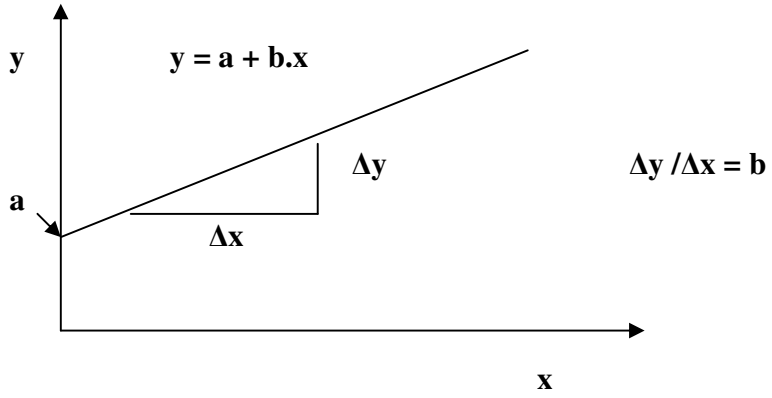
Introducción a la metodología de la ciencia

El método científico: Aceptaremos como definición operativa y a nuestros fines que el Método científico se basa en tres conceptos básicos: 1) Observación; 2) Experimentación y 3) Validez del Principio de causalidad.
El principio de causalidad: En idénticas circunstancias las mismas causas producen los mismos efectos.
Los sistemas experimentales: Clasificaremos a los sistemas experimentales (basados en la experiencia) en "determinados" o "aleatorios".
Son sistemas determinados aquellos en que las causas son plenamente conocidas y en consecuencia los resultados previsibles. Por ejemplo el movimiento de los planetas y otros astros en el vacío es predecible con gran precisión
En los sistemas aleatorios actúan simultáneamente un gran número de causas (prácticamente infinitas) sin que ninguna predomine sobre las otras, lo que hace imposible predecir el resultado. A esta situación se la denomina también **"fenómeno al azar".** El ejemplo clásico es al lanzar la bolilla en la ruleta.
Debemos aclarar aquí que esto no implica la no validez del principio de causalidad sino que simplemente nos es imposible conocer al sinnúmero de causas que simultánea y cronológicamente llevan al resultado final.
Datos experimentales: Pueden ser cualitativos (por ejemplo el color de la piel) o cuantitativos. Estos a su vez se dividen en continuos (por ejemplo la duración de una prueba determinado) o discretos (número de empleados en una empresa).
Probabilidad teórica: Clásicamente se la define como la relación entre casos favorables y casos posibles. Por ejemplo al arrojar un dado decimos que la probabilidad de obtener un determinado resultado es de 1/6 = 0,166 (o lo que es lo mismo de 16,6%). Esta definición solo es válida si sabemos "a priori" que todas las alternativas posibles tienen la misma probabilidad de ocurrir. Esto frecuentemente no es así en las ciencias experimentales y especialmente en el campo de la Psicología, por lo que la probabilidad teórica generalmente no es empleada.
Probabilidad experimental: Se la define como casos favorables sobre número de experiencias realizadas. Como se ve es un resultado obtenido "a posteriori" de la experiencia. Veamos un ejemplo. Pedimos a un grupo de 80 personas resolver un problema en un tiempo determinado. Al finalizar la prueba 54 personas lograron el objetivo: La probabilidad experimental en este caso fue de 54/80 = 0,675 ó 67,5%.
Lo que si podemos decir es que si ese grupo de personas es representativo de la población al cual pertenece (ver más adelante) el resultado obtenido en la muestra (las 80 personas) se puede extender (dentro de ciertos límites) a la población en estudio

CAPITULO II

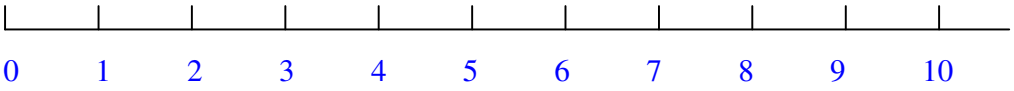
Un mínimo de Matemáticas

Por definición una **constante** es invariable (por ejemplo el número "pi").Un **parámetro** es un dato que se fija y mantiene invariable durante el experimento, mientras que las **variables** son los datos experimentales que mencionamos en el capítulo I y que pueden presentar distintos valores.
La forma más precisa de relacionar dos variables es poder definir una **función** (ecuación) que las relaciona. En el ejercicio 1 de la actividad IV se analizó la relación entre el número de empleados supervisados (variable que llamaremos x) y el grado de stress en los gerentes comerciales que los supervisan (variable que llamaremos y). Se llegó a la conclusión que se puede describir la situación como una relación lineal. Veamos entonces un poco las características de dicha relación, representada por una recta.

