Revista Mexicana de Cardiología

Volumen 26, Número 1
Enero - Marzo 2015
pp 53 - 55

Comunicado breve

Breve repaso de temas de estadística aplicada a la investigación médica

Cuauhtémoc Acoltzin Vidal*

BÚSQUEDA DE DIFERENCIAS ENTRE RESULTADOS DE INVESTIGACIÓN

Se ha comentado que si las variables dependiente e independiente son nominales, el análisis de los resultados permite identificar asociación entre ellas, mejor dicho entre los grupos estudiados que, estrictamente hablando, representan subconjuntos de un mismo universo.

Cuando la variable independiente –es decir aquella que se conoce y se puede manipular– es nominal (se le llama por su nombre, por ejemplo: intervención o control; medicamento o placebo; curación o muerte) y la variable dependiente –cuyo comportamiento es desconocido y casi siempre es el desenlace buscado– es numérica, el análisis estadístico va dirigido en busca de diferencias entre los resultados de los grupos que se comparan. Hay variedades para comparar una muestra con la población general, dos muestras entre sí, una sola muestra con dos resultados (se dice pareada) o más de dos muestras.

Resulta obvio que para indicar una prueba de diferencias hay que partir del tipo de variables según su nivel de medición: nominal y numérica.

El siguiente paso es observar la manera en que se distribuyen. Para esto se hace un histograma de cada grupo, es decir, se ordenan los datos de menor a mayor, acumulando los resultados coincidentes que se suman formando columnas alineadas de acuerdo con los intervalos de cada grupo de datos (pueden ser por cada dato o por clases diseñadas por el investigador, por ejemplo, en el caso de distribuir las edades es frecuente hacer grupos de cada cinco años. Aunque no hay una regla obligada, se sugiere calcularlos mediante raíz cuadrada del número de datos –o individuos participantes– ya que eso garantiza que todas las clases sean iguales.

Si el histograma adquiere forma acampanada y simétrica se dice que la distribución es normal –o paramétrica (recuérdese que se llaman parámetros a los resultados obtenidos en la población general)–, pero si es asimétrica se dirá que la distribución es libre –o no paramétrica (que puede ser sesgada hacia arriba o hacia abajo según se acerque o aleje del cero; rectangular o fractal).

En el caso de la distribución normal se usan pruebas llamadas paramétricas que están representadas por las de un grupo peculiar llamado «t» de Student y por el análisis de varianza (ANVA o ANOVA) que tienen requerimientos muy estrictos: como se parte de que las distribuciones son normales y, por lo tanto iguales, se exige que los datos sean numéricos, que el tamaño de las muestra sea igual y que las varianzas sean homogéneas.

Breves aclaraciones: teniendo los datos habrá que sumarlos y dividir el resultado entre el número de observaciones (casos, individuos o lo que fuere) y con eso se obtendrá la media (valor medio o media aritmética) que es medida de tendencia central y es la que se comparará. También se calculan medidas de dispersión que son la desviación estándar y su predecesora: la varianza. Ésta se calcula sumando las diferencias de cada dato con la media, elevadas al cuadrado; y luego la desviación estándar que es la raíz cuadrada de la varianza (por facilidad de manejo aritmético la diferencia de cada dato con la media se ha elevado al

* Médico Cirujano, Cardiólogo y Maestro en Ciencias Médicas. Universidad de Colima.

Rev Mex Cardiol 2015; 26 (1): 53-55

cuadrado, por lo tanto la raíz cuadrada evita este factor de error). Ambos datos se pueden calcular fácilmente con calculadora científica o desde el paquete Excel[§].

El análisis de las varianzas ya busca diferencias, aunque el investigador espera que sean iguales, es decir homogéneas. En primera instancia se divide la menor entre la mayor y si el resultados se acerca a la unidad se dirá que son homogéneas y se seguirá adelante con el análisis.

Las pruebas «t» consisten en dividir la diferencia de las medias de los grupos entre la dispersión de los datos (representada como error estándar de la media que resulta de la desviación estándar y la raíz cuadrada del número de datos, y tiene por objeto hacer inferencia a la población general).

En el caso de más de dos muestras, el ANOVA (cuyo cálculo requiere un procedimiento especial) puede ser suficiente pues si no hay diferencia entre varianzas el análisis podrá terminar; pero si la hay se hará lo que se conoce como análisis *post hoc* que consiste en hacer prueba «t» a cada par de grupos.

