

---

# PROGRAMA DE ACTUALIZACION CONTINUA Y A DISTANCIA EN UROLOGIA

---

Comité de Educación Médica Continua  
Sociedad Argentina de Urología

## Módulo 6 - Fascículo 1 - 2001

**El ABC de los ensayos clínicos controlados**  
*Dr. Hernán Claudio Doval*

**Director**

*Dr. Jorge H. Schiappapietra*

**Secretario**

*Dr. Carlos A. Acosta Güemes*

**Asesor**

*Dr. Elías J. Fayad*

**SAU**

SOCIEDAD ARGENTINA DE UROLOGIA

# EL ABC DE LOS ENSAYOS CLINICOS CONTROLADOS

Dr. Hernán Claudio Doval

Director del Grupo de Estudio, Docencia e Investigación Clínica (G.E.D.I.C.)

***Para observar tienes que aprender a comparar. Para comparar necesitas haber observado. La observación genera un conocimiento y el conocimiento es necesario para observar. Observa mal, el que no sabe hacer nada con lo que haya observado.***

***Para el manzano tiene un ojo más agudo el cultivador de frutas que el paseante, pero no ve exactamente al hombre quién no sepa que el hombre es el destino del hombre.***

**- Bertolt Brecht. "1934"**

Como las fronteras y parcelizaciones del "conocimiento" (en latín *scientia*) son imaginarias, es un escritor, claro que un escritor como Bertolt Brecht que piensa, y hace pensar, el que a mi entender escribió las palabras más transparentes, sencillas y a la vez bellas sobre el propio proceso de conocimiento.

Es un lugar común en medicina decir, que para afirmar que una "intervención controlada" es efectiva o mejor que otra, es necesario **comparar** esa intervención con una "no intervención" (control no tratado), una "intervención simulada" (control con placebo) u "otra intervención" (tratamiento alternativo).

Aquí la palabra clave central es **comparar**, que aparece encabezando el aforismo de Bertolt Brecht, claro que esta comparación sucede entre el grupo que realizó la intervención y el grupo control al final del proceso de un ensayo controlado.

¿Por qué no analizamos como inicia el proceso el investigador?

Todos decimos que el investigador adquiere uno por uno los "datos" de cada individuo, establecidos en el protocolo, y los que reúne paciente-mente constituyendo uno o más grupos de personas.

Acá debemos detenernos brevemente en la introducción de la palabra "datos", deriva del latín *datum*, significa lo que se da, el antecedente, la cosa o hecho anterior que sirve para juzgar los hechos posteriores; por lo tanto cuando decimos "adquirir los datos", los investigadores nos adscribimos a la línea filosófica materialista que reconoce la realidad objetiva fuera de nosotros (el conocimiento de la "cosa en sí" de los filósofos). Pero ¿qué queremos decir cuando mencionamos "adquirir los datos"?

Muchos dirán que es examinar las características, observar; o sea algo así como escudriñar con diligencia y cuidado, atentamente. Pero hacer esto solamente no es suficiente, como dice la primera línea de B. Brecht; **"para observar tienes que aprender a comparar"**. O sea que para observar se debe fijar la atención en dos o más objetos para descubrir sus relaciones o estimar sus diferencias o semejanzas; otra forma de decirlo es manifestar que comparar es un proceso por el cual se asigna una cualidad a un atributo esencial de una persona, siempre con respecto a otra, y que puede expresarse en palabras y/o números.

Cuando expresamos que un paciente tiene fiebre, estamos haciendo una comparación cualitativa con palabras; pero si decimos que tiene una temperatura de 39 grados centígrados estamos haciendo una comparación cualitativa numérica.

En gran parte de los estudios paraclínicos de la medicina, como los análisis de laboratorio, se pueden realizar "comparaciones numéricas" es decir mediciones; ya que medir es comparar cuántas veces una cantidad

contiene a su respectiva unidad (decir que existe una glucemia de 120 mg %, es decir que 100 ml de sangre contienen una magnitud de glucosa que es comparativamente una cantidad de 120 veces la unidad de 1 mg), para hacer esto también resulta cierto que **"para comparar necesitas haber observado"**. Las comparaciones numéricas o mediciones son interesantes porque permiten el manejo y utilización algebraica de dichos números que en el caso de las **mediciones** en escalas de proporciones se restan, se dividen, etc., etc.).

Sin embargo, la mayoría de los datos y los juicios de la práctica clínica, son comparaciones cualitativas que son intrínsecas a la vida humana. Si se los ignora, como se hace en gran parte actualmente, se excluiría de la atención científica al dolor del paciente, el malestar, tensión, insomnio, ansiedad, pesares y alegrías y otras cualidades y calidades de vida; por lo cual se borrarían las características específicamente humanas que distinguen las personas de los animales, los organismos microbianos y las moléculas químicas. Además se perdería importante información clínica pronóstica y terapéutica, como los distintos tipos o grados de severidad de los síntomas, tiempo de evolución de la enfermedad, velocidad de progresión o enfermedades asociadas.

Sin embargo, los buenos clínicos utilizaron y utilizan, de manera usual, una gran recolección de los atributos clínicos y humanos, que distinguen a cada paciente, y que les permiten tomar una decisión médica correcta en cada uno individualmente.

Para nuestra tranquilidad esas comparaciones cualitativas se pueden realizar, ya sea con palabras, por lo cual la escala que utiliza se denomina "nominal". Por ejemplo el sexo que es masculino o femenino, el estado civil, el grupo sanguíneo. Muchas de estas escalas nominales son dicotómicas (binominales), como el estado vital de una persona que puede estar viva o muerta únicamente; asignando un número a cada término nominal y contando, podemos decir de un grupo que ocurrió un 10 % de mortalidad (10 de 100 pacientes murieron) o que tienen un 90 % de sobrevivencia (90 de cada 100 pacientes están vivos), que son dos formas de decir lo mismo con los dos términos dicotómicos excluyente de esa escala nominal.

Gran parte de las escalas que utilizan en la clínica son de orden (escalas ordinales), en la cual el número solamente tiene un significado de posición, en qué orden se encuentra si tomamos como punto de partida o de unidad la cualidad de uno de ellos y le asignamos a los restantes una cifra que corresponda a su orden relativo (en esta escala no tienen validez los procedimientos matemáticos). Los Grados I,II,III,IV de la capacidad funcional en la insuficiencia cardíaca o coronaria, constituye un típico ejemplo de transformación numérica a una escala ordinal, o el puntaje de vitalidad del recién nacido instituido en el índice de Apgar, o cuando clasificamos la motilidad regional de la pared miocárdica en el ecocardiograma de -1 a 3 (llamando: -1 a la diskinesia, 0 a la akinesia, 1 a la hipokinesia, 2 a la normokinesia, 3 a la hiperkinesia).

