LA VARIACION Y SU SIGNIFICADO

JULIO CESAR ANGEL GUTIERREZ

RESUMEN

El artículo "La variación y su significado" es una recopilación de ideas sobre el concepto de la dispersión, que han surgido como consecuencia de mi experiencia docente en las áreas de Estadística y Control de Calidad.

Su contenido hace referencia al concepto de variación, no sólo desde el punto de vista estadístico, sino también desde los puntos de vista de la administración y la ingeniería.

Igualmente, se describen y explican las principales medidas de dispersión, con énfasis especial en la desviación estándar, la cual se constituye en el error más frecuente en cuanto a su cálculo e interpretación.

Todas las ideas expuestas en este artículo hacen referencia al mundo real, en donde prevalece más la heterogeneidad que la homogeneidad.

INTRODUCCION

Se encuentra la variación en todos los momentos de nuestra vida. Los gerentes y administradores de empresas viven enfrentados al reto de la heterogeneidad para la toma de decisiones objetivas en un mundo en donde prevalece más la variación que la uniformidad.

Un análisis exhaustivo de la variación se constituirá en una poderosa herramienta para extraer conclusiones razonables y tomar decisiones sólidas en todos los campos del mundo real.

En este artículo se presenta la variación mediante características que pueden expresarse cuantitativamente, con el fin de que ella pueda ser relacionada con el promedio y otros aspectos de las distribuciones.

1. EL MAL USO DE LA ESTADISTICA

Hay un viejo refrán que dice: "Las cifras no mienten, pero los mentirosos hacen las cifras".

El primer ministro británico Disraeli clasificó las falsedades en orden creciente de ignominia como "mentiras".

Como todo en la vida, muchas herramientas útiles y eficaces como la estadística pueden ser efectivamente mal usadas.

El mal uso de la estadística no siempre es el resultado de un inescrupuloso. En algunos casos son también algunos investigadores bien intencionados que no conocen los métodos estadísticos suficientemente pero que, sin embargo, los utilizan como un soporte a sus investigaciones.

JULIO CESAR ANGEL GUTIERREZ. Ingeniero Industrial U de A. Magister en Matemáticas Aplicadas EAFIT. Profesor del Departamento de Informática y Sistemas.

La importancia de la estadística no decrece porque ella sea mal utilizada a veces, ya sea por personas astutas o por investigadores poco competentes aunque bien intencionados.

2. LA DISPERSION Y LOS PROMEDIOS

La utilidad de un promedio no puede ser exagerada. A veces, carece totalmente de significado. Decir que el promedio de una característica o variable es igual a determinado número puede constituirse en un engaño para el interlocutor, si se tiene en cuenta que es necesario describir otras características de los datos para entender algo de su significado. La variación es una de esas características imprescindibles para interpretar un conjunto de datos, o una condición necesaria pero no suficiente para el buen manejo de la información.

Cuando se dice que el promedio del ingreso familiar en cierta ciudad es de $2.500.000.oo anualmente, muy poco se conoce acerca del comportamiento de los ingresos en esa ciudad, a menos que se especifique el tipo de promedio calculado, la variación de los datos con relación al promedio y la distribución de los mismos.

La importancia de la variación fue expresada vivamente por Darell Huff en su libro "How to lie with statistics":

> "Deposite poca fe en su promedio, en un gráfico o en una tendencia, cuando dichas cifras importantes brillan por su ausencia. De otro modo, se encuentra usted tan ciego como un hombre que escoge un lugar para acampar guiándose por un informe de la temperatura media solamente".

3. MODELOS DE VARIACION

Un estudio del número de accidentes de trabajo ocurridos en una empresa durante un año, por cada 1000 horas- hombre, puede llegar a concluír que la variación no está originada por cambios en las condiciones básicas que afectan el índice de accidentes. Este tipo de variación, que existe cuando los factores causales permanecen inalterados, recibe el nombre de variación inherente dentro de un sistema de causas constantes.

El modelo de variación para el número de accidentes por cada 1000 horas - hombre puede ser transformado si se modifican las causas que lo determinan. Una modificación de las causas puede lograrse implementando una campaña de seguridad en la planta, con la cual se podrá esperar que la mayor frecuencia o índice de accidentes disminuya considerablemente. En este caso se atenuarán las causas que generan los accidentes o se eliminarán algunas de ellas.

La campaña de seguridad no sólo cambia la dispersión, sino que también reduce el promedio.

La diferencia entre la variación inherente dentro de un sistema de causas constantes y los cambios en el modelo de variación que se presentan como resultado de los cambios en las condiciones causales, es de gran importancia en el análisis de datos.