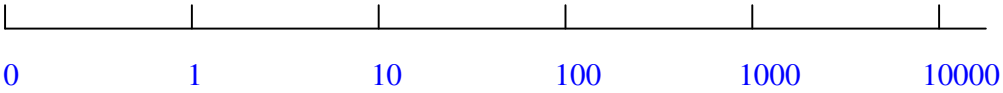


En ecuación de la recta (y = a + b . x) las variables en estudio son x (variable "independiente" e y (variable "dependiente). "a" es un parámetro que indica el punto en que la recta corta el eje de las "y" mientras que la relación Δy /Δx nos define el parámetro "b" que nos da la "inclinación de la recta. Evidentemente si conocemos "a" y "b" podemos trazar la recta en cuestión

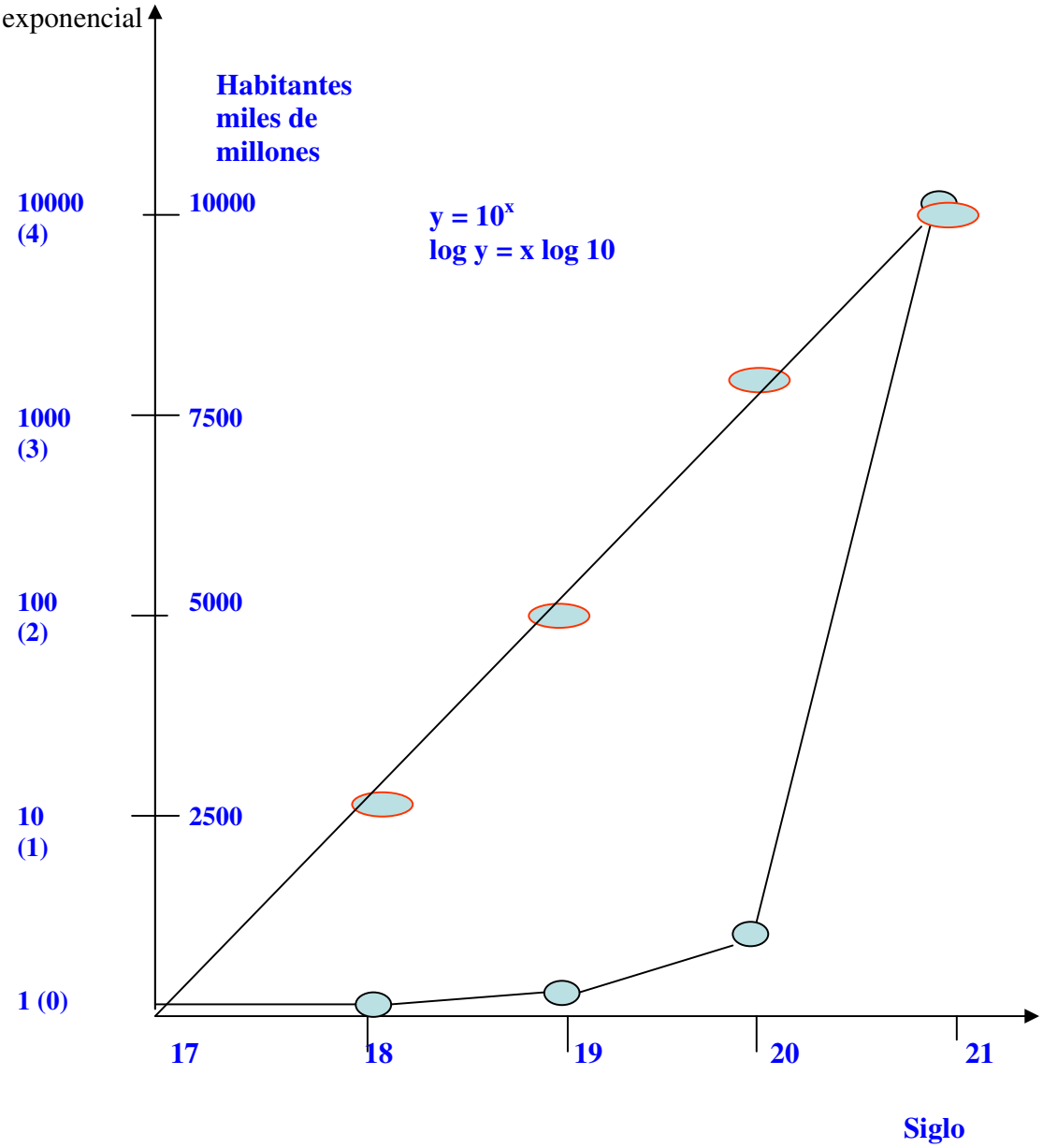
Concepto de medición:
La definición clásica dice: medir es comparar. Para ello tomamos un "patrón" determinado con el que comparamos la magnitud a medir.
Escalas:
Para medir o representar magnitudes son útiles las escalas. Para construirlas volcamos sobre una recta o semirrecta los valores correspondientes. La más simple es la escala lineal, en la cual las distancias son proporcionales a los valores de los números