Ahora estamos en condiciones de decir que de las propiedades que se miden en escalas nominales y/u ordinales se ocupa la Estadística No Paramétrica; a su vez la Estadística Paramétrica se encarga de estudiar las propiedades numéricas que son medidas en escalas de intervalo o proporción y tienen distribución gaussiana o normal.

De esta forma **"la observación genera un conocimiento y el conocimiento es necesario para observar"**. Estamos de acuerdo que si podemos medir o transformar numéricamente atributos ya podemos generar conocimientos elementales de las cosas o hechos, pero ¿por

qué Brecht dice en la misma frase “... **y el conocimiento es necesario para observar**”. Es cierto, para observar y comparar con escalas numéricas debemos conocer sus unidades, para medir una longitud debemos conocer qué es un metro, un decímetro, un centímetro, etc. (unidades arbitrarias establecidas por consenso en el pasado). Para poder utilizar las escalas nominales u ordinales de la clínica debemos definir por consenso con otros y aceptar cada posición que establezcamos en la escala; y ese es un conocimiento necesario para permitirnos observar. *O sea los hechos de los datos no existen sin el conocimiento de una teoría de la observación.*

Cuando con más cuidado se definan los atributos de las variables operativas de un trabajo, el dato se puede hacer “consistente”, es decir “repetible por el mismo observador y reproducible por otro”, es decir intentamos que el dato obtenido sea lo más *preciso* posible.

Estas características son los ingredientes básicos que convierten a los considerados “*datos blandos*” de la clínica en similares a los “*datos duros*” paraclinicos.

Si se logra la “**precisión**” (consistencia), la “**validez**” (seguridad) del dato es relativamente fácil de conseguir, solo se necesita un standard aceptado contra el cual “controlar” el resultado de la medición. Tales estándares, pueden muchas veces ser desarrollados fácilmente, como un “consenso de conocedores” de los procedimientos de medición. Como todas las mediciones “duras” actuales, que tienen un estandar reconocido por consenso (hora, metro, litro, gramo, etc.).

A la adquisición de cada dato individual “consistente”, Alvar R. Feinstein lo llama “*mensuración*” (que es un sinónimo de medición de una unidad); si luego reunimos los datos individuales en “grupos” (uno, dos o más), estamos efectuando una “*cuantificación*” de los datos (o sea una expresión numérica de una magnitud), para ello nos valemos de la así llamada “**estadística descriptiva**” (porcentaje, mediana, rango, media, desvío standard, etc.).

Si la “cuantificación del grupo” que formamos, es de una “*muestra representativa*” extraída de la **población** en estudio, o sea que ha existido un muestreo aleatorio (por azar) que permitió que todos los individuos de la población tuvieran la misma posibilidad de formar parte de la muestra; las afirmaciones que surjan serán aplicables por la estadística “inferencial” (inductiva), a la población de la cual proviene la muestra (Figura 1).

Así, si conociéramos la media y el desvío standard de la altura de los recién nacidos de una muestra representativa, podríamos estimar “aproximadamente” cuál es el rango de altura que comprende al 95% de los recién nacidos recolectados retrospectivamente durante el período de medición. O sea con la lógica de la inferencia hacemos “**retrodicciones**” (o sea predicciones de lo que ya pasó), pero tenemos una inhabilitación lógica para utilizar esos mismos datos y hacer la “**predicción**” de la altura en los recién nacidos a posteriori del estudio estadístico. La predicción inductiva solo es posible en un mundo cerrado, repetitivo e inmutable; pero en el ejemplo considerado todos sabemos que no es así y la

altura de los recién nacidos aumentó con el tiempo en muchos países, quizás debido a mejoras en la nutrición de las embarazadas. Este fenómeno hace que las sociedades de pediatría tengan que actualizar periódicamente los diagramas de percentilos de crecimiento de los niños.

Por lo cual es necesario reconocer que la inducción predictiva no es un proceso lógico y se impone por la expectativa que crea la repetición de sucesos, haciendo que el próximo hecho se espere que resulte igual a los anteriores, si no se modifican las circunstancias en que se desarrolla; siendo posible la predicción aproximada durante un cierto tiempo.

Si bien, como ya dijimos, al resumir las características de una muestra (por medio de la estadística descriptiva), se genera por inferencia conocimientos de la población de donde se extrae, los ensayos controlados intentan crear otro tipo de conocimientos. O sea, saber si un evento o intervención sobre un grupo, en “*comparación científica*” con otro grupo control de la misma población (muestras representativas y similares), muestra una diferencia que se puede atribuir al azar y por lo tanto resulta probable que ambas muestras son iguales y sigan perteneciendo a la misma población. En ese caso el evento o intervención no causa ninguna modificación y por lo tanto no es efectivo. O es posible que la diferencia sea verdadera (no debida al azar) y ya no pertenezcan a la misma población, formando parte de dos universos distintos; por lo cual el evento o la intervención sí tuvieron efecto.

El investigador desea atribuir la diferencia medida al evento natural o a la intervención planificada. Para que esto sea cierto, es imprescindible que el grupo que actúa de control pertenezca a la misma población (intervención controlada), para evitar la aparición de factores que “*confundan*” los resultados. Ya que si se encontraran otras variables que fueran (numéricamente) distintas entre el grupo con intervención y el control, debido a un proceso metodológico inadecuado, las diferencias previas en la constitución de los grupos podrían justificar, por este solo hecho, la diferencia encontrada posteriormente en los resultados.

Los factores de orden metodológico que pueden “**confundir**” los resultados, podrían deberse a:

a) Grupo control o con tratamiento alternativo que “**no es concurrente**” con el grupo en que se realiza la intervención; o sea es un “control histórico”.

El mejor resultado en el grupo con intervención podría deberse a la identificación de la enfermedad con nuevos métodos que permiten un diagnóstico con la enfermedad menos avanzada y un mejor pronóstico, o a la existencia de nuevos tratamientos coadyuvantes o de apoyo que no se encontraban antes; esto crea un “**sesgo de aprovechamiento**” (se llama sesgo a toda desviación sistemática).

b) La existencia de “**distintos criterios de admisión**”, puede hacer que el grupo con la intervención tenga un mejor pronóstico antes de decidir comenzar con la terapéutica; esto produciría un grosero “**sesgo de susceptibilidad**”.

