La fiabilidad de las mediciones clínicas: el análisis de concordancia para variables numéricas

Pita Fernández S, Pértegas Díaz S,
Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario-Universitario Juan Canalejo. A Coruña (España)
Actualización 12/01/2004.

Introducción

La medición es un proceso inherente tanto a la práctica como a la investigación clínica. Mientras que algunas variables son relativamente sencillas de medir (como el peso o la tensión arterial) otras comportan cierto grado de subjetividad que hace especialmente difícil su medición, como la intensidad de dolor o el concepto de calidad de vida. En cualquier caso, el proceso de medición conlleva siempre algún grado de error. Existen factores asociados a los individuos, al observador o al instrumento de medida que pueden influir en la variación de las mediciones [1,2]. En la medida de la temperatura corporal, por ejemplo, pueden aparecer errores en el registro debidos tanto al estado del paciente, como a defectos en el termómetro utilizado o a la objetividad del observador.

Cualquier estudio epidemiológico debe garantizar la calidad de sus mediciones, no sólo porque condicionará en gran medida la validez de sus conclusiones, sino por la importancia de las decisiones clínicas que se apoyen en esa investigación[3]. La calidad de una medida depende tanto de su validez como de su fiabilidad[1,4]. Mientras que la validez expresa el grado en el que realmente se mide el fenómeno de interés, la fiabilidad indica hasta qué punto se obtienen los mismos valores al efectuar la medición en más de una ocasión, bajo condiciones similares. El que una medida sea muy precisa no implica, sin embargo, que sea necesariamente válida. Así, si se realizan dos mediciones consecutivas de la presión arterial de un paciente con un esfigmomanómetro mal calibrado los valores obtenidos seguramente serán parecidos, aunque totalmente inexactos.

En los estudios que tratan de evaluar la validez de una medida se comparan sus resultados con los obtenidos mediante una prueba de referencia (gold standard) que se sabe válida y fiable para la medición del fenómeno de interés[5]. Cuando el objetivo se centra en la fiabilidad de una medición, se repite el proceso de medida para evaluar la concordancia entre las distintas mediciones. En un estudio de la fiabilidad pueden valorarse los siguientes aspectos[1]:

a. **Repetibilidad:** indica hasta qué punto un instrumento proporciona resultados similares cuando se aplica a una misma persona en más de una ocasión, pero en idénticas condiciones.
b. **Concordancia intraobservador:** tiene por objetivo evaluar el grado de consistencia al efectuar la medición de un observador consigo mismo.
c. **Concordancia interobservador:** se refiere a la consistencia entre dos observadores distintos cuando evalúan una misma medida en un mismo individuo.
d. **Concordancia entre métodos de medición:** cuando existen diferentes métodos de medida para un mismo fenómeno, es interesante estudiar hasta qué punto los resultados obtenidos con ambos instrumentos son equivalentes.

La concordancia entre variables es de sumo interés en la práctica clínica habitual[6-9]. La concordancia entre mediciones puede alterarse no sólo por la variabilidad de los observadores, sino por la variabilidad del instrumento de medida o por el propio proceso a medir si se realiza en momentos diferentes. Las técnicas de análisis de la concordancia dependen del tipo de variable a estudiar. El índice estadístico más utilizado, para el caso de variables cualitativas, es el coeficiente kappa[10]. Si las variables son cuantitativas, se utiliza habitualmente el coeficiente de correlación intraclase[2,6,11]. El concepto básico subyacente del coeficiente de correlación intraclase fue introducido originalmente por Fisher como una formulación especial de la r de Pearson, basándose en un modelo de análisis de la varianza[12]. Las dificultades para interpretar desde el punto de vista clínico los valores de este coeficiente y otras desventajas metodológicas han hecho que algunos autores propongan métodos alternativos para estudiar la concordancia de este tipo de variables. Así, Bland y Altman (1995)[13] proponen un método gráfico y muy sencillo, basado en el análisis de las diferencias individuales, que permite determinar los límites de

concordancia y visualizar de forma gráfica las discrepancias observadas. Recientemente, otros métodos de análisis de concordancia han sido propuestos[14-16]. A continuación, se procederá a una descripción detallada de algunas de estas técnicas de análisis.

