Tema 61

Desigualdad de Tchebyshev. Coeficiente de variación. Variable normalizada. Aplicación al análisis, interpretación y comparación de datos estadísticos

61.1 Desigualdad de Markov

Teorema 1 ***(Desigualdad de Markov)****: Si $g(X)$ es una función medible de una variable aleatoria X tal que $g(X) \geq 0$, entonces:*

$$\forall k > 0 \qquad P\{g(X) > k\} \leq \frac{Eg(X)}{k}$$

siempre que $\exists Eg(X)$.

Demostración:

Sea $S : \{x \in \mathbb{R} : g(x) > k\}$. Entonces:

$$Eg(X) = \int_{\mathbb{R}} g(x)\, dF_X(x) = \int_S g(x)\, dF_X(x) + \int_{S^c} g(x)\, dF_X(x) \geq$$

$$\geq \int_S k dF_X(x) = kP\{g(X) > k\}$$

y por tanto:

$$\forall k > 0 \qquad P\{g(X) > k\} \leq \frac{Eg(X)}{k}$$

como queríamos demostrar. $\square$

61.2 Desigualdad de Tchebyshev

61.2.1 Acotación de Tchebyshev

En las misma hipótesis del teorema de Markov, si tomamos $g(X) = (X - \alpha_1)^2$ y $k = r^2\sigma^2$, siendo α_1 y σ^2 respectivamente la media y la varianza de la variable aleatoria X, tendremos:

$$P\left\{(X - \alpha_1)^2 > r^2\sigma^2\right\} \leq \frac{E\left[(X - \alpha_1)^2\right]}{r^2\sigma^2} = \frac{\sigma^2}{r^2\sigma^2} = \frac{1}{r^2}$$

es decir,

$$P\{|X - \alpha_1| > r\sigma\} \leq \frac{1}{r^2}$$

Así:

$$P\{|X - \alpha_1| > r\sigma\} \leq \frac{1}{r^2}$$

61.2.2 Desigualdad de Tchebyshev generalizada

Teorema 2 ***(Desigualdad de Tchebyshev)****: Si g es una función medible no negativa de la variable aleatoria X y $\exists E\left[g(X)^k\right]$, entonces:*

$$P\{g(X) > t\} \leq \frac{E\left[g(X)^k\right]}{t^k}$$

Demostración:

Sea $S = \{x \in \mathbb{R} : g(x) \leq t\}$, que es un conjunto medible. Entonces:

$$Eg(X)^k = \int_{\mathbb{R}} g(x)^k\, dF_X(x) = \int_S g(x)^k\, dF_X(x) + \int_{S^c} g(x)^k\, dF_X(x) \geq$$

$$\geq \int_{S^c} g(x)^k\, dF_X(x) \geq \int_{S^c} t^k dF_X(x) = t^k P\{X \in S^c\}$$

siendo esta desigualdad estricta si $P\{X \in S^c\} \neq 0$. Es decir,

$$Eg(X)^k \geq t^k P\{g(X) > t\} \Rightarrow P\{g(X) > t\} \leq \frac{E\left[g(X)^k\right]}{t^k}$$

como queríamos demostrar. $\square$

61.2.3 Casos particulares

Si hacemos $g(X) = |X - EX|$ y $k = 4$, entonces:

$$P\{|X - EX| > t\} \leq \frac{\mu_4}{t^4}$$

Así, dada una variable aleatoria X con desviación típica finita σ, la probabilidad de que esté situada la masa en un intervalo de longitud $2t\sigma$ centrado en su media, es mayor o igual que $1 - \frac{1}{t^2}$:

$$P\{EX - 2t\sigma \leq X \leq EX + 2t\sigma\} \geq 1 - \frac{1}{t^2}$$

En particular:

$$P\{EX - 2\sigma \leq X \leq EX + 2\sigma\} \geq \tfrac{3}{4}$$

$$P\{EX - 3\sigma \leq X \leq EX + 3\sigma\} \geq \tfrac{8}{9}$$

61.3 Variable aleatoria normalizada

Definición 3 *Una variable aleatoria se dice tipificada, normalizada o estandarizada sii tiene esperanza matemática cero y desviación típica uno.*

Proposición 4 *Sea X una variable aleatoria con*

$$EX = \mu \quad y \quad Var(X) = \sigma^2$$

Definimos la variable aleatoria $Y = \frac{X-\mu}{\sigma}$. Entonces, la variable Y tiene esperanza matemática cero y desviación típica uno.

Demostración:

Se tiene que

$$EY = E\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}E[X-\mu] = \frac{1}{\sigma}(EY - \mu) = \frac{1}{\sigma}(\mu - \mu) = 0$$

y por otra parte

$$Var(Y) = E[(Y - EY)]^2 = E\left[Y^2\right] \qquad (1)$$

ya que $EY = 0$ según se demostró anteriormente.