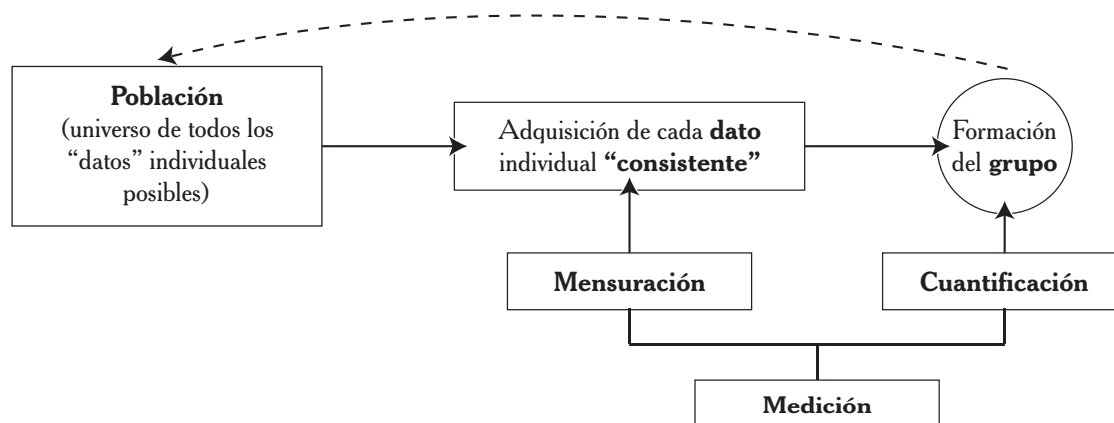


Fig. 1

c) Si el “**método o el criterio para evaluar los resultados son distintos**” (ausencia de objetividad, igualdad de estudios y criterios en el seguimiento y evaluación sin “doble ciego”) ese “**sesgo de detección**” podría ser la causa de la diferencia.

d) A pesar de los criterios anteriores, si **NO** se realiza una asignación del tratamiento de manera que sea “**impredecible**” para el paciente y el médico, en un proceso aleatorio que se llama “**randomización**” (en castellano deberíamos llamarlo “**aleatorización**”), puede aparecer un “**sesgo de susceptibilidad**” difícil de detectar en variables que no se midieron; como asignar la intervención a pacientes más jóvenes o con menos severidad de la enfermedad, etc., que ya configura un grupo con mejor pronóstico desde el momento previo a la intervención.

Por lo tanto además de una “**comparación cuantificada**” de los grupos, es necesario que el procedimiento permita aislar, como diferente, solamente la intervención, para así poder valorar su efecto. Para ello la comparación debe ser “**controlada**”, en el sentido que se puedan balancear en los grupos esas cuatro características, y evitar los “**sesgos**” (desvíos sistemáticos) que harían el ensayo poco atractivo y muy dificultoso para una demostración científica del efecto de la intervención.

Si bien los tres primeros criterios (grupos concurrentes, con criterio de admisión similar, método y/o criterios de evaluación de resultados similares) se puede configurar con un método científico que no implica necesariamente la **randomización**; la cuarta característica de “**asignación impredecible del tratamiento**”, para evitar los “**sesgos de susceptibilidad**” que son más difíciles y a veces imposibles de detectar, solo se puede lograr con un “**proceso de aleatorización**”.

Sin embargo, si utilizamos la **randomización** en la asignación de los distintos grupos, como *subproducto* de este proceso se obtiene la igualdad de las tres primeras características. Ya que entonces necesariamente los grupos deben ser comparados concurrentemente (grupos paralelos), todos los pacientes son diagnosticados con métodos iguales y admitidos con criterios de severidad similares (criterios de inclusión y exclusión), la observación de la respuesta se arregla para períodos iguales de seguimiento, con los mismos criterios objetivos (punto final o resultado) y si es necesario con un procedimiento llamado de “**doble ciego**”, donde ni el paciente ni el médico conoce si está o no con el tratamiento que se desea evaluar, impidiendo los sesgos del médico y del paciente (Figura 2).

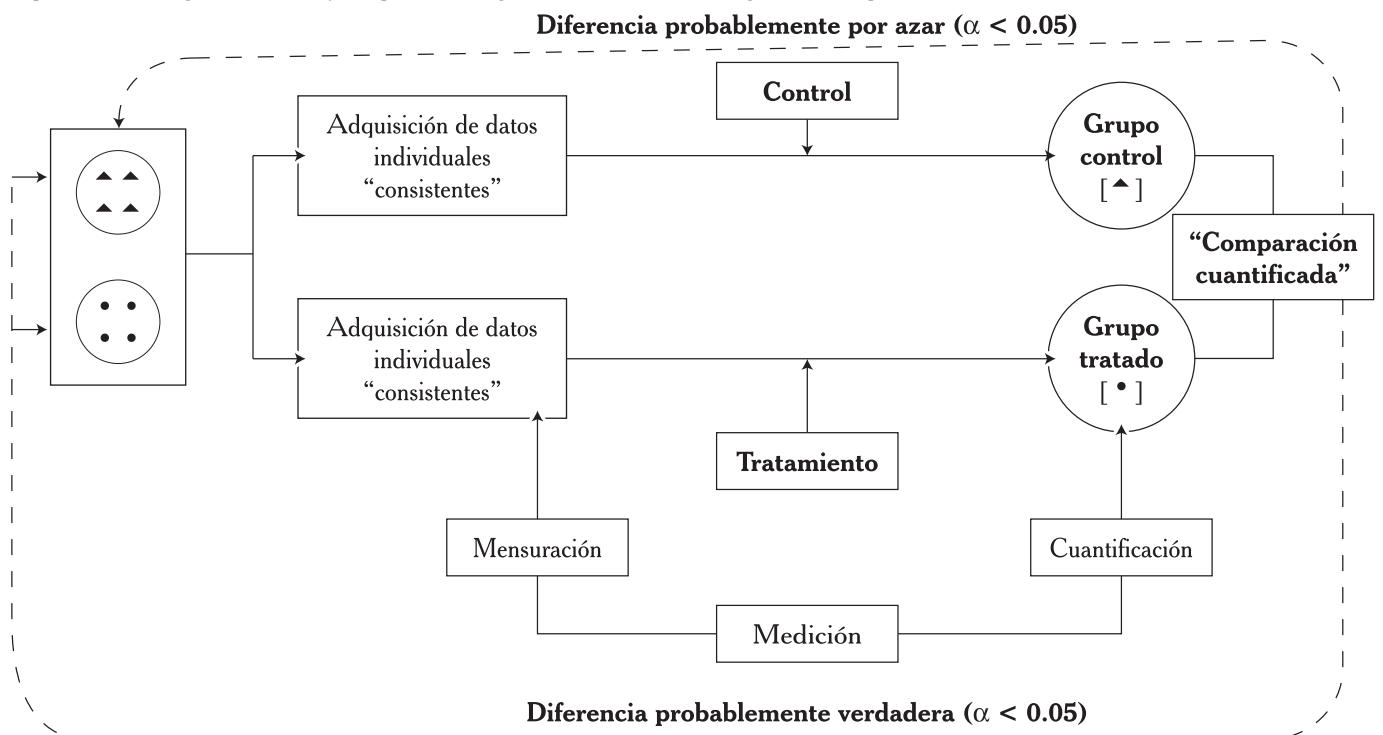
Una vez adquiridos todos los datos individuales “consistentes” de la misma población para el grupo control y el que efectuó la intervención, si al realizar la **comparación cuantificada** hallamos, una diferencia en la variable estudiada, debemos reconocer por un procedimiento de la “**estadística probabilística**” cuál es la posibilidad que se presente la diferencia que encontramos solamente debido al azar, si ambos grupos fueran parte de la misma población (se establece el error Tipo I que se va a aceptar colocando el coeficiente “ $\alpha$ ”, que se establece generalmente como  $< 0.05$  ( $< 5\%$ ), y se calcula el conocido índice “**p**” de significación), y si la intervención no hubiera tenido ningún efecto *estadísticamente significativo*, estaríamos diciendo que tendría una posibilidad mayor del 5% ( $p > 0.05$ ) de que la diferencia obtenida se deba al azar.