El coeficiente de correlación intraclase

Para el caso de variables cuantitativas, es frecuente que el análisis de la concordancia se aborde mediante técnicas estadísticas inapropiadas. Con frecuencia ha sido utilizado el cálculo del coeficiente de correlación de lineal (r) de Pearson como índice de concordancia. Sin embargo, ésta no resulta una medida adecuada del grado de acuerdo entre dos mediciones, ya que si dos instrumentos miden sistemáticamente cantidades diferentes uno del otro, la correlación puede ser perfecta (r=1), a pesar de que la concordancia sea nula. Consideremos como ejemplo los datos de la Tabla 1, en la que se comparan las mediciones de tensión arterial con dos instrumentos diferentes. El instrumento B mide sistemáticamente 1mm Hg más que el instrumento A. Al representar gráficamente la correlación entre ambas mediciones, se objetiva que la correlación es la máxima posible (r=1), a pesar de que ninguna de las mediciones ha concordado (Figura 1). No se debe olvidar que el coeficiente de correlación de Pearson no proporciona información sobre el acuerdo observado, y solamente mide la asociación lineal entre dos variables[17]. Así mismo, al calcularse a partir de los pares ordenados de mediciones, si varía el orden también cambia el valor del coeficiente[17], mientras que un cambio en las escalas de medida no afecta a la correlación pero sí afecta a la concordancia. A su vez, debemos mencionar que la idea de que si el coeficiente de correlación entre dos medidas es significativamente diferente de cero la fiabilidad es buena, es incorrecto. El coeficiente de correlación lineal puede ser muy pequeño y resultar significativo si el tamaño muestral es suficientemente grande. Por último, tampoco la comparación de medias mediante un test t de Student con datos apareados es una técnica adecuada para este tipo de análisis[1].

Desde el punto de vista matemático, el índice más apropiado para cuantificar la concordancia entre diferentes mediciones de una variable numérica es el llamado coeficiente de correlación intraclase (CCI)[2,6,11]. Dicho coeficiente estima el promedio de las correlaciones entre todas las posibles ordenaciones de los pares de observaciones disponibles y, por lo tanto, evita el problema de la dependencia del orden del coeficiente de correlación. Así mismo, extiende su uso al caso en el que se disponga de más de dos observaciones por sujeto.

Sin embargo, una de las principales limitaciones del CCI es la dificultad de su cálculo, ya que debe ser estimado de distintas formas dependiendo del diseño del estudio[18]. La forma de cálculo más habitual se basa en un modelo de análisis de la varianza (ANOVA) con medidas repetidas (Tabla 2). La idea es que la variabilidad total de las mediciones se puede descomponer en dos componentes: la variabilidad debida a las diferencias entre los distintos sujetos y la debida a las diferencias entre las medidas para cada sujeto. Esta última, a su vez, depende de la variabilidad entre observaciones y una variabilidad residual o aleatoria asociada al error que conlleva toda medición. El CCI se define entonces como la proporción de la variabilidad total que se debe a la variabilidad de los sujetos.

En la actualidad el valor del CCI puede obtenerse de modo directo con algunos programas informáticos como el SPSS. Otra forma sencilla de obtener el valor del CCI es a partir de una tabla ANOVA para medidas repetidas. Como ejemplo, en la Tabla 3 se representan los datos de un estudio hipotético en el que se tomó la tensión arterial sistólica en 30 pacientes utilizando dos métodos diferentes. Si se representan gráficamente estos datos, indicando el coeficiente de correlación r=0,997 una asociación prácticamente lineal (Figura 2). A partir de la tabla ANOVA correspondiente (Tabla 4), el CCI se puede calcular como:

$$CCI = \frac{k \cdot SC_{ENTRE} - SS_{TOTAL}}{(k-1) \cdot SS_{TOTAL}}$$

donde k es el número de observaciones que se toman en cada sujeto. En el ejemplo:

$$CCI = \frac{k \cdot SS_{ENTRE} - SS_{TOTAL}}{(k-1) \cdot SS_{TOTAL}} = \frac{2 \times 73597{,}683 - 73940{,}183}{73940{,}183} = 0{,}991$$

Como toda proporción, los valores del CCI pueden oscilar entre 0 y 1, de modo que la máxima concordancia posible corresponde a un valor de CCI=1. En este caso, toda la variabilidad observada se explicaría por las diferencias entre sujetos y no por las diferencias entre los métodos de medición o los diferentes observadores. Por otro lado, el valor CCI=0 se obtiene cuando la concordancia observada es igual a la que se esperaría que ocurriera sólo por azar. A la hora de interpretar los valores del CCI, toda clasificación es subjetiva, si bien resulta útil disponer de una clasificación como la que proponen otros autores[6] (Tabla 5).