4. LA VARIACION EN PROCESOS Y PRODUCTOS

Los diferentes procesos y productos realizados en una empresa pueden parecer iguales, pero en realidad no lo son. Las variaciones en las características de calidad pueden ser muy significativas. Variables como el tiempo, la resistencia, el peso, la longitud, etc., determinan procesos o productos aceptables cuando ellas están dentro de un intervalo de especificación, como por ejemplo $N \pm T$, pero los resultados de las medidas o inspecciones son diferentes en la práctica, si concebimos un alto nivel tecnológico en los equipos de medición.

Cuando se habla de ensambles o de productos formados por varios componentes, también se habla de partes intercambiables, las cuales deben ser producidas con cierta uniformidad, Esto significa que con relación a las características de calidad variables, ellas se identifican con un valor nominal, pero que realmente son admitidas con alguna variación o tolerancia. A medida que dicha tolerancia se hace menor, la calidad del producto o componente se hace mayor, determinando también una mayor uniformidad en el producto o proceso.

Es frecuente en la práctica del muestreo obtener distribuciones de datos con una variación significativamente mayor que la variación real de los mismos; esto ocurre muy a menudo cuando se toman las muestras en momentos en los cuales están variando las condiciones o causas que afectan las características de la variable en estudio. En estas situaciones es muy frecuente obtener distribuciones bimodales. Otras veces se toman muestras en poblaciones mal definidas o cuyo método de muestreo no es el adecuado, obteniendo una subvaloración o sobrevaloración en la variabilidad de los datos.

> Si la desviación de los datos con relación a la media es grande, entonces el valor del promedio no tiene mérito; en caso contrario el promedio es altamente significativo.

5. UTILIDAD DE LAS MEDIDAS DE DISPERSION

La variación entre los datos de un conjunto de valores numéricos es lo que se llama la dispersión, la cual puede ser medida de diferentes maneras. La más importante de ellas es la variación de los datos con relación a la media aritmética.

Cuando se mide la variación o dispersión de un conjunto de datos, generalmente se habla de la palabra desviación relacionada con un valor determinado, lo cual significa que estamos calculando la diferencia entre un dato cualquiera y en valor de referencia prefijado. Si X_1 representa un valor de una variable aleatoria y A es cualquier valor de referencia, entonces la desviación de la variable con relación al valor prefijado A es la diferencia $(X_1\text{-}A)$.

De igual manera, la desviación de un valor X_i de la variable aleatoria X, con relación a su media aritmética es $(X_1 - X)$.

Una medida de dispersión puede ser útil en dos sentidos:

- Para medir el grado de dispersión entre todos los valores de una serie de datos. Si la dispersión es muy pequeña, podemos hablar de homogeneidad de los datos; si es muy grande, podemos hablar de falta de control, en cuyo caso es necesario encontrar y estudiar sus causas.

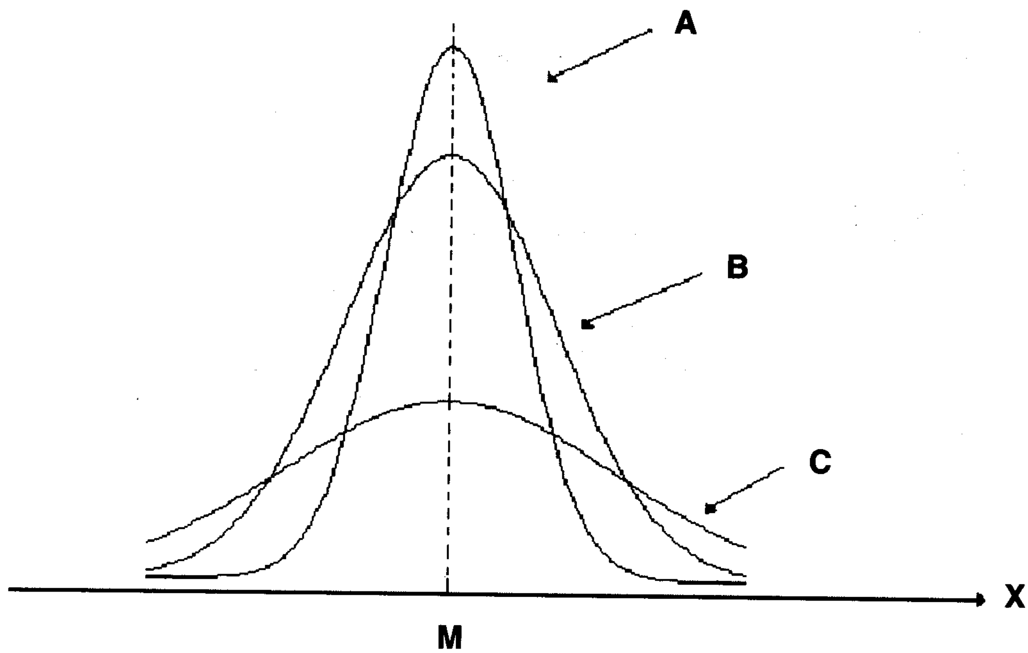
- Puede ser usada una medida de dispersión conjuntamente con un promedio para apreciar la homogeneidad de un conjunto de datos. Si la desviación de los datos con relación a la media es grande, entonces el valor del promedio no tiene mérito; en caso contrario el promedio es altamente significativo.