O sea la estadística, y en esta situación la estadística probabilística, puede ser definida como un cuerpo de métodos para aprender de la experiencia con muestra de la población; trasladando esa experiencia por inferencia, con cierta aproximación, al resto de situaciones iguales no estudiadas.

Como decíamos para evaluar las diferencias entre dos o más grupos de mediciones, utilizamos el valor de “**p**”, que significa la “**probabilidad**” de obtener un resultado tan extremo como el observado si la diferencia se debe únicamente a las naturales variaciones en la medición o en la respuesta del sujeto; o sea la probabilidad que la desigualdad resulte solamente por azar cuando en realidad forma parte de la misma población.

El cálculo y la definición del valor de “**p**” dependen implícitamente de cuatro conceptos, que serían: medición, dirección, hipótesis de nulidad e hipótesis alternativa.

En el primer paso del cálculo de “**p**” debemos considerar la “**escala que ese utilizó en al medición**”. Si la misma es continua (mm Hg, metro, Kg, etc.) podemos utilizar pruebas para distribuciones normales (distribución de Gauss) o “cuasi” normales como el conocido test “**t**” de Student para datos independientes o apareados, o en distribución continua no paramétrica la prueba de “**Wilcoxon**” para datos independientes o relacionados, también podemos utilizar esta prueba, la prueba de los “**signos**”, o la prueba de “**Mann-Whitney**” para escalas ordinales (de rango), y la prueba de “**Chi-cuadrado ( $X^2$ )**” para escalas nominales; éstas son algunas de las pruebas habitualmente utilizadas.



**Fig. 2**

En segundo lugar debemos “conocer la dirección”, esto significa plantearnos la hipótesis y calcular la probabilidad solamente para conocer si la intervención lo hace mejor o lo hace peor que el grupo control, y a esto se lo llama “*p de un lado*”; o plantear la hipótesis que “no conocemos en qué sentido” se desviará (¿mejor o peor?) del valor control de referencia, y entonces calculamos una “*p de dos lados*”.

Lo que hace la “*p*”, es calcular cuál es la probabilidad que la diferencia hallada se deba solamente al azar, y por lo tanto sea cierta la “hipótesis de nulidad” de las diferencias; ésta es la tercera condición.

Cuando nosotros obtenemos una “*p*” pequeña (por ejemplo  $p = 0.01$ ), sólo podemos afirmar que esa diferencia por azar ocurrirá más raramente (1 de cada 100 estudios realizados en iguales condiciones); a su vez cuando el valor de “*p*” es grande (por ejemplo  $p = 0.40$ ) esa diferencia se presentará frecuentemente por azar (40 de cada 100 estudios).

Pero en realidad si observamos un resultado que es poco probable que ocurra por azar, podemos adoptar dos razonamientos contrapuestos, podríamos seguir pensando que la “hipótesis de nulidad” es correcta y se ha presentado el resultado improbable, o por el contrario que la “hipótesis de nulidad” es falsa y la diferencia existe realmente.

Como se ve, adoptar uno u otro razonamiento no es un problema de la estadística, de la matemática, ni siquiera de la lógica, sino un problema de “opinión”.

Por lo tanto, el planteo de cuan pequeña debería ser la probabilidad que ocurra la diferencia por azar, para aceptar que la “hipótesis de nulidad” está equivocada, y en un cuarto paso aceptar como válida la “hipótesis alternativa” de que la diferencia es real, es un problema de convicción o sea de opinión.

Dependiendo, entre otras cosas, del valor clínico del estudio y la representabilidad de la muestra, algunos colocan el límite de probabilidad para rechazar la nulidad en una  $p < 0.05$ , y otros, o en otras situaciones, utilizan una  $p < 0.01$ .

En realidad lo que indica la “*p*”, es la probabilidad que ocurra un “Error tipo I”, si se concluye que hay una diferencia cuando en realidad ésta no existe. Esto significa que pende sobre nuestras cabezas, como espada de Damocles, la posibilidad de cometer un “Error tipo I” cuando decimos que una intervención es efectiva, claro que la posibilidad de equivocarnos es menor cuando el valor (“*p*”) es más pequeño.

Esta situación es insoluble; cuando establecemos un “límite de significación” para rechazar la hipótesis de nulidad, lo que estamos haciendo es delimitar cuál es el grado de error que aceptaremos como adecuado para los fines prácticos.

Ahora bien, si en “estadística probabilística” podemos graduar cual es nuestra posibilidad de equivocarnos (error tipo I) cuando decimos que hay una diferencia; la situación dialéctica inversa, decir que no existe diferencia porque la probabilidad es  $> 0.05$  ¿puede ser una afirmación errónea?