En la escala logarítmica las distancias son proporcionales a los logaritmos de los números



Estas son útiles para representar fenómenos que crecen en forma geométrica o exponencial



Confección e interpretación de gráficos. Histograma.

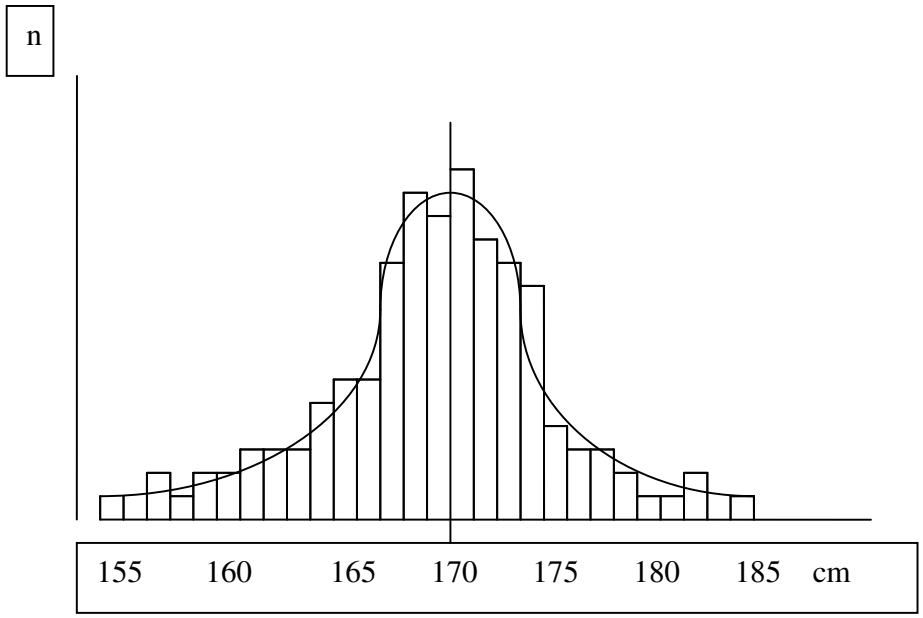
El "error" en una medición. Error de apreciación. Error accidental o estadístico. Error sistemático.
Toda medición tiene asociado "errores". Lo importante es detectarlos y "acotarlo" (decir cuanto vale)
Error de apreciación: Es el dado por la precisión del instrumento de medida (p.e. menor división de lectura. Ejemplo en psicología: Precisión con respecto al momento que ocurrió determinado evento)
Error sistemático: El debido a un error en la calibración del instrumento de medida. (En psicología "sesgo" del observador)
Error accidental: El debido a múltiples factores, independientes entre sí y sin que ninguno de ellos sea preponderante sobre los otros (al medir varias veces y en distintas circunstancias el coeficiente intelectual de una persona aplicando una determinada batería de estudio, podemos veremos diferencias que no son atribuibles a un factor en particular

CAPÍTULO III

.

Una aproximación conceptual a la Bioestadística.