Hasta ahora, se ha presentado la forma más habitual de cálculo del CCI. Para su cálculo en otras situaciones, así como para la obtención de intervalos de confianza, puede recurrirse a referencias más especializadas[6,18,19].

A pesar de ser la medida de concordancia más adecuada par el caso de variables numéricas, el CCI presenta ciertas limitaciones. Junto a la dificultad inherente a su cálculo, el hecho de que se trate de una prueba paramétrica limita su uso al caso en el que se verifiquen las hipótesis necesarias. A saber: variables distribuidas según una normal, igualdad de varianzas e independencia entre los errores de cada observador. Así mismo, el valor del CCI depende en gran medida de la variabilidad de los valores observados: cuanto más homogénea sea la muestra estudiada, más bajo tenderá a ser el valor del CCI. Pero quizás lo que más ha limitado la difusión del uso del CCI en la literatura médica es la carencia de interpretación clínica, que ha propiciado la aparición de otros métodos de análisis, mucho más intuitivo y fácilmente interpretables, que se exponen a continuación.

Análisis de las diferencias individuales: método de Bland y Altman

Un sencillo procedimiento gráfico para evaluar la concordancia entre dos sistemas de medida es el propuesto por Bland y Altman[13]. Dicho procedimiento consiste en representar gráficamente las diferencias entre dos mediciones frente a su media. Utilizaremos para ilustrar dicha metodología las mediciones de tensión arterial sistólica obtenidas por medio de un esfigmomanómetro de mercurio en el brazo y la obtenida por medio de un monitor autoinflable electrónico en el dedo índice. Dichas mediciones fueron realizadas a 159 alumnos de las escuelas universitarias de enfermería de A Coruña y Ferrol.

La correlación existente entre ambas mediciones ($r=0,202$; $p<0.05$) se presenta en la Figura 3, donde se objetiva una correlación positiva y estadísticamente diferente de cero. Si se representan en un diagrama de dispersión en el eje de ordenadas las diferencias entre ambos procedimientos, y en el eje de abscisas el promedio de ambas mediciones, se obtiene la Figura 4. En dicha figura objetivamos que muy pocas mediciones han concordado (diferencia igual a cero). Por el contrario, la mayoría de las veces el aparato electrónico digital ha proporcionado valores superiores al esfigmomanómetro de mercurio, de hecho la media de dichas diferencias (electrónico – mercurio) es positiva (22,5). Además, dicha gráfica permite objetivar que la discordancia se incrementa a medida que se obtienen valores más elevados de TAS. Por lo tanto, las diferencias no son homogéneas a lo largo del eje horizontal. La distribución de las diferencias se puede a su vez valorar realizando un histograma de las mismas (Figura 5), donde se objetiva claramente el predominio de diferencias positivas mostrando por lo tanto cómo el aparato electrónico claramente proporciona valores más elevados que el esfigmomanómetro de mercurio. Es evidente por lo tanto que la falta de homogeneidad de las diferencias, así como la magnitud de la misma, invalida la utilización del monitor digital del dedo índice como método en este estudio para tomar la tensión arterial.

Un aspecto muy importante de la metodología de Bland y Altman es que proporciona además unos límites de concordancia a partir del cálculo del intervalo de confianza para la diferencia de dos mediciones. Como es bien sabido, el intervalo de dos desviaciones estándar alrededor de la media de las diferencias incluye el 95% de las diferencias observadas. Estos valores deben compararse con los límites de concordancia que se hayan establecido previamente al inicio del estudio para concluir si las diferencias observadas son o no clínicamente relevantes.