6. LA DISPERSION Y LA DISTRIBUCION DE DATOS

Dos conjuntos de datos pueden tener el mismo promedio y, no obstante, tener distribuciones muy diferentes. Esto indica que un promedio es sólo parte de la información que se necesita para conocer bien un conjunto de datos.

Se consideran tres conjuntos de datos A, B y C cuyas distribuciones se muestran en la siguiente figura:

FIGURA 1
Distribuciones con medias iguales y dispersiones diferentes



Como puede observarse, la media de los tres conjuntos de datos es la misma M; sin embargo, las dispersiones son bien diferentes. La curva A tiene menor dispersión que la B y ésta, a su vez, menor dispersión que la C.La dispersión de una distribución es una característica muy importante de entender y medir porque ella suministra información con relación a la confiabilidad del promedio, y porque se requiere la capacidad de reconocer que los datos están o no muy dispersos para poder abordar otros problemas.

A los analistas financieros les interesa la dispersión de las ganancias en su empresa. Cuando la dispersión de las ganancias es muy alta, se tiene un riesgo más alto para los accionistas y acreedores que cuando las ganancias permanecen relativamente estables. De manera análoga, los expertos en control de calidad analizan la dispersión para determinar los niveles de calidad de un producto.

Por esta razón, es por lo que se habla en calidad de los conceptos de exactitud y precisión.

7. EXACTITUD Y PRECISION DE LAS MEDIDAS

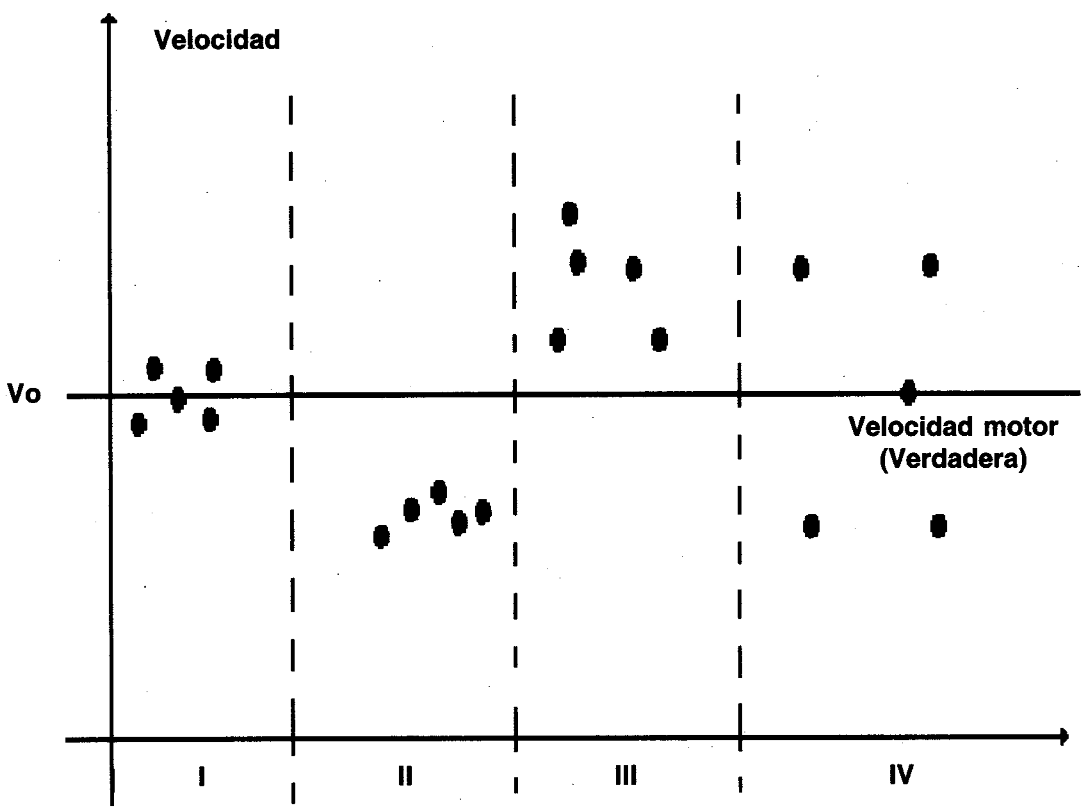
Cuando se realizan mediciones, se debe distinguir entre exactitud y precisión. La exactitud expresa cuán cerca están las medidas respecto del valor "verdadero" o "real" de la magnitud que se mide. El término "precisión" (o repetibilidad) se refiere al grado en el que las mediciones concuerdan entre sí.