Si la distribución es libre se usan pruebas llamadas no paramétricas, sólo que éstas comparan las distribuciones y no los valores obtenidos en las mediciones. No tienen valor medio porque la curva no es normal. Se calculan entonces la mediana (M) y la moda (Mo). Para la primera se acomodan en línea todos los datos, de menor a mayor y se cuentan (sin apilarlos ni reunirlos en clases como se hizo para el histograma); el total de observaciones se divide entre dos y la cifra situada en ese orden (o sea en la mitad) será la mediana. La moda es aquélla que reúne más observaciones. Cada mitad de la fila se divide por en medio para tener cuatro cuartas partes llamadas cuartiles (identificadas por Q: Q1, Q2, Q3 y Q4), limitadas por el dato al que le toque el orden. Para analizar una muestra se usa la prueba de Kolmogórov-Smirnov (ante muestras pequeñas compara con la distribución de la población de la que pudiera ser sólo una fracción), para dos la de Mann-Whitney, para tres la de Kruskal-Wallis y para muestras pareadas la del signo o de Wilcoxon.

A partir de la desviación estándar –en las pruebas paramétricas– y del recorrido intercuartílico –en las no paramétricas– se calcula el intervalo de confianza de 95%, cuya importancia es que da idea al lector del margen de variación del resultado que podrá encontrar al aplicar el procedimiento probado en la población a su alcance en el mundo real.

En la investigación clínica es más frecuente obtener datos con distribución libre, lo que representa una dificultad para definir los límites de normalidad. La solución puede ser aumentar el tamaño de la muestra hasta incluir miles de observaciones, lo que no resulta fácil excepto para estudios multicéntricos. Más a la mano quedan recursos de análisis como la transformación *Z* y el score *Z*. La transformación *Z* consiste en modificar los datos de una distribución libre para acomodarlos artificialmente como curva normal: se calculan valor medio y desviación estándar; se hacen coincidir los resultados con los lugares que ocuparían si la curva fuera normal. Se sabe que a 1.96 desviaciones estándar de cada lado de la media se ubica el 95% de las observaciones; que a dos desviaciones estándar el 95.4%, a 2.576 el 99% y a tres desviaciones estándar el 99.7%; es decir: casi todas. Por lo tanto, aquellos valores que se sitúen más allá serán calificados como fuera de lo normal. A esta prueba también se le conoce como umbral de normalidad.

Si en vez de tener todos los datos de un estudio se cuenta sólo con uno y se quiere saber si está dentro de lo normal, se recurrirá al score *Z,* aplicando una fórmula en la cual se divide la diferencia entre el valor medio poblacional y el dato, entre la desviación estándar de la población. El resultado indicará el número de desviaciones estándar que está, el dato, separado de la media.

[§] Cálculos desde Excel: abra un libro (los archivos de Excel se llaman libros) nuevo. A partir de la celda 2B escriba una serie de datos numéricos en forma de columna. Deje el cursor en la celda que está debajo de la última cifra. Solicite función (*f*x): abrirá una ventana que dice "Insertar Función" para elegir (si no está disponible en el menú, se escribe en la ventana especial que hay para ello y se presiona ir). Se acepta. Abra otra ventana que dice "Argumentos de Función". Se eligen todos los datos arrastrando el cursor sobre ellos, usando para ello el *mouse*. Aceptar. En la celda preparada aparecerá el resultado. De la misma manera se pueden calcular: suma, promedio, var y desvest. Se puede hacer la prueba T si se tienen dos series de datos (colocados en columnas B y C con igual número de anotaciones). Se solicita prueba T. El paquete ofrece la ventana de "Argumentos de Función". Se selecciona la primera serie en la primera ventana y la segunda en la siguiente; en la tercera ventana se tiene que definir si se desea analizar la diferencia en los dos extremos de la curva o sólo en una. Por último se anota (con un código que la misma ventana ofrece al pie) si las varianzas son iguales. En ese momento aparece el valor «t» que deberá compararse con la tabla de P en el renglón correspondiente a los grados de libertad que se calculan restando uno al número de observaciones de los dos grupos sumados.

Rev Mex Cardiol 2015; 26 (1): 53-55

BIBLIOGRAFÍA COMPLEMENTARIA

1. Cañedo L. Investigación clínica. México, D.F., Nueva Editorial Interamericana; 1987.
2. Dawson-Saunders B, Trapp RG. Bioestadística médica. México, Editorial El Manual Moderno S.A. de C.V.; 1993. (capítulos 7 y 8).
3. Castilla Serna L, Cravioto J. Estadística simplificada para la investigación en ciencias de la salud. México, Trillas; 1991.
4. Liebovith LS et al. Nonlinear properties of cardiac rythm anormalities. Physical Review E. 1999; 59: 3312-3319.

Dirección para correspondencia:

Cuauhtémoc Acoltzin Vidal
E-mail: jose_rafael_c_acoltzin@yahoo.com