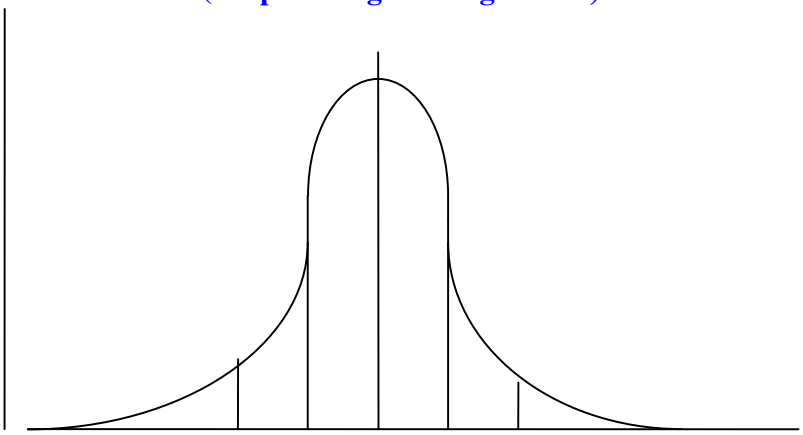
La distribución de Gauss o Normal: Es muy frecuente que las variables numéricas "aleatorias" (influenciadas por el azar" se distribuyan siguiendo la denominada función de Gauss, con su característica forma de "campana:



Nota: La "forma" de la campana se muestra algo "deformada" por problemas de diseño.

En realidad los datos experimentales son agrupados para construir un histograma (como se muestra en la figura precedente) y si la forma se aproxima a la de la función de Gauss aceptamos que estamos frente a una "distribución normal" o "gaussiana". En este caso se ha tomado el peso en 1000 jóvenes de 22 años tomados al azar entre los inscriptos en el padrón electoral de Buenos Aires y los resultados son representados. Es evidente que este caso podemos aceptar una distribución "normal" de la variable peso

La distribución de Gauss se caracteriza por dos parámetros: **la media o promedio (Xm) y la Desviación Estándar (SD por la sigla en inglés o σ)**



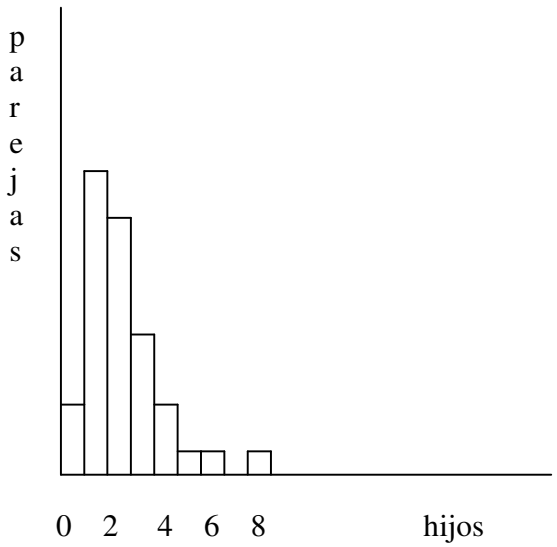
2SD 1SD Xm 1SD 2SD

La expresión Xm ± SD representa el 68% de la superficie bajo la curva dibujada por la función de Gauss. Si tomamos la expresión tendremos el 95 % del área y si Xm ± 3 DS el 99% del área (se reitera que por problemas de diseño la "campana" está algo deformada en nuestro ejemplo).

Regresando a nuestro ejemplo: Si aceptamos en base a lo observado que la distribución de los pesos de nuestros jóvenes fue normal, tendremos al 95% de ellos en el intervalo Xm ± 2 SD. Como la nuestra es representativa podemos además decir que en ese intervalo esperamos encontrar a los jóvenes de 22 años inscriptos en el padrón electoral de Buenos Aires, con un 95% de probabilidad. Hay un 5% de probabilidad de hallarlos fuera de ese intervalo.

El **parámetro Z** mide la distancia entre un valor y el promedio en unidades de desvío estándar. Esto implica que para un valor situado a 2,3 veces el desvío estándar le corresponde un Z de 2,3

No todas las variables se distribuyen siguiendo la función de Gauss. **Existen otras distribuciones.** Veamos un ejemplo: el número de hijos por pareja se distribuye de la siguiente forma:



Población y muestra aleatoria representativa: Una muestra es representativa de una población cuando fue tomada "al azar" (muestra aleatoria). Esto implica que todos los integrantes de la población tuvieron la misma probabilidad de formar parte de la muestra..