El concepto de exactitud en control de calidad se evalúa por la diferencia entre la media de un proceso y el valor nominal de la especificación. En la práctica, es posible cambiar el centramiento del proceso mediante procedimientos de ajuste para lograr la exactitud deseada.

El mejoramiento de la precisión se logra con la aplicación de las mejores técnicas administrativas por parte de la dirección, tales como calidad total o reingeniería.

En la siguiente figura se muestra el resultado de 4 tacómetros que se utilizaron para medir la velocidad de un motor de velocidad constante, en donde cada instrumento realizó 5 mediciones:

FIGURA 2
Exactitud y precisión de las mediciones



> La exactitud expresa cuán cerca están las medidas respecto del valor "verdadero" o "real" de la magnitud que se mide. El término "precisión" se refiere al grado en el que las mediciones concuerdan entre sí.

Se observa que el tacómetro I tiene buena precisión porque los 5 valores están agrupados estrechamente y, buena exactitud porque los 5 valores están cerca del valor verdadero V_o.

El tacómetro II es preciso pero no muy exacto.

El tacómetro III no muestra exactitud ni precisión satisfactorias.

El tacómetro IV es exacto como el tacómetro I, pero es poco preciso.

En consecuencia, puede decirse que el tacómetro I tiene las dos características deseables.

Los errores de "precisión" a menudo reciben el nombre de errores aleatorios o accidentales, los cuales se pueden valorar estadísticamente. Los errores de exactitud se conocen como errores sistemáticos, y suelen reducirse por medio de la calibración de equipos. De esta manera se mejorará el funcionamiento de un instrumento de medición con exactitud deficiente, aunque posea buena precisión.

En casos en los cuales un error sistemático no haya sido corregido o eliminado completamente, éste se convertirá en parte del error aleatorio o de incertidumbre y, por lo tanto, así será considerado.

No se puede usar un rifle nuevo en prácticas de tiro que requieran precisión, si las pruebas señalan que es grande la desviación desde un blanco definido (precisión). Por otra parte, si la desviación es pequeña, pero la distancia promedio de los impactos desde el blanco (exactitud) es grande, es posible incorporar una corrección en la puntería al tratar de acercar los impactos al centro del blanco.

8. LA VARIACION Y LA CALIDAD

En los procesos de control se analizan en detalle las causas de variación de un proceso o producto. Se habla de causas de variación aleatorias, no asignables o casuales, como aquellas causas que generan una variación natural y que pueden considerarse como inherentes al proceso o producto.

Las causas de variación no aleatorias se llaman asignables y son las que generan efectos indeseables que son objeto de control. Tales causas deben ser detectadas y reducidas o eliminadas, si se quieren mejorar los niveles de calidad.

Las causas aleatorias de variación también pueden ser reducidas o eliminadas, pero para ello se requieren estrategias muy definidas como cambio de tecnología, motivación al personal, rediseño de procesos, cambio de materias primas, rediseño de productos, capacitación, etc.

9. LA VARIACION Y EL CONTROL DE PROCESOS

La variación es un fenómeno propio de todos los procesos productivos que se observa en el momento de comparar lo real con lo deseado o esperado de las características de calidad. La comparación nos lleva a concluir que una materia prima, proceso o producto es aceptado o rechazado, o que se debe tomar alguna acción correctiva en el tiempo.

La calidad de un producto terminado se mide por su variación con respecto a lo deseado y depende de la variación existente en las características de calidad de la materia prima, los materiales, los procesos y la mano de obra.

El concepto de variación nos lleva a pensar que no hay dos artículos iguales por más cuidado que se ponga en la obtención de sus características. En otras palabras, la característica de calidad que no sea variable, no es de interés en el control de los procesos.