Otros métodos de análisis

Distintos autores han propuesto algunas técnicas alternativas para el análisis de la concordancia para mediciones numéricas, principalmente desde un punto de vista gráfico, que vienen a complementar el método de Bland y Altman[14-16]. Una propuesta sencilla y muy reciente se basa en construir una gráfica, similar a las de Kaplan-Meier que se utilizan en el análisis de supervivencia, donde en el eje horizontal se representa la diferencia absoluta entre dos mediciones para cada sujeto y en el eje vertical la proporción de casos en los que las discrepancias igualan al menos cada una de las diferencias observadas[16]. La gráfica se construye así igual que en un análisis de supervivencia, donde ningún caso es censurado, y el papel de la variable "tiempo" lo juega aquí la diferencia absoluta entre las mediciones.

Si retomamos el ejemplo anterior (Tabla 3), en la Figura 6 se muestra el análisis de las diferencias individuales según la metodología de Bland y Altman. Del gráfico se deduce claramente que el método B proporciona con frecuencia valores más bajos de tensión arterial, con una diferencia media de -3,23. De modo complementario, en la Tabla 6 se muestra la magnitud, en términos absolutos, de las dos mediciones de tensión arterial en cada paciente, así como el porcentaje acumulado de casos en los que se supera cada una de estas diferencias. A partir de estos datos puede construirse fácilmente la Figura 7, en la que se muestra el desacuerdo existente entre ambos métodos. Dicho gráfico permite evaluar si la diferencia tiene o no alguna relevancia desde un punto de vista clínico. Así, por ejemplo, si establecemos como aceptable un margen de error entre las mediciones de 2 mmHg se obtiene un porcentaje de acuerdo de un 20%, mientras que la concordancia alcanza el 90% si se admiten diferencias de hasta 8 mmHg, lo cual resulta aceptable desde un punto de vista clínico.

Al igual que el método propuesto por Bland y Altman, el principal atractivo de esta alternativa es que permite expresar sus resultados gráficamente, relacionándolos con unos límites de concordancia preestablecidos según criterios clínicos antes del estudio, lo que los hace especialmente atractivos para los profesionales sanitarios. Así mismo, permite contrastar si el grado de acuerdo depende de alguna otra covariable, construyendo gráficos independientes, uno para cada nivel de la variable. Incluso es posible utilizar el test del log-rank para testar la existencia de diferencias significativas entre esas curvas. No obstante, al trabajar con las diferencias absolutas, este método, al contrario que el de Bland y Altman, no permite observar si existe una diferencia sistemática a favor de alguna de las dos técnicas u observadores, y tampoco comprobar si la magnitud de dicha diferencia se modifica en relación a la magnitud de la medida.

En definitiva, el problema del análisis de la concordancia en el caso de variables numéricas puede abordarse según diferentes metodologías. Lejos de recomendar el uso estándar de alguna de estas técnicas, más bien deben considerarse como métodos de análisis que ofrecen información complementaria. En cualquier caso, es conveniente insistir una vez más en la conveniencia de garantizar la validez y fiabilidad de los instrumentos de medida utilizados habitualmente en la práctica e investigación clínica. No debemos olvidar que un estudio bien diseñado, ejecutado y analizado fracasará si la información que se obtiene es inexacta o poco fiable[1].

Bibliografía

1. Argimon Pallán JM, Jiménez Vill J. Métodos de investigación clínica y epidemiológica. 2ª ed. Madrid: Harcorurt; 2000.
2. Hernández Aguado I, Porta Serra M, Miralles M, García Benavides F, Bolúmar F. La cuantificación de la variabilidad en las observaciones clínicas. Med Clin (Barc) 1990; 95: 424-429. [Medline]
3. Sackett DL. A primer on the precision and accuracy of the clinical examination. JAMA 1992; 267: 2638-2644. [Medline]
4. Latour J, Abraira V, Cabello JB, López Sánchez J. Métodos de investigación en cardiología clínica (IV). Las mediciones en clínicas en cardiología: validez y errores de medición. Rev Esp Cardiol 1997; 50(2): 117-128. [Medline] [Texto completo]
5. Pita Fernández S, Pértega Díaz S. Pruebas diagnósticas. Cad Aten Primaria 2003; 10: 120-124. [Texto completo]
6. Fleiss JL. The design and analysis of clinical experiments. New York: Wiley; 1986-