Inferencia estadística: Si una muestra es aleatoria y representativa los datos obtenidos con ella se pueden extender, a nivel de probabilidades a la población a la cual pertenecen (dentro de un cierto error). A esto se le llama **"inferencia estadística**"

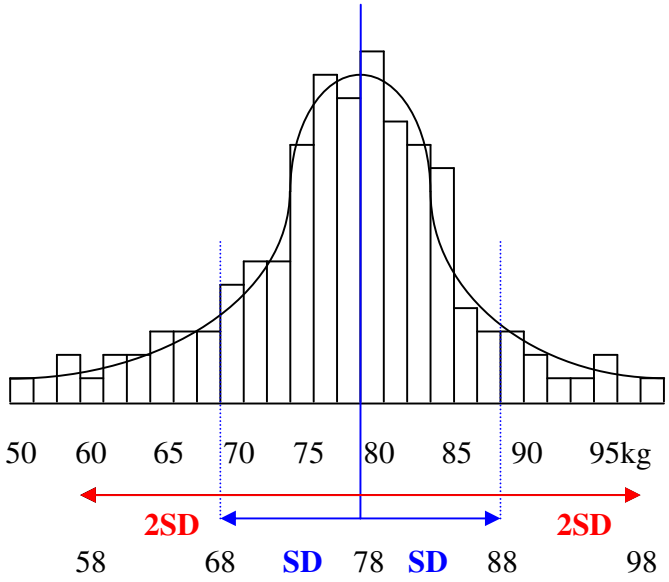
Probabilidad estadística: Recordemos aquí que la probabilidad se puede medir como fracción de 1 o como porcentaje. Así una probabilidad del 5% se puede también expresar como 0,05 y una del 38% como 0,38. Es común al hablar de probabilidades usar los símbolos "mayor" (>) o "menor" (<). Luego la expresión " $p < 0,05$" indica que la probabilidad de lo analizado es menor que el 5 %.

Probabilidad de observar determinados valores. Si una variable se distribuye normalmente y conocemos el valor de Xm y de SD podemos calcular el Z para determinado valor. A partir de una tabla (tabla de Z) es posible conocer el porcentaje de datos que se encuentran por encima (o por debajo) del mismo

CAPÍTULO IV

El error estándar de la media (SEM).

Regresemos al ejemplo del peso medido en 1000 jóvenes de 22 años tomados al azar entre los inscriptos en el padrón electoral de Buenos Aires y cuyos resultados son representados en el gráfico adjunto. Aceptemos también que el valor medio (Xm) calculado fue de 78 Kg y la Desviación Estándar (SD) de 10 Kg (Este ejemplo es simulado con fines didácticos. Recordar también que el gráfico presentado no responde fielmente a una campana de Gauss)



De lo antes expresado también sabemos que en el intervalo Xm ± SD (entre 68 y 88 Kg) esperamos encontrar al 68% de los casos estudiados (en el ejemplo 680 casos) y en el intervalo Xm ± 2SD (entre 58 y 98 Kg) esperamos encontrar al 95% de los casos estudiados (en el ejemplo 680 casos).

La siguiente pregunta que podemos hacernos es: ¿Cuán representativa de la media de la población en estudio es la media (Xm) observada en la media estudiada? Intuitivamente podemos aceptar que cuando más grande sea la muestra, más nos acercaremos al verdadero valor de la media de la población, pero existe un método relativamente sencillo para estimarla: **calcular el error estándar de la media (SEM, a partir de la sigla en inglés)**. El SEM se calcula como:

$$SEM = SD \sqrt{n}$$

Utilizando el SEM podemos definir ahora un nuevo intervalo:

$$Xm \pm SEM$$

Que define donde esperamos encontrar a la media de la población con una probabilidad del 68% (0,68). Si pasamos a la expresión

Xm ± 2 SEM

Definimos un intervalo dentro del cual esperamos encontrar a la media de la población con una probabilidad del 95%

Comparar valores medios. El error Standard de la diferencia (SEMdiff)

Para desarrollar el tema partiremos de un ejemplo: Se midió el umbral al dolor en 60 pacientes con antecedentes de tentativa de suicidio internados en hospitales psiquiátricos (muestra experimental**; exp**). En una escala de 1 a 100 el valor medio observado fue de 38 ($\mathbf{Xm_{exp}}$) con un desvío estándar de 5,3. A partir de estos datos podemos calcular el SEM:

$$\mathbf{SEM_{exp} = SD_{exp} / \sqrt{n} = 5,3 / 7,7 = 0,68}$$

Con este resultado podemos decir que la media de la población de la que se ha tomado la muestra en cuestión se halla en el intervalo**:**

38 ± 2 x 0,7 es decir entre 36,6 y 39,4 con un 95 % de probabilidad.