Sólo las características variables son de interés, cuyo entorno y frecuencia deben ser analizados exhaustivamente.

10. EL MITO DE LA DESVIACION ESTANDAR

En todas las investigaciones en donde se involucra el manejo de datos cuantitativos, la media y la desviación estándar juegan un papel importante. Tanto la una como la otra son calculadas e interpretadas sin ningún tipo de cuestionamientos, porque se cree que con ellas se puede describir perfectamente una muestra o población.

La calidad de un producto se mide por su variación con respecto a lo deseado y depende de la variación existente en las características de la calidad de la materia prima, los materiales, los procesos y la mano de obra.

El caso de la desviación estándar es muy interesante. Todo el mundo la calcula e interpreta con una facilidad sorprendente, sin tener en cuenta que ella es sólo un ente matemático o estadístico que no permite explicarse en función de los datos.

Suponga que X es una variable aleatoria que representa la edad de los estudiantes de una universidad. Suponga además, que se toma una muestra al azar de 100 estudiantes y se obtiene que la media de la muestra es $\overline{X}$= 21 años y su desviación estándar es S = 2 años.

¿Qué significa en la práctica que la desviación estándar sea de 2 años?

Algunas interpretaciones comunes son las siguientes:

a. La mayoría de los estudiantes de la universidad tienen una edad entre 19 y 23 años.

b. La mayoría de los estudiantes de la muestra tienen una edad entre 19 y 23 años.

c. El promedio de alejamiento de las edades de los estudiantes de la universidad con relación a su media ($\overline{X}$ = 21 años) es de 2 años.

d. La diferencia promedio de las edades de los estudiantes de la muestra con relación a la media ($\overline{X}$ = 21 años) es de 2 años.

e. La raíz cuadrada de la suma de los cuadrados de las desviaciones de las edades de los estudiantes de la muestra, con relación a su media aritmética, dividido por sus grados de libertad es de 2 años.

¿Cuál de las interpretaciones anteriores es cierta?

En general, ninguna de ellas.

Las interpretaciones a) y b) son las que utilizan casi todas las personas que manejan datos. **Pero... ¿qué significa la mayoría?**

Si se considera la mayoría como algo más del 50% de los datos, equivaldría a decir que 51 estudiantes o más tienen una edad entre 19 y 23 años, lo cual es ambiguo y, en general, incorrecto.

Las interpretaciones c) y d) coinciden con un error frecuente de muchas personas. Interpretan la desviación estándar como si se tratase de la desviación media, $DM = \frac{1}{n}\sum|x_i - \overline{x}|$, tanto para el caso muestral como para el caso poblacional.

¿Se entiende en la práctica la interpretación dada en e)?

No. Sinembargo es una lectura descriptiva de la fórmula que define correctamente la desviación estándar muestral.

Los razonamientos anteriores nos permiten concluir que la desviación estándar es sólo un ente matemático y estadístico que sólo tiene sentido en situaciones muy particulares, y que fue definido de esta manera por tener propiedades muy deseables para los procesos de inferencia estadística. Para el caso de una distribución normal podríamos estimar que el 68.27% de los estudiantes de la universidad tienen una edad entre 19 y 23 años.

En términos prácticos la desviación estándar es una medida de dispersión comparativa, en el sentido de que una muestra o población con desviación estándar de 4 años es más dispersa o tiene mayor variación que otra muestra o población con desviación estándar de 3 años; sin embargo, los valores puntuales de la desviación estándar carecen de un significado práctico.

11. LA VARIABILIDAD Y LA ADMINISTRACION DE RIESGOS

De manera obvia, la variabilidad o dispersión se constituye en un riesgo en el sentido de que ella existe, pero en la mayoría de los casos es indeseable.

Cuando se identifica un valor nominal en una medición, como decir 1 kilo de arroz o una tubería de ½ pulgada, lo ideal sería que todos los paquetes pesara exactamente 1 kilo, y que todos los tubos tuvieran un diámetro de ½ pulgada; pero estos valores no son otra cosa que la identificación del producto, debido a que todo valor nominal tiene inherente una tolerancia que define la especificación del producto aceptable. Decir que se desea comprar 1 kilo de arroz puede ser equivalente a decir que se va a comprar una cantidad de arroz comprendida entre 980 y 1020 gramos.

De igual manera, decir que se va a comprar o fabricar una tubería de ½ pulgada, puede ser equivalente a decir que se va a comprar o producir una tubería con un diámetro entre 0.48 y 0.52 pulgadas.