7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174. [Medline]
8. Ripolles Orti M, Martín Rioboo E, Díaz Moreno A, Aranguren Baena B, Murcia Simón M, Toledano Medina A, Fonseca Del Pozo FJ. Concordancia en la medición de presión arterial entre diferentes profesionales sanitarios. ¿Son fiables los esfigmomanómetros de mercurio? Aten Primaria 2001; 27(4): 234-43. [Medline] [Texto completo]
9. Divison JA, Carbayo J, Sanchis C, Artigao LM. Concordancia entre las automedidas domiciliarias y la monitorización ambulatoria de la presión arterial. Med Clin (Barc). 2001; 116(19): 759. [Medline]
10. López de Ullibarri Galparsoro I, Pita Fernández S. Medidas de concordancia: el índice Kappa. Cad Aten Primaria 1999; 6: 169-171.
11. Prieto L, Lamarca R, Casado A. La evaluación de la fiabilidad en las observaciones clínicas: el coeficiente de correlación intraclase. Med Clin 1998: 110(4): 142-145. [Medline]
12. Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. J Clin Epidemiol. 1991; 44(4-5): 381-90. [Medline]
13. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-310. [Medline]
14. Monti KL. Folded empirical distribution function curves – mountain plots. Am Stat 1995; 49: 342-345. [ISI]
15. Krouwer JS, Monti KL. A simple, graphical method to evaluate laboratory assays. Eur J Clin Chem Clin Biochem 1995; 33: 525-527. [Medline]
16. Luiz RR, Costa JL, Kale PL, Werneck GL. Assessment of agreement of a quantitative variable: a new graphical approach. J Clin Epidemiol 2003; 56(10): 963-967. [Medline]
17. Bland JM, Altman DG. Measurement error and correlation coefficients. BMJ 1996; 313: 41-42. [Medline] [Texto completo]
18. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966; 19: 3-11. [Medline]
19. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996; 1: 30-46.

Tabla 1. Ejemplo teórico sobre mediciones de Tensión Arterial Sistólica con dos instrumentos diferentes.

Instrumento A	Instrumento B
110	111
120	121
130	131
140	141
150	151
160	161
170	171
180	181
190	191
200	201

Tabla 2. Tabla ANOVA para medidas repetidas.

Fuente de variación		Grados de libertad	Suma de cuadrados	Media cuadrática
Entre sujetos		n-1	$SC_{ENTRE}= k\sum_{i=1}^{n}\left(\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot}\right)^2$	$\frac{SC_{ENTRE}}{n-1}$
Intra sujetos	Observador	k-1	$SS_{OBS}= n\sum_{j=1}^{k}\left(\overline{X}_{\cdot j}-\overline{X}_{\cdot\cdot}\right)^2$	$\frac{SC_{OBS}}{k-1}$
	Residual	(n-1)(k-1)	$SS_{RES}= \sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij}-\overline{X}_{i\cdot}-\overline{X}_{\cdot j}+\overline{X}_{\cdot\cdot}\right)^2$	$\frac{SC_{RES}}{(n-1)(k-1)}$
Total		nk-1	$SC_{TOTAL}= \sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij}-\overline{X}_{\cdot\cdot}\right)^2$	$\frac{SC_{TOTAL}}{nk-1}$

n: número de sujetos.
k: número de observaciones por sujeto.

Tabla 3. Resultados de la medición de la presión arterial sistólica (TAS) en 30 pacientes, utilizando dos métodos diferentes.