Como control de lo anterior se repitió el estudio en 60 voluntarios aparentemente normales (muestra control; **cont**) con los siguientes resultados: Promedio 33 ($\mathbf{Xm_{cont}}$)**,** desviación estándar 5,2
Con este resultado podemos calcular

$$\mathbf{SEM_{cont} = SD_{cont} / \sqrt{n} = 5,2 / 7,7 = 0,67}$$

Es decir que la media de la población de la que se ha tomado la muestra en cuestión se halla en el intervalo**:**

33 ± 2 x 0,7 es decir entre 31,6 y 34,4 con un 95 % de probabilidad.

Evidentemente hay una diferencia entre la muestra control y la experimental. ¿La misma es extensiva a las poblaciones correspondientes? Frente a ello hay dos probabilidades 1) Que la diferencia se deba al azar, es decir que no haya una causa específica que la justifique (hipótesis de nulidad) o que 2) Efectivamente el nivel de sensibilidad al dolor es más bajo en los pacientes estudiados que en los controles (hipótesis alternativa). Para poder decidir entre ambas hipótesis utilizamos un nuevo parámetro: el **Error Estándar de la Diferencia ($\mathbf{SEM_{diff}}$)** que se calcula como:

$$\mathbf{SEM_{diff} = \sqrt{(SEM_{exp})^2 + (SEM_{cont})^2}}$$

En nuestro caso

$$SEM_{diff} = \sqrt{(0{,}68)^2 + (0{,}67)^2} = 0{,}95$$

Para utilizar el SEM_{diff} efectuamos el cociente

$$(Xm_{exp} - Xm_{cont}) / (SEM_{diff})$$

al que llamaremos "Z de la diferencia" (Zdiff). En nuestro caso

$$Zdiff = (38 - 33) / (0{,}95) = 5{,}26$$

Aplicaremos un procedimiento "práctico y empírico" para utilizar el Zdiff ya que su justificación teórica excede los límites de este opúsculo. Diremos que si el valor de Zdiff es mayor que 2 la probabilidad de que la diferencia se deba al azar (hipótesis de nulidad) es menos que el 5% ($p < 0{,}05$). Este es así en nuestro caso, por lo que aceptamos la hipótesis alternativa: la diferencia no es debida al azar.

Cuando una diferencia no es debida al azar decimos que la diferencia es estadísticamente significativa,

El concepto de "doble prueba ciega"

El concepto de "doble prueba ciega" implica: 1) El experimentador debe desconocer si el individuo que evalúa pertenece al grupo control o al experimental. 2) El sujeto motivo del estudio debe asimismo desconocer si pertenece a uno de estos grupos.
Lo anterior es relativamente fácil de implementar si estudiamos el efecto de un fármaco. Por ejemplo:
En un estudio sobre el tratamiento de la depresión en adultos mayores (N Engl J Med 16;354(11) 1189-90 (2006)) se testeó el efecto de la paroxetina asociada a la psicoterapia interpersonal en pacientes de más de 70 años. Un grupo de 116 pacientes fueron distribuidos al azar en dos grupos: 1. (Grupo experimental): Recibían paroxetina y psicoterapia y 2 (Grupo control): Recibían placebo y psicoterapia (el placebo tiene la misma apariencia pero no contiene la droga). Los terapeutas no sabían cuales eran los pacientes que recibían droga o placebo y la misma situación se daba para los pacientes. El estudio es informado como:" Test 2 por 2, al azar, doble ciego, controlado por placebo" Se informa que el riesgo de recaída fue del 36% en el grupo 1 y del 63% en el grupo 2, siendo la diferencia significativa ($p > 0.02$).
En otros casos, por ejemplo en el estudio sobre la sensibilidad al dolor es más complejo cumplir con los requisitos de la "doble prueba ciega". Sin embargo en este caso dado el carácter más objetivo de la prueba el punto es menos crítico.