Véamos un problema: si a 10 pacientes les medimos la presión arterial antes y después de un tratamiento, y encontramos una reducción de la media de las diferencias (*d*) de 3.3 mm Hg y el desvío standard de las diferencias (DSd) es de 5.9 mm Hg, el calculado con el test “*t*” de Student para datos apareados es de  $p \cong 0.1$ ; podría decirse que la intervención no es efectiva. Sin embargo obtenidos iguales resultados ( $d \pm DSd$  de 3.3 mm Hg  $\pm$  5.9 mm Hg), en 30 pacientes en lugar de los 10, la probabilidad “*p* sería  $< 0.005$ , y no hubiéramos expresado ninguna duda en la franca efectividad de la intervención.

Este ejemplo demuestra claramente que cuando decimos que no existe diferencia podemos cometer un error, que se lo denomina “Error tipo II”; aquél que cometemos cuando se concluye que no hay diferencia y en realidad ésta existe.

Se puede calcular la posibilidad de este error por un índice denominado  $\beta$  (beta); los estadígrafos consideran aceptable un “error tipo II”, con una probabilidad  $\beta$  entre 0.1 a 0.2.

Ahora reconocemos de manera incontestable, que penden sobre nuestras cabezas los dos tipos de errores cuando tomamos una decisión sobre si una intervención se diferencia del control. Y entramos en un dilema de hierro, si manteniendo iguales las condiciones del ensayo (número de las muestras, incidencia del evento en la población control y diferencia con la intervención), queremos disminuir el **Error tipo I** y exigimos una significación estadística más pequeña, en el mismo momento aumentamos la posibilidad de cometer un **Error Tipo II**; en forma simétrica si intentamos disminuir el error  $\beta$  o sea aumentar el “Poder del ensayo (que es el complemento aritmético de  $\beta$  (Poder =  $1 - \beta$ )) para demostrar una diferencia, aumenta matemáticamente la posibilidad de cometer un error tipo I (encontrar una diferencia cuando en realidad no existe).

Podemos concluir, que la existencia de una diferencia clínicamente importante entre los grupos investigados no puede ser rechazada por un ensayo clínico de bajo “Poder”, aunque la diferencia no sea estadísticamente significativa.

En el ejemplo desarrollado anteriormente, el poder para demostrar esa diferencia clínica en los 10 pacientes iniciales era sólo de 30%, pero al aumentar a 30 el número de pacientes, el poder sube a un nivel aceptable del 80% y entonces si se demuestra que esa misma diferencia es significativa.

En la literatura médica de los últimos años, además de la habitual publicación del índice de probabilidad de error tipo I, se está considerando con mayor atención la interpretación apropiada de los estudios clínicos negativos. Varias revisiones que se han publicado, demostraron que gran parte de los estudios clínicos incluyen un número reducido de pacientes, para ser capaces de detectar resultados clínicamente importantes; por lo cual las revistas médicas más relevantes comenzaron a exigir el cálculo del índice de error tipo II ( $\beta$ ), o su complemento el Poder del ensayo clínico que se envía a publicar.

El error tipo II, debido a un bajo “Poder” del ensayo clínico, conlleva varios problemas, algunos pueden ser graves, como considerar no efectiva una medicación durante muchos años, hasta que un último ensayo clínico con gran número de pacientes, en general multicéntrico, demuestra un efecto clínico de importancia. Esta situación se presentó una y otra vez en la práctica clínica. En la evaluación de la estreptoquinasa endovenosa en el infarto agudo de miocardio, los megaensayos multinacionales de los últimos años demostraron que los estudios iniciales previos tenían un número de pacientes ridículamente bajo (con bajo poder), para poder demostrar una disminución del 20 al 40 % de la mortalidad en la fase aguda.

Otro de los problemas graves de los estudios con bajo poder, es de ética médica, o de conciencia personal, al someter a los pacientes a los inconvenientes y riesgos de un “ensayo clínico controlado”, que ya de antemano tiene muy escasas posibilidades de demostrar una diferencia clínica importante.

Si se consideran los factores con que se calcula el Poder de un ensayo clínico, es evidente que la única variable que depende de la voluntad del investigador es el tamaño de la muestra o de las muestras. El “*n*” (número de personas en la muestra) tan descuidado y accesorio en el pasado, se ha convertido en una variable cuidadosamente estipulada en el diseño de las actuales investigaciones.

Es conocido o debería ser conocido, que “significación estadística” no necesariamente quiere decir lo mismo que “significación clínica”.

Acabamos de tomar nota que una disminución de la presión arterial que terminó siendo significativa ( $p < 0.05$ ) y que podría tener cierto valor clínico, no tenía significación estadística en los 10 pacientes inicialmente estudiados, claro que ahora podemos reconocer que el Poder para demostrar esa diferencia era bajo (del 30%); otra manera de darse cuenta que retenía un potencial valor clínico sería utilizando un método de estimación estadístico como el “intervalo de Confianza”

Las pruebas de significación tienen la limitación que sólo nos dicen si

una intervención es mejor o peor que otra o un grupo control, sin embargo, el principal propósito de un ensayo clínico sería estimar la magnitud de la mejoría de un tratamiento respecto a otro; necesitamos responder a la pregunta ¿cuánto mejor? Sin embargo debido a que la diferencia en un ensayo tiene una variación o error random con respecto a la diferencia real de toda la población, no se puede contestar a la pregunta ¿cuánto mejor? con una estimación única puntual, sino con un intervalo que reúna los resultados de 90.95 o 99% de todos los resultados “random” posibles. Cuando el intervalo comprende el resultado de 95 de 100 estudios con igual diseño, tipo y número de pacientes, hablaremos de un “intervalo de Confianza del 95%”.

El cálculo de intervalo de confianza para los 10 pacientes iniciales del estudio que estamos considerando era IC95% - 7.5 a + 0.5 mm Hg o sea la diferencia hallada de 3.3 mm Hg era la mejor estimación de una diferencia real de la población que en el 95% de las veces iría entre un leve aumento de 0.9 mm Hg o una interesante disminución de 7.5 mm Hg. Si hubiéramos obtenido esa diferencia (-3.3 mm Hg) con 30 pacientes sería altamente significativa ( $p < 0.05$ ) y por lo tanto el intervalo de Confianza del 95 % se encontraría entre una reducción de -1.1 mm Hg a -5.5 mm Hg (los dos límites del IC 95% siguen la misma dirección, son negativos y no tocan cero).

Por lo tanto ensayos clínicos controlados estadísticamente no significativos, pero que tienen amplios IC 95% resultan inconclusos; ya que un nuevo estudio con mayor número de pacientes y menor error tipo II (mayor Poder), podría demostrar la existencia de un resultado clínicamente relevante.