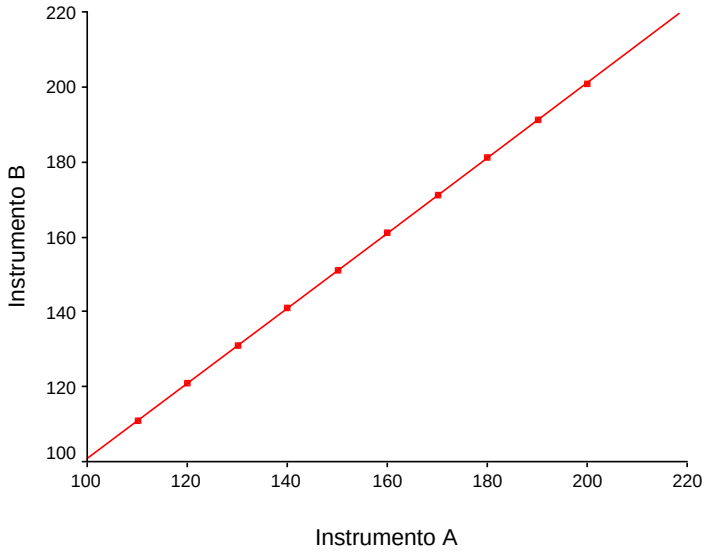
TAS Método A	TAS Método B	Diferencia
80	83	-3
85	83	2
90	94	-4
95	93	2
100	100	0
105	103	2
110	112	-2
115	114	1
120	121	-1
125	127	-2
110	111	-1
120	123	-3
130	128	2
140	148	-8
110	113	-3
130	132	-2
135	139	-4
140	144	-4
145	152	-7
150	157	-7
155	156	-1
160	171	-11
165	164	1
170	179	-9
175	181	-6
180	184	-4
185	190	-5
190	196	-6
195	203	-8
200	206	-6

Tabla 4. Tabla ANOVA para las mediciones de tensión arterial.				
Fuente de variación		**Grados de libertad**	**Suma de cuadrados**	**Media cuadrática**
Entre sujetos		29	73597,683	2537,851
Intra sujetos	Observador	1	156,817	156,817
	Residual	29	185,683	6,403
Total		59	73940,183	

Tabla 5. Valoración de la concordancia según los valores del Coeficiente de Correlación Intraclase (CCI).	
Valor del CCI	**Fuerza de la concordancia**
>0,90	Muy buena
0,71-0,90	Buena
0,51-0,70	Moderada
0,31-0,50	Mediocre
<0,30	Mala o nula

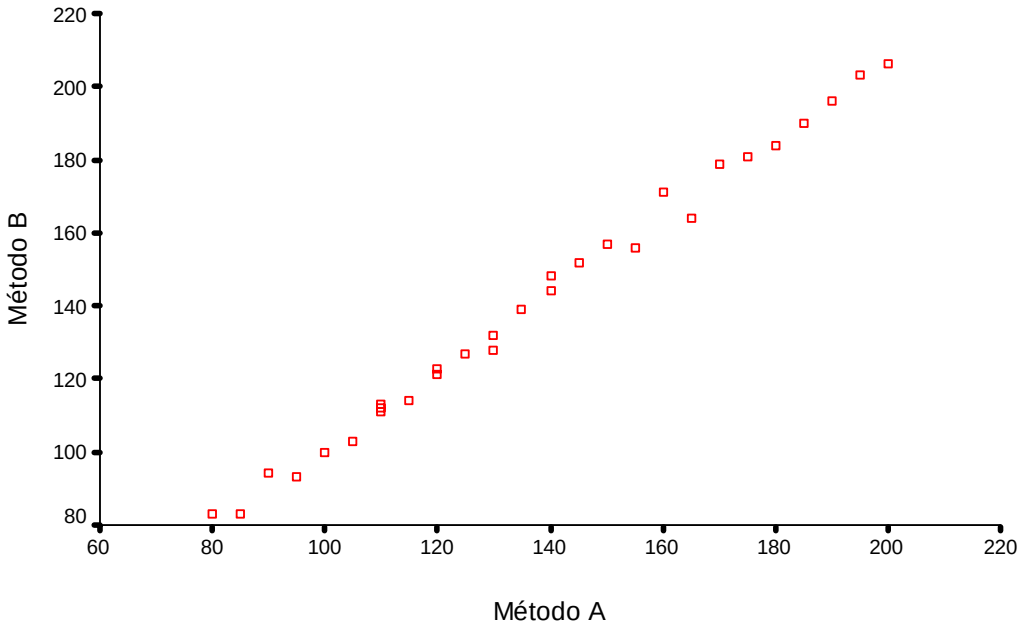
Tabla 6. Distribución de la diferencia absoluta entre las mediciones de tensión arterial en 30 pacientes.		
Diferencia absoluta	**Frecuencia**	**Porcentaje acumulado**
0	1	3,3%
1	5	20,0%
2	7	43,3%
3	3	53,3%
4	4	66,7%
5	1	70,0%
6	3	80,0%
7	2	86,7%
8	2	93,3%
9	1	96,7%
11	1	100,0%

Figura 1. Correlación entre los instrumentos A y B para la medición de la Tensión Arterial Sistólica.