En administración de riesgos se definen estos, para algunas aplicaciones, como unas condiciones en las cuales existe la posibilidad de desviaciones adversas con relación a unos resultados deseables o esperados. En este caso se habla de la dispersión o variabilidad como un factor negativo.

Si se tiene en cuenta que el riesgo está relacionado con una variación por fuera de las especificaciones, también se puede decir que la variabilidad dentro de ellas es permitida, y por lo tanto no constituye un riesgo.

En todos los casos es necesario tener en cuenta, no sólo la probabilidad de obtener un resultado que tiene una determinada desviación con relación a lo esperado, sino tambien la severidad o consecuencias económicas de cada resultado indeseable. Para tomar una decisión relacionada con la detección de un error o defecto, es necesario evaluar las consecuencias económicas de separarlo o dejarlo pasar.

12. PRINCIPALES MEDIDAS DE DISPERSION

Como se ha dicho, la dispersión es la cantidad de variación, desperdigamiento o diseminación de un conjunto de datos, con relación a un valor prefijado.

Las principales medidas de variabilidad o dispersión son las siguientes:

12.1 La amplitud o rango (R):

Es la medida de dispersión más sencilla, que se define como la diferencia entre los valores máximo y mínimo provenientes de una muestra o de una población.

Si X es una variable aleatoria bajo estudio y X1, X2, X3...Xn es una muestra, entonces la amplitud o el rango R está dado por R= Xmáx- Xmin.

El rango tiene sus ventajas y desventajas como medida de dispersión.

Entre las ventajas se citan las siguientes:

- Es fácil de calcular.
- Es fácil de comprender.
- Es útil para estimar la desviación estándar en los procesos de control estadístico.

Entre las desventajas, se citan las siguientes:

- Es muy sensible al tamaño de la muestra.
- Sólo incluye dos valores de la muestra o población.
- Es muy fácil obtener valores distorsionados, de tal manera que lo invaliden.
- No da información acerca del comportamiento de los datos entre el valor máximo y el valor mínimo.

En ocasiones se habla del recorrido, de rango o amplitud, pero calculando el valor absoluto de la diferencia.

Suponga que se toma una muestra de 100 estudiantes de una universidad de manera aleatoria. Después de consultar la edad de esos estudiantes se obtiene que: $\overline{X}$ = 21 años , R = 13 años.

¿Qué se puede decir acerca de la edad de estos 100 estudiantes con los estadísticos anteriormente calculados? Poco. Muy poco. No obstante, se puede concluir que los datos ocurren en un entorno con una amplitud de 13 años que contiene la media.

El rango ha sido utilizado con éxito en control de calidad para la estimación de la desviación estandár

12.2 La desviación media absoluta (DM):

La desviación media se define como el promedio de las distancias de los datos con relación a la media aritmética, esto es: $DM = \frac{1}{n}\sum|X_i - \overline{X}|$

Debe notarse que la desviación media sólo toma en cuenta la magnitud de las desviaciones con respecto a la media, y no su signo. Además, esta medida utiliza todos los datos, lo cual hace un poco más difícil su cálculo.

La ventaja básica de la desviación media es su fácil interpretación, aunque en la práctica sea confundida frecuentemente con la desviación estándar.

> La medición del riesgo, se hace en muchos casos, por la probabilidad de que el resultado se aleje una determinada distancia del valor esperado.

Estadísticamente, dispersión significa el grado de alejamiento de los valores muestrales o poblacionales con relación a un valor de referencia tal como el origen de los datos, la media aritmética, la mediana, etc.

Si en el ejemplo de los 100 estudiantes se sabe que $\overline{X}$ = 21años y DM = 2 años, se puede concluir que las edades de los 100 estudiantes se alejan de 21 años en un valor promedio de 2 años.

12.3 La desviación estándar (S ó σ)

La desviación estándar o típica se define como la raíz cuadrada de la suma de los cuadrados de las desviaciones de los datos con relación a la media aritmética,dividida por sus grados de libertad, esto es: $S = \sqrt{\Sigma (X_i - \overline{X})_2 / (n-1)}$ para el caso de una muestra. Si se trata de la desviación estándar poblacional (parámetro) entonces la media aritmética X se cambia por la media poblacional μ y el número de grados de libertad (n - 1) se cambia por el número N de elementos de la población.