Una situación en espejo, sería el hallazgo de una pequeña diferencia estadísticamente significativa porque se utilizó un gran número de pacientes (quizás miles), como por ejemplo una disminución de la presión arterial de 0.8 mm Hg con un IC 95% de -0.6 a -1.0 mm Hg. Con estos resultados un laboratorio puede demostrar que su fármaco produce un descenso *estadísticamente significativo* de la presión arterial, pero un clínico también podría decir que esa disminución de magnitud mínima, no tiene ninguna *significación clínica* para sus pacientes, ya que conoce que la simple dieta hiposódica baja la presión, por lo menos, tres veces más.

Por lo cual indefectiblemente, todo el desarrollo estadístico y matemático es siempre dependiente del pensamiento u opinión clínica que se plantea la siguiente pregunta: ¿cuál será el “n” (número) de pacientes en el ensayo que quiero realizar, para encontrar una diferencia con la intervención que me parezca de significado clínico, aceptando cierto error tipo I y II determinados y estimando de la mejor manera posible la probable incidencia del evento a comparar en la población control? O sea planificar un ensayo para demostrar el efecto de una intervención nunca es suficiente por si solo; si al mismo tiempo no se establece cuál es la diferencia que esperamos encontrar, cuál es el error que vamos a aceptar si afirmamos que una diferencia existe (error tipo I) o no existe (error tipo II), y en cuánto estimamos la incidencia del resultado en la población control; fijadas estas condiciones podemos calcular el “n” de cada muestra.

Conociendo la fórmula matemática para establecer el tamaño de la muestra:

$Z_{\alpha/2}$ : valor de una curva normal para un nivel de significación  $\alpha$  de 2 lados (para 0.05 = 1.96, si se utiliza 0.01 = 2.58).

$Z_{\beta}$ : valor de una curva normal para un nivel de probabilidad  $\beta$  de un lado (para 0.20 = 0.84, si se utiliza 0.1 = 1.28).

Pi: proporción de los que responden en el grupo con intervención.

Pc: proporción de los que responden en el grupo control.

P:  $(P_i + P_c)/2$ , cuando ambos grupos tienen igual número.

$\delta$ : es el “desvío standard” del promedio de los desvíos standard en el grupo control.

$\gamma$ : diferencias de las medias del grupo con intervención y del grupo control.

## Variables Dicotómicas

$$n = \left( \frac{Z_{\alpha/2} \sqrt{2P(1-P)} + Z_{\beta} \sqrt{P_i(1-P_i) + P_c(1-P_c)}}{P_i - P_c} \right)^2$$

## Variable Continuas

$$n = \frac{2\delta^2 (Z_{\alpha/2} + Z_{\beta})^2}{\gamma^2}$$

## ASUNCIONES

- El tamaño de las dos muestras son iguales (n es común para cada muestra).
- Las muestras son randomizadas y no relacionadas (los sujetos del grupo control no están “apareados” al grupo con intervención).
- Las variables continuas tienen una distribución normal con varianza conocida.
- Las variables dicotómicas tienen una distribución binomial aproximada a la distribución normal (esto es verdad cuando  $nP$  y  $n(1-P)$  son cada uno mayores de cinco).
- La varianza en ambos grupos, control y con intervención, se pueden asumir que son iguales.
- No se realizan pruebas múltiples de significación.

**Por lo tanto el tamaño de la muestra para comparar dos grupos de pacientes va a depender de los siguientes factores:**

- ¿Cuál es el tipo de resultado que se está evaluando?. Si el resultado es nominal o mejor dicho binomial dicotómico (por ejemplo vida o muerte) se utiliza una de las fórmulas, si el resultado es continuo se utiliza la otra fórmula para el cálculo “n” de cada muestra.
- ¿Cuál es el riesgo de error que se va a aceptar? Por ejemplo:  
2a): Error tipo I  $\alpha/2$  0.05 = “1.96” de desviación en la curva normal (2 lados).  
2b): Error tipo II  $\beta$  0.2 = “0.84” de desviación de la curva normal (1 lado).
- ¿Cuál sería una diferencia clínicamente importante entre el grupo control y el grupo al que se realizó la intervención, según su opinión?
- ¿Cuál es la variación de la variable en estudio en el grupo control? Se considera el promedio de los desvíos standards conocidos de grupo control, cuando se analizan variables continuas.
- ¿Cuál es la relación del número de pacientes entre el grupo que se realizó la intervención y el grupo control? La relación 1/1 entre dos grupos, es la que utiliza el número total menor, y es la que se usa en la mayoría de los estudios.

Si se observan las fórmulas anteriores, el “n” necesariamente aumenta si se disminuyen el error tipo I ( $\alpha/2$ ), el error tipo II ( $\beta$ ) la diferencia que se quiere buscar ( $P_i - P_c$  o  $\gamma$ ), disminuye la proporción de evento en el grupo control ( $P_c$ ), y se aumenta la dispersión ( $\delta$ ).

Cuando se utilizan variables binomiales dicotómicas los eventos solo ocurren en algunos de los pacientes, ya sea en el grupo control o con intervención (solo algunos tienen un accidente cerebrovascular cuando el punto final es el evento neurológico), en cambio se mide el evento o resultado en todas las personas cuando la variable es continua (medición hasta

centígramos en cada participante cuando el punto final es el peso corporal, etc.); en esta última situación por lo tanto se necesita un “n” de muestras mucho menor que en la situación anterior.

Por todas estas consideraciones, solo se necesitan algunas decenas de pacientes para demostrar la disminución de la Presión Arterial con un fármaco en un ensayo clínico, que además se realiza en un corto tiempo (variable continua), pero sin embargo para demostrar que el mismo fármaco disminuye la incidencia de eventos cerebrovasculares se necesitan alrededor de 10 000 pacientes durante 3 a 5 años de seguimiento (variable binominal dicotómica con incidencia baja de eventos en el grupo control), y aún ensayos tan numerosos como el anterior no demostraron disminución de la mortalidad total.

Por suerte, para conocer rápidamente el tamaño de las muestras no es imprescindible efectuar los cálculos matemáticos, porque podemos aplicar los nomogramas diseñados por Mark Y. Young y col. (ver bibliografía)

Los nomogramas fueron diseñados para un nivel de significación  $\alpha=0.05$  (para dos lados) y  $\beta=0.20$  (un lado) y son dos: el realizado para variables dicotómicas, en el eje horizontal se localiza el porcentaje de cambio que se desea detectar, se extiende una línea vertical hasta interceptar la diagonal correspondiente a la frecuencia de eventos (en porcentaje) en el grupo control, y por fin llevado horizontalmente hasta el eje vertical muestra el tamaño requerido para cada uno de los grupos (control e intervención).