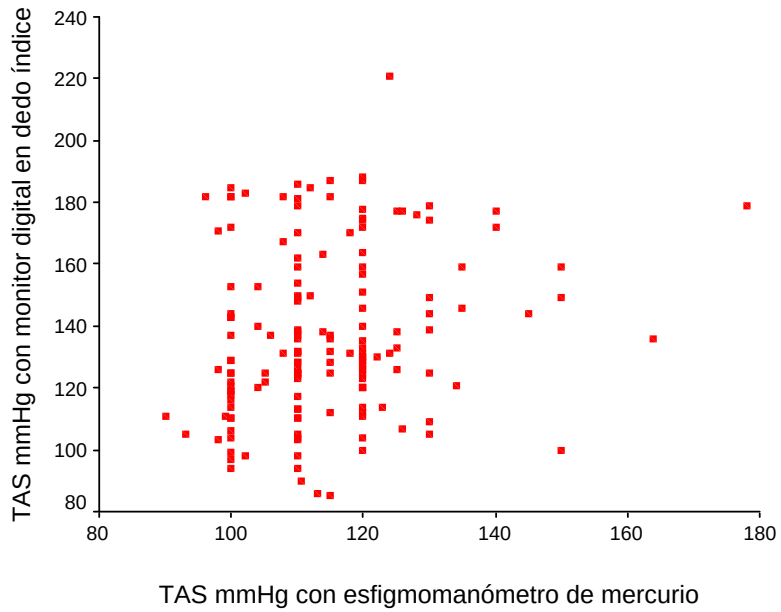
Coeficiente de correlación de Pearson r=1

Figura 2. Mediciones de tensión arterial sistólica en 30 pacientes según dos métodos de medición.



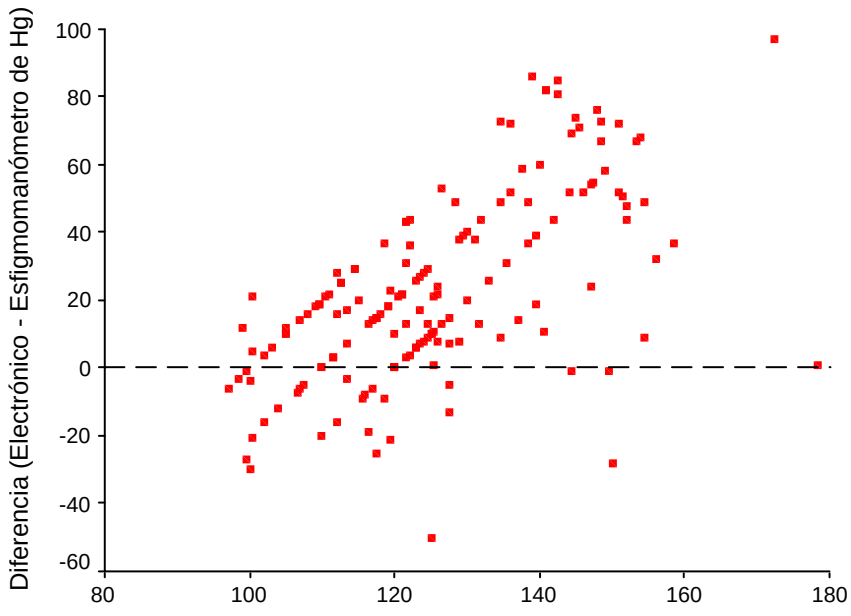
Coeficiente de correlación lineal r= 0.997

Figura 3. Correlación entre los valores de Tensión Arterial Sistólica medida con esfigmomanómetro de mercurio en brazo dominante y monitor digital en dedo índice.



