2. REPASO DE ESTADÍSTICA

ESTADÍSTICA APLICADA

Necesidad de la Estadística.

Necesidad de **razonamientos inductivos** a partir de datos: Se hacen afirmaciones acerca de un colectivo de individuos u objetos, habiendo observado en realidad sólo una parte de ellos.

Definición de Estadística:

> *Conjunto de métodos para recoger, clasificar, representar y resumir datos, así como para hacer inferencias científicas a partir de ellos.*

Ejemplo 1.0: SNED Sistema Nacional de Evaluación del Desempeño Docente

Estadística Descriptiva. Con el estudio de ciertos estadísticos se conocen magnitudes que representan a la globalidad de los datos disponibles de forma resumida.

Inferencia Estadística. La segunda fase es la formulación y confirmación de hipótesis, Se cuantifica el grado de certidumbre con el que se pueden establecer afirmaciones sobre los datos: Se obtienen conclusiones a partir de una información incompleta

Población: conjunto de los objetos (individuos, observaciones, etc.) que se desea observar. Puede ser finito o infinito.

Notación: Ω

Ejemplo 1.1: Los habitantes del Gran Santiago, de 18 años o más.

Muestra: parte de la población Ω seleccionada para un experimento

El análisis de la información será de gran ayuda para la **toma de decisiones** y la realización de investigaciones

No hay que olvidar que los datos disponibles suministrarán una información parcial del proceso en estudio y aunque la estadística valide unas hipótesis, el investigador deberá dar un **significado real** a las conclusiones en el contexto correspondiente.

Algunos conceptos básicos:

Observación: una observación es un objeto individual que nos sirve como fuente de datos para la realización de nuestra investigación. Reciben diferentes denominaciones: Unidades muestrales, Individuos, Observaciones, Casos, Objetos, Unidades experimentales:

$$\omega \in \Omega$$

Variable: Es una característica del individuo que puede tomar distintos valores. Cuando medimos algo representamos por un modelo numérico aquello que medimos:

$$X : \Omega \to Q$$

(Q es el conjunto de todos los valores posibles que puede tomar X sobre los elementos de Ω)

Ejemplo 1.2: La altura de una persona: asignamos un número a cada persona. Las medidas físicas, como altura y peso, se miden con un instrumento físico. Otras propiedades abstractas tales como razonamiento, inteligencia se miden indirectamente.

Valor: son los