El diseño para variables continuas es similar, salvo que en el eje horizontal se localiza la diferencia de las medias de los dos grupos que se considera clínicamente importante detectar y la línea vertical intercepta las diagonales correspondientes al desvío standard promedio de grupo control, y por fin llevado horizontalmente hasta el eje vertical, indica el tamaño requerido para cada grupo.

Hasta llegar a este punto del desarrollo metodológico y estadístico (pruebas de significación), una asunción importante que hemos mantenido implícita, es que los “ensayos clínicos controlados” estaban conduciendo para probar únicamente una hipótesis.

Cuando en un estudio se utilizan múltiples puntos finales surgen dos problemas, uno metodológico y otro estadístico.

Cuando se utilizan varios resultados finales éstos deberían tener una relación metodológica, aunque sea implícita, a una teoría que los abarca, o en sí mismo estar encadenados formando una teoría global. Por lo cual si algunos puntos finales son probados y otros rechazados la interpretación es ambigua, y declarar que la intervención se diferencia o no del control dependerá de un comentario u opinión a posteriori (ad hoc) de interpretación subjetiva. En realidad si todos los puntos finales tienen similar jerarquía, y se es exigente en la metodológica, al rechazar alguno de los resultados se refutaría la teoría que los sustenta.

El problema estadístico que se plantea, es que la utilización de test de significancia separados para la comparación de cada punto final (que se llama error de comparación), aumenta el riesgo de la aparición de algún resultado falsamente positivo (a esto se lo llama error de experimento).

Ya hemos discutido extensamente que aceptamos como razonable un error tipo I por comparación  $<0.05$ , pero si efectuamos 20 comparaciones de puntos finales distintos, como cada comparación tiene  $1/20$  probabilidad de error ( $0.05 = 5\%$ ), al multiplicar por las 20 comparaciones ( $20 \times 1/20 = 20/20 = 1$ ) tenemos la posibilidad que una de las comparaciones presente una diferencia estadísticamente significativa del control aunque en la realidad no existiera ninguna diferencia.

Una solución simple a este problema, es multiplicar cada valor de P (error por comparación) por el número de puntos finales que se está analizando, si este valor de P “ajustado” (método de Bonferroni) es  $<0.05$  se acepta que la intervención en forma global con todos sus puntos finales es distinta al grupo control error por experimento).

Por ejemplo, si se evaluarán 10 mediciones distintas de resultados y las

10 separadas tuvieron una  $P = 0.02$  (error por comparación) el valor de P ajustado por Bonferroni será  $P A = 0.2$  ( $10 \times 0.02 = 0.20$ ) indicando que no hay diferencia entre los grupos.

Sin embargo este simple aumento del valor de P es una sobrecorrección para puntos finales múltiples, particularmente si las diferentes mediciones de resultados de los pacientes están fuertemente asociados entre sí. En el ejemplo anterior si los 10 puntos finales estuvieran perfectamente correlacionados - si se da 1 de las mediciones se puede predecir exactamente, sin error, los otros 9 resultados -, en esta situación los resultados de las 10 pruebas serán idénticos, y podría seleccionarse un solo resultado para aceptar o rechazar el efecto de la intervención; en este ejemplo uno de los resultados con  $P: 0.02$  debería ser apropiado para juzgar que la intervención es estadísticamente significativa, sin embargo el ajuste de Bonferroni dará un P “ajustado” =  $0.20$  indicando una falsa ausencia de diferencia.

La proliferación de puntos finales es una forma poco satisfactoria de interpretar los resultados de los ensayos clínicos. Es preferible reducir el número de puntos finales medidos y especificar por adelantado, en forma prospectiva, el punto final prioritario (punto final principal) que se utilizará para aceptar o rechazar el resultado del ensayo clínico controlado y los puntos finales secundarios.

De esta manera el punto final principal se puede analizar estadísticamente en forma única, no ambigua, con un método estadístico simple y de gran valor probatorio. Los otros puntos finales pueden ser analizados secundariamente, con el valor de P “ajustado” (Bonferroni y otros métodos) para comparaciones múltiples.

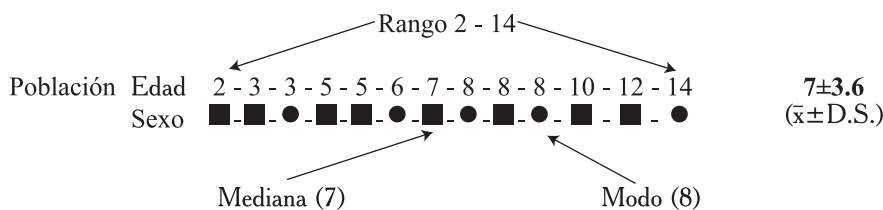
Es de suma importancia insistir que el punto final primario del ensayo clínico controlado debe ser especificado en forma prospectiva en el diseño del protocolo del estudio. La selección posterior (ad hoc) de puntos finales buscando donde la diferencia de la intervención es más importante, es una engañosa e inaceptable estrategia que invariablemente sobrevalúa la diferencia del tratamiento, se dice que siempre se puede conseguir una variable significativa si se “*torturan a los datos*”.

Es importante para los clínicos practicantes participar en el diseño, elaboración y realización de “ensayos clínicos controlados”, que respondan a las preguntas que surgen de practicar la medicina, porque como concluye Bertolt Brecht:

***Observa mal el que no sabe hacer nada  
con lo que haya observado.  
Para el manzano tiene un ojo más agudo  
el cultivador de frutas que el paseante,  
pero no ve exactamente al hombre quién  
no sepa que el hombre es el destino del hombre.***