Bien se sabe que la desviación estándar no tiene una interpretación práctica de carácter puntual para sus valores y que sólo permite una interpretación comparativa. Sin embargo, tiene algunas propiedades deseables como indicador de la variabilidad en un conjunto de datos. Entre ellas:

- Tiene en cuenta todos los valores muestrales o poblacionales.
- Se puede calcular por métodos abreviados.
- Su valor es mínimo, en el sentido de que al cambiar la media aritmética de su fórmula por cualquier número real, su resultado será siempre mayor.

$$S = \sqrt{\Sigma\left(X_i - \overline{X}\right)_2} / (n-1) \leq \sqrt{\Sigma\left(X_i - X_k\right)^2 / (n-1)} \text{ con } X_k \in \Re$$

- El valor esperado o esperanza matemática de la desviación estándar al cuadrado coincide con el cuadrado de la desviación estándar poblacional.

  Esta propiedad es de gran utilidad en la inferencia estadística.

12.4 La varianza (S^2 ó σ^2):

Es una medida de dispersión equivalente a la desviación estándar elevada al cuadrado. Para efectos de interpretación tiene más dificultades que la desviación estándar, debido a que sus unidades de medida son diferentes a las de los datos por estar elevadas al cuadrado y, como ella, sólo puede entenderse desde un punto de vista comparativo. Sus ventajas y desventajas están directamente relacionadas con las de la desviación estándar.

12.5 El rango intercuartil (RI):

El rango intercuartil está definido como el rango de los datos resultantes después de eliminar el 25% de datos superiores y el 25% de datos inferiores. En otros términos se han eliminado los datos que están por debajo del primer cuartil y por encima del tercer cuartil, obteniendo la siguiente expresión: RI = Q3 - Q1

Con esta medida se obtiene un indicador de dispersión para los datos más representativos de la distribución y que no incluyen los datos extremos.

Así como el rango intercuartil elimina el 50% de los datos, también podría ser útil en algunas aplicaciones el hecho de eliminar otro porcentaje diferente de datos en cada uno de los extremos. Los estadísticos así obtenidos reciben el nombre correspondiente con el apelativo de recortado. Se habla así de la media recortada, la desviación estándar recortada, etc.

Su utilización es muy frecuente en estudios sociales y biológicos.

12.6 El coeficiente de variación (CV)

El coeficiente de variación está definido como la relación entre la desviación estándar y la media aritmética de los datos. Al igual que la desviación estándar, tiene dificultades para su interpretación puntual. Si en un conjunto de datos (muestra) obtenemos que CV = S / $\overline{X}$ = 0.65, sólo podemos concluir que la desviación estándar es el 65% del valor de la media aritmética, pero no podemos decir algo acerca de que la muestra tenga una alta, mediana o baja variabilidad.

No obstante, se puede calcular el coeficiente de variación para dos muestras y concluir que el coeficiente de variación mayor coincide con la muestra de mayor variabilidad relativa.

Por ejemplo, si se desea comparar la variabilidad entre la población de los pesos de todos los productos fabricados por una empresa y la población de los volúmenes envasados por una máquina envasadora en una fábrica de gaseosas, entonces el coeficiente de variación será una medida comparativa muy importante para la dispersión entre las dos poblaciones.

13. LA VARIACION Y LOS GRADOS DE LIBERTAD

Al calcular la variación de un conjunto de datos, generalmente se incluye el concepto de grados de libertad, como ocurre en los casos de la varianza y la desviación estándar, los cuales son de gran utilidad en la estadística inferencial.

El número de grados de libertad puede considerarse como el número de variables que pueden variar libremente en un conjunto de ellas y bajo ciertas condiciones. Si se consideran 3 variables X, Y, Z con la condición de que la suma de ellas es 15, entonces el número de grados de libertad es igual a 2, porque si se conoce, por ejemplo, que X = 7, Y = 3 (2 grados de libertad) entonces se tendrán perfectamente determinado el conjunto de variables; porque si X + Y + Z = 15 (una restricción), con X = 7 y Y = 3, entonces Z = 5.

> La desviación media es una medida de dispersión que muchas personas interpretan erróneamente como si se tratase de la desviación estándar.

De otra manera, se podría decir que el número de grados de libertad es igual al número de variables menos el número de restricciones de las variables.

Si se toma una muestra X1, X2, X3...Xn de la cual se sabe que X = 10, basta con conocer el valor de (n - 1) variables, para que el valor de la varianza sea automáticamente determinado, puesto que el enésimo punto es fijo. En otros términos, se puede decir que son n variables con una restricción, $\sum\left(X_i - \overline{X}\right) = 0$, por lo tanto se tendrán (n - 1) grados de libertad para el cálculo de la varianza, esto es:

$$S^2 = \sum\left(X_i - \overline{X}\right)^2 / (n-1)$$

Una segunda interpretación de los grados de libertad consiste en decir que equivale al número de comparaciones que pueden efectuarse en una muestra de tamaño n. Cada una de las n observaciones de una muestra pueden compararse con otras (n - 1) observaciones.