CAPÍTULO V

Estudios con muestras pequeñas.

Al comparar dos muestras con un n mayor de 30 **a**plicamos un procedimiento "práctico y empírico" para utilizar el Zdiff: Dijimos que si el valor de Zdiff es mayor que 2 la probabilidad de que la diferencia se deba al azar (hipótesis de nulidad) es menos que el 5% ($p < 0,05$). En realidad y para proceder con precisión debemos utilizar la denominada "tabla de Z" que nos dice los niveles de probabilidad asociados a cada valor de Z:

Tabla 1: Fracción de la curva normal y valores de probabilidad asociados a distintos valores de Z

Z	Fracción del área desde El centro hasta ± Z	Probabilidad (2 colas)
0	0,00	1,00
0,5	0,38	0,24
1	0,68	0,32
1,5	0,86	0,14
2,0	0,95	0,05
2,5	0,98	0,02
3,0	0,99	0,01
3,5	0,999	0,001
4,0	0,9999	0,0001
…..	……	……
∞	1,00	0,00

Por supuesto existen y pueden ser intercalados todos los valores que se deseen entre Z= 0 y Z = ∞

El parámetro t

Ahora bien. Intuitivamente podemos aceptar que si el número de casos (n) es menor que 30 las probabilidades asociadas a un determinado valor de Z son tanto menos seguras cuanto menor es el valor de n. Hablamos entonces de "muestras pequeñas**" para las cuales, al comparar 2 de ellas el nivel de probabilidad de que la diferencia entre ellas se deba al azar no dependera exclusivamende del valor de Z diff sino también de n. Cuando trabajemos con menos de 30 casos a Zdiff lo llamaremos t, los que nos indica que debemos utilizar una nueva tabla de probabilidades (tabla de "t") en la que los valores dependen del número de casos.**

En el ejemplo en que se comparó en el capítulo cuatro el umbral al dólar de pacientes internados y un grupo control con un n igual a 60 en ambos casos se llegó a un valor de

$$Zdiff = (Xm_{exp} - Xm_{cont}) / (SEM_{diff})$$

$$Zdiff = (38 – 33) / (0,95) = 5,26$$

Con la nueva tabla de Z que hemos incorporado ahora podemos decir que la probabilidad de que la diferencia se debe al azar es no solo menor del 5 % (0,05) sino del 1 por diez mil (0,0001). Pero supongamos que en vez de 60 hubieramos trabajado con 10 casos en cada muestra y que los valores medios y las desviaciones estándar hubieran sido las mismas (**Xm_{exp}** = 38 con un desvío estándar de 5,3; **Xm_{cont} = 33,** desviación estándar 5,2)

Evidentemente los valores de SEM hubieran sido distintos:

$$SEM_{exp} = SD_{exp} / \sqrt{n} = 5,3 / 3,2 = 1,65$$

$$SEM_{cont} = SD_{cont} / \sqrt{n} = 5,2 / 3,2 = 1,62$$

$$SEM_{diff} = \sqrt{(1,65)^2 + (1,62)^2} = 2,31$$

$$t (Zdiff) = (38 – 33) / (2,31) = 2,16$$

Para averiguar la probabilidad debemos ahora ir a la tabla de t

---

Tabla 1: Valores de probabilidad asociados a distintos valores de t

---

n-1→	5	10	15	20	25	30
t↓						
1	0,4	0,4	0,4	0,4	0,4	0,4
2	0,2	0,1	0,1	0,1	0,07	0,05
2,5	0,1	0,07	0,05	0,05	0,02	0,02
2,7	0,05	0,05	0,02	…..	…..	0,015
3	0,05	0,02	….	0,015	…..	0,010
4	0,02	0,01	0,001	…..	…..	0,0001
5	0,01	0,001	…..	….	…..	…..

n-1: Número de casos menos 1 (Grados de libertad, ver más adelante)
Tabla modificada y simplificada a partir de Fisher and Yates, 1938)

Como podemos ver la probabilidad de que la diferencia se debiera al azar es ahora, en nuestro ejemplo, 0,05 (5%) por lo que aún podemos desechar que la diferencia se deba alazar, pero con una probabilidad de error mucho mayor que en el caso de n = 60.