PARÁMETROS ESTADÍSTICOS QUE DEBEN SELECCIONARSE ACORDE CON LA ESCALA DE MEDICIÓN

Escala		Parámetros	Ejemplo
Nominal		<b>Frecuencia:</b> es el valor que corresponde a la cantidad de veces que el dato ha sido contado.	8■ y 5●
		<b>Modo:</b> es el valor que más frecuentemente se presenta en la medición.	8■
		<b>Razón:</b> es una división, donde dividendo y divisor pueden ser cualquier número.	■/●=1.6
		<b>Proporción:</b> es una razón, en la cual el dividendo es uno de los sumandos cualquiera de una suma y el divisor el total de la suma.	$\frac{■}{●+■} = 0,615$
Ordinal	Estimación por puntos	<b>Modo:</b> (ya definido)	8
	Medidas de dispersión	<b>Mediana:</b> es aquel valor de los datos que está ubicado en una posición tal que tanto por delante, como por detrás de él, queda el mismo número de observaciones	7
	Estimación por intervalos	<b>Rango:</b> intervalo comprendido entre los valores mínimo y máximo que alcanzan los datos.	2 - 14
Intervalo y proporción, (*)	Estimación por puntos	<b>Media Aritmética (promedio):</b> La suma de los datos( $\sum x$ ) dividido el número (n) de observaciones: $\bar{x} = \frac{\sum x}{n}$	7
	Medidas de dispersión	Es una "distribución normal" la dispersión de 1 D.S alrededor de la Media comprende 67% de la población ( $\pm 2$ D.S. comprende el 95% de la población). <b>Varianza:</b> $\frac{\sum (x - \bar{x})^2}{n - 1}$ <b>Desvío Standard:</b> $\sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$	V: 12.7 D.S.: 3.6
	Estimación de intervalos	Media Desvío Standard ( $\bar{x} \pm$ D.S.)	$7 \pm 3.6$



Estadística NO PARAMETRICA: Escalas Nominal, Ordinal, Intervalo y Proporción que NO tienen “distribución normal”.  
PARAMETRICA: Escalas de intervalo y Proporción si existe “distribución normal”.

**Error Standard (E.S.)** =  $\frac{D.S.}{\sqrt{n}} = \frac{3.6}{\sqrt{13}} = 1.0$

Es el “Desvío Standard de la Media” y la  $\bar{x} \pm 2$  E.S. ( $7.0 \pm 2.0$ ) indican que aproximadamente el 95% de las medias de distintas muestras extraídas de la misma población oscilarán entre 5 a 9 años de edad (sería el “intervalo de confianza” del 95% de la estimación de la Media de la Población).

## **BIBLIOGRAFÍA**

- Bailar III JC, Mosteller F. Medical Uses of Statistics. Massachusetts Medical Society. 1992
- Pocock SJ. Clinicals Trials. A Practical Approach. John Wiley & Sons. 1993
- Shapiro SH, Louis TA. Clinicals Trials. Issues and Approaches. Marcel Dekker, Inc. 1983
- Kahn HA, Sempos CT. Statistical Methods in Epidemiology. Oxford University Press. 1989
- Shott S. Statistics for Health Professionals. W. B. Saunders. 1990
- Dawson-Saunders B, Trapp RG. Basic and Clinical Biostatistics. Appleton & Lange. 1990
- Swinscow TDV. Statistics at Square One. BMJ Group. 1999
- Wassertheil-Smoller S. Biostatistics and Epidemiology. A primer for health professionals. Springer-Verlag. 1990
- Young MY, Bresnitz FA, Strom BL. Sample size nomograms for interpreting negative clinical studies. Annals of Internal Medicine. 1983;99: 248-51.
-

### Preguntas de Evaluación

Las siguientes preguntas corresponden al presente fascículo, "El ABC de los ensayos clínicos controlados" del MODULO 6: INTRODUCCION A LA INVESTIGACION.

El médico deberá registrar en él las respuestas elegidas y remitir la hoja por correo o fax al Comité de Educación Médica Continua, Sociedad Argentina de Urología, Pasaje de la Cárcova 3526, (1172) Buenos Aires. Tel./fax: 4963-8521/4336/4337.

El requisito para aprobar el módulo consistirá en contestar correctamente por lo menos el 75% del total de las preguntas del módulo, para ello tendrá un máximo de 60 días a partir del momento en que recibió el fascículo. Luego de ese lapso en uno de los próximos fascículos figurarán las respuestas correctas, de esta manera el médico podrá realizar su autoevaluación e ir comprobando los resultados de su aprendizaje.

Cualquier consulta y/o aclaración en relación con las preguntas, dirigirse a la dirección indicada previamente.

- 1.- En un estudio clínico controlado, el grupo en el que se realiza la intervención se puede comparar con:
  - a) ..... Control no tratado (no intervención)
  - b) ..... Control con placebo (intervención simulada)
  - c) ..... Control con tratamiento alternativo (otra intervención)
- 2.- La mayoría de los datos y juicios de la práctica clínica son comparaciones:
  - a) ..... Cuantitativas
  - b) ..... Cualitativas
  - c) ..... Numéricas
  - d) ..... Algebraicas
- 3.- El índice de Apgar, instituido para conocer la vitalidad del recién nacido, es una:
  - a) ..... Escala ordinal
  - b) ..... Escala de proporciones
  - c) ..... Escala nominal
  - d) ..... Escala binominal
- 4.- El Error Tipo I en la estadística probabilística indica:
  - a) ..... Cual es la posibilidad que se presente la diferencia por la intervención
  - b) ..... Que podemos cometer un error cuando se concluye que no hay diferencia y en realidad esta existe.
  - c) ..... Que podemos cometer un error cuando se concluye que hay diferencia y en realidad esta no existe.
  - d) ..... Cual es la posibilidad que no se presente la diferencia por la intervención
- 5.- El Error Tipo II en la estadística probabilística indica:
  - a) ..... Cual es la posibilidad que se presente la diferencia por la intervención
  - b) ..... Que podemos cometer un error cuando se concluye que no hay diferencia y en realidad esta existe.
  - c) ..... Que podemos cometer un error cuando se concluye que hay diferencia y en realidad esta no existe.
  - d) ..... Cual es la posibilidad que no se presente la diferencia por la intervención
- 6.- Para el número de pacientes en la muestra (n) de un ensayo clínico controlado, se debe estimar:
  - a) ..... La diferencia de significación clínica con la intervención
  - b) ..... Que Error Tipo I y II voy a aceptar
  - c) ..... La incidencia del punto final (resultado) en la población control
  - d) ..... Todos los anteriores
- 7.- El punto final con el cual se debe calcular el número de pacientes de la muestra y se va a interpretar los resultados de un ensayo clínico, es preferible que sea:
  - a) ..... Punto final único o principal
  - b) ..... Punto final principal y secundario
  - c) ..... Todos los puntos finales que se planifiquen
  - d) ..... Todos los anteriores son ciertos

Apellido y Nombre: ..... N° inscripto: .....

Dirección: ..... Código: .....

Ciudad: ..... Provincia: .....

Tel. ó fax: ..... E-mail: .....