De la igualdad (1) resulta:

$$Var(Y) = E\left[\left(\frac{Y-\mu}{\sigma}\right)^2\right] = \frac{1}{\sigma^2}E\left[(Y-\mu)^2\right] = \frac{1}{\sigma^2}\sigma^2 = 1$$

como queríamos demostrar. □

61.4 Coeficiente de variación de Pearson

Sea X una variable aleatoria tal que

$$EX = \mu$$
$$Var\,(X) = \sigma^2$$

donde la desviación típica σ indica la dispersión de los valores de X con relación a la media μ.

Un problema que se plantea al intentar comparar las distribuciones de dos variables aleatorias X e Y es que tanto la media como la desviación típica de estas distribuciones dependen de las unidades de medida de las variables X e Y. Para soslayar esta dificultad se introduce una medida denominada coeficiente de variación de Pearson que no depende de las unidades de medida usadas, de tal manera que entre dos distribuciones diremos que posee menos dipersión aquella cuyo coeficiente de Pearson sea menor.

Definición 5 *Dada una variable aleatoria X con $EX = \mu$ y desviación típica σ, se define el coeficiente de variación de Pearson por:*

$$CV = \frac{\sigma}{\mu}$$

Este coeficiente se usa frecuentemente para medir la dispersión en las distribuciones de frecuencias ya que es muy usual que cuando se tienen dos o más distribuciones de frecuencias éstas no vengan dadas en las mismas unidades.

61.5 Aplicaciones al análisis estadístico de datos

61.5.1 De la desigualdad de Tchebyshev

Al estudiar la distribución de probabilidad binomial en el tema 65 se ha comprobado que la variable aleatoria

$$X = \frac{1}{n} Y_n$$

que indica la frecuencia relativa del suceso aleatorio A, de probabilidad p en n pruebas independientes puede tomar los valores:

$$\frac{r}{n} : r = 0, 1, ...n$$

y sus probabilidades vienen dadas por

$$P\left[X = \frac{r}{n}\right] = \binom{n}{r} p^r q^{n-r}$$

donde $q = 1 - p$.

Se compruba sin dificultad que:

$$\mu = EX = E\left[\tfrac{1}{n}Y_n\right] = \tfrac{EY_n}{n} = \tfrac{np}{n} = p$$

$$\sigma^2 = Var\left(X\right) = Var\left(\tfrac{1}{n}Y_n\right) = \tfrac{Var(Y_n)}{n^2} = \tfrac{npq}{n^2} = \tfrac{pq}{n}$$

Aplicando la desigualdad de Tchebyshev a la variable aleatoria $Y = \frac{1}{n}Y_n$ resulta:

$$P\left[\left|\frac{1}{n}Y_n - p\right| < k\sqrt{\frac{pq}{n}}\right] \geq 1 - \frac{1}{k^2} \qquad (1)$$

Si tomamos $\frac{1}{k^2} = \alpha$, se puede tomar n suficientemente grande para que dado $\varepsilon > 0$ se cumpla que

$$k\sqrt{\frac{pq}{n}} = \frac{1}{\sqrt{\alpha}}\sqrt{\frac{pq}{n}} < \varepsilon \qquad (2)$$

para lo cual ha de suceder que

$$n > \frac{pq}{\alpha\varepsilon^2}$$

y al sustituir (2) en (1) queda

$$P\left[\left|\frac{Y_n}{n} - p\right| < \varepsilon\right] \geq 1 - \frac{pq}{n\varepsilon^2} \qquad (3)$$

La desigualdad (3) se conoce con el nombre de Teorema de BERNOUILLI y nos indica que hay una probabilidad tan próxima a 1 como queramos de que, tomando n suficientemente grande, pero fijo, se tenga

$$\left|\frac{Y_n}{n} - p\right| < \varepsilon$$

Cuando la probabilidad p es desconocida, el número $T = pq = p\left(1-p\right) = p - p^2$ se puede acotar por $\frac{1}{4}$ que es el mayor valor que puede tomar la expresión $T = pq$.

61.5.2 De la variable aleatoria normalizada

La distribución de probabilidad normal es de uso muy frecuente y su función de densidad de probabilidad es

$$f\left(x\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad \forall x \in \mathbb{R}$$

donde μ y σ representan la media y la desviación típica respectivamente.

En el caso de que $\mu = 0$ y $\sigma = 1$, la función de densidad es

$$f\left(x\right) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

llamada distribución normal tipificada.

Para el cálculo de probabilidades correspondientes a fenómenos que siguen la ley de distribución normal es preciso tipificar las variables aleatorias, ya que todas las tablas proporcionan probabilidades de la distribución $\mathcal{N}(0,1)$.

61.5.3 Del coeficiente de variación

En el apartado 3 del presente tema se ha explicado la utilidad del coeficiente de variación de Pearson para comparar dos distribuciones analizando la dispersión de cada una de ellas.