($p<0,001$) La tabla nos enseña también que con 15 casos el valor "necesario" de t para llegar a una $p < 0,05$ es de 2,5 en vez de 2.

Concepto de "grados de libertad"

Para entender el concepto lo mejor es un ejemplo. Supongamos que introducimos en una bolsa cuatro bolillas: Una roja, una verde, una azul y una amarilla. Las retiramos ahora de a una y salen, en las tres primeras ocasiones, la roja, la azul y la verde. Ahora tenemos la certeza, sin sacarla, que la que queda en el interior es la amarilla. Con cuatro bolillas tuvimos "tres grados de libertad" (n-1). Otro caso: Supongamos que se forman 10 parejas (n=20) para participar en un juego de sociedad (Juan y Alicia, Pedro y Susana, Alfredo y Susana, etc). Las parejas se hallan en un cuarto contiguo y se incorporan al juago una a una. Cuando hayan entrado 9 parejas (18 personas) ya sabemos, sin que entren, cuales son las otras dos. En este caso los "grados de libertad" fueron "20 – 18 = 2). Finalmente: Si me dan el valor medio entre n valores y luego los valores correspondientes a n-1 de esos datos puedo calcular sin problemas el dato faltante. Otra vez los grados de libertad han sido "n-1"

Estudios con datos apareados

Una vez más recurriremos a un ejemplo. Supongamos que tenemos 10 datos correspondientes al número de veces que aparecen palabra relacionadas con el concepto de "depresión" (previamente definidas) en el discurso libre en un grupo de pacientes. Los datos son:

8; 14; 22; 9; 18; 20; 25; 6; 23; 6

Calculamos la media y el desvío estándar y el SEM: Xm =15,1; SD = 7,41; SEM = 2,34
Nos dicen que en otro grupo los valores fueron:

6; 12; 20; 10; 14; 15; 24; 6; 18; 5

Calculamos la media y el desvío estándar y el SEM : Xm = **13,0; SD=6,42; SEM = 2,03**

Nos preguntan si la diferencia entre los promedios de ambos grupos es significativa. Para saberlo aplicamos (n = 10) una prueba de t:

$$\mathbf{SEM_{diff} = \sqrt{(2,34)^2 + (2,03)^2} = 3,09}$$

$$\mathbf{t\ (Zdiff) = (14,6 - 13,1) / (3,09) = 0,48}$$

Sin siquiera mirar la tabla de t podemos decir que la diferencia no es estadisticamente significativa.

Sin embargo ha habido un error en la información. Nos dicen que en realidad los datos corresponden a los mismos pacientes, antes y despues de 1 año de tratamiento psicoterapéutico. Es decir que son pareados, Por lo que podemos ahora construrir la tabla siguiente a partir delos datos:

8; 14; 22; 9; 18; 20; 25; 6; 23; 6
6; 12; 20; 10; 14; 15; 24; 6; 18; 5

Paciente	Antes	Después	Diferencia
1	**8**	**6**	**-2**
2	**14**	**12**	**-2**
3	**22**	**20**	**-2**
4	**9**	**10**	**+1**
5	**18**	**14**	**-4**
6	**20**	**15**	**-5**
7	**25**	**24**	**-1**
8	**6**	**6**	**0**
9	**23**	**18**	**-5**
10	**6**	**5**	**-1**

Calculamos ahora la media, el desvío estándar y el SEM de las diferencias: Xmd = -2,1; SDd = 1,92; SEMd = 0,60. En este caso t se calcula como:

T = Xmd / SEMd = 2,1 / 0,68 = 3.08

Ahora la tabla de t nos dice que la diferencia es estadisticamente significativa con una p> 0,05

Bibliografía central:

Aron A, Aron E.E:
Estadística para Psicología
Pearson-Prentice Hall, 2001-2006

Klimovsky, Gregorio
Las desventuras del conocimiento científico: una introducción a la epistemología - 4a ed.
Buenos Aires : Editorial A-Z , 1999

Parisi Mario
Introducción al Estudio de las Variables Aleatorias
Campus Sociedad Argentina de Fisiología (a partir del 16-4-2007)
www.safisiol.org.ar .

Steel, Robert G. D..; Torrie, James H.
Bioestadística: principios y procedimientos - 2a ed.
Mc Graw Hill - Interamericana , 1992