Coeficiente de correlación de Pearson r=0,202 (p<0.05)

Figura 4. Diferencias en los valores de Tensión Arterial Sistólica medidos con esfigmomanómetro de mercurio en brazo dominante y monitor digital en dedo índice. Método de Bland y Altman.



Diferencia media=22,5 (D.T.=27,6); IC 95%=(18,1;26,8)

Figura 5. Histograma de las diferencias entre el monitor electrónico y el esfigmomanómetro de mercurio para la medición de Tensión Arterial Sistólica.

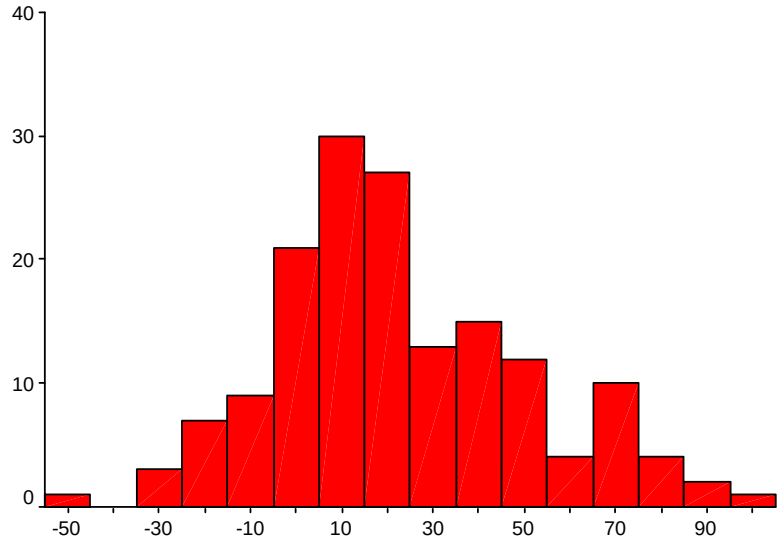


Figura 6. Diferencias en los valores de tensión arterial sistólica (TAS) según dos métodos de medida A y C en relación con su promedio.

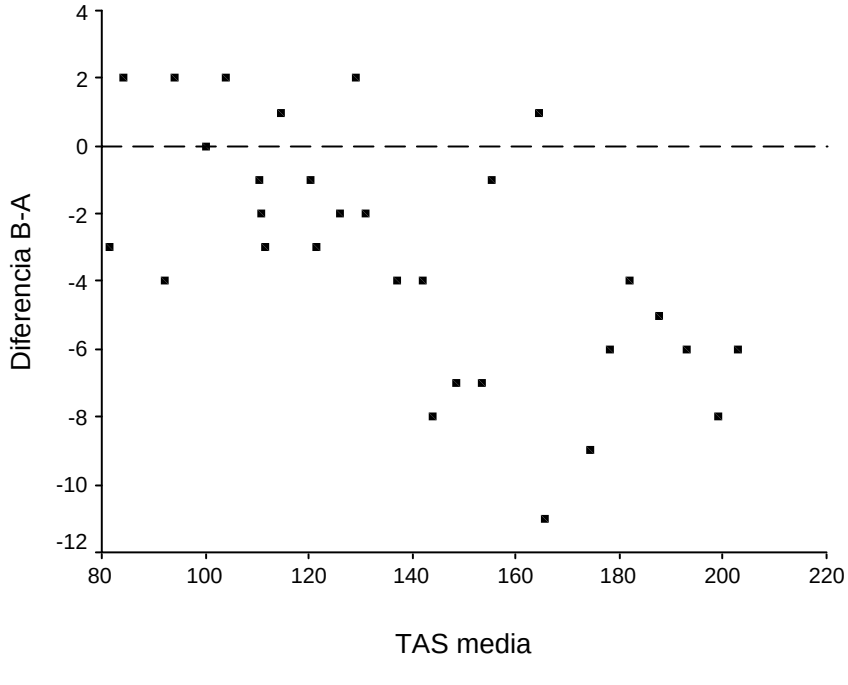


Figura 7. Porcentaje de discordancia entre dos métodos de medición de la tensión arterial sistólica (A y B)