Una tercera interpretación de los grados de libertad puede ser visualizada considerando un punto que puede moverse libremente en un espacio tridimensional. Dicho punto podría localizarse en las coordenadas (x, y, z) del espacio. Si se limita su movimiento a un un solo plano, entonces sólo se tendrían 2 grados de libertad.

En algunas aplicaciones de ingeniería se incluyen, no solo los movimientos de un punto en la dirección de los ejes, sino también los movimientos de rotación alrededor de cada uno de ellos. Esto daría origen a seis grados de libertad.

Como puede concluirse, el número de grados de libertad de un conjunto de variables está directamente relacionado con su dispersión.

En el caso particular de la varianza se dice que ésta aumenta a medida que el número de restricciones sea mayor, lo cual implica que existe una disminución en el número de grados de libertad.

Un grado de libertad también se define como una comparación entre los datos, independientemente de otras que se realicen en el análisis. Cada una de las observaciones de una muestra al azar de tamaño n, se puede comparar con otras (n-1) observaciones; de esto se puede concluir que haya (n-1) grados de libertad.

En caso de que se conozca la media real μ, y se desee compararla con las n observaciones de una muestra, cada una de ellas podrá ser comparada con μ, de manera independiente; de ahí, que cuando se calcule la varianza muestral con μ conocido, se divida el término $\sum (x_i - \mu)^2$ entre n.

13.1 Descomposición de los grados de libertad

Una práctica usual en estadística consiste en tomar K muestras independientes de tamaño nj en igual número de poblaciones, con el objetivo de analizar los datos y tomar decisiones.

Para la solución de estos problemas, es necesario calcular las diferentes varianzas o medias cuadráticas asociadas con los datos. Ellas son:

La varianza total, que se calcula con base en las desviaciones de todos los datos de las K muestras con relación a las media aritmética de todos ellos. Para este caso, se tiene que el número de grados de libertad es N-1, correspondientes á N variables (N= n1+ n2 +n3 +...+ nk) y una restricción.

La varianza entre muestras, que se calcula con base en los desviaciones de las medias aritméticas de cada una de las K muestras con relación a la media aritmética de todos los datos.

El número de grados de libertad para esta varianza es K-1, correspondientes a las K medias aritméticas, menos la restricción correspondiente a la media aritmética de todos los datos.

Varianza dentro de las muestras, que se calcula con base en las desviaciones de los datos de todas las muestras con relación a las medias aritméticas de cada una de las poblaciones. El número de grados de libertad es N-K, correspondientes N variables y K restricciones para cada una de las medias de las muestras.

Con los razonamientos anteriores se obtiene que el número de grados de libertad de la varianza total se descompone en dos partes; una para la varianza entre muestras, y otra para la varianza dentro de las muestras, esto es: $N - 1 = (K - 1) + (N - K)$

De esta manera puede observarse que los componentes de la variabilidad total no tienen el carácter de promedios, debido a que sus denominadores no toman valores que coincidan con el número de datos. Esto corrobora aún más, el hecho de que las varianzas sólo constituyen un ente matemático de difícil interpretación, pero con propiedades deseables para la inferencia estadística.

BIBLIOGRAFIA

Downie, N.M. y R. W. Heath. "Métodos estadísticos aplicados". 5ª ed. México: Harla, 1986.

García, P. Fernando y Fernando Garzo Pérez. "Estadística". Madrid: McGraw Hill, 1988.

Kelejian, Harry H. Y Wallace E. Oates. "Introducción a la econometría". 3ª ed. México, 1994.

Mesa, E. Antonio., Julio César Angel C. Y José Antonio Riascos. "Control de calidad". Medellín: ed. ACCC, 1981.

Naiman Arnold, Rosenfeld Robert y Gene Zirkel. "Introducción a la estadística". 3ª ed. México: Mc Graw Hill, 1987.

Neter, John y William Wasserman. "Fundamentos de estadística". 3ª ed. México: CECSA, 1975.

Sánchez A., Javier I. "Manual práctico de estadística". 4ª ed. Medellín: Universidad Nacional, 1990.

Schefler, William C. "Bioestadística". 2ª ed. México: Fondo Educativo Interamericano, 1981.

Shao, Stephen P. "Estadística para economistas y administradores de empresas". 8ª ed. México: Herrero Hermanos, Sucs. S.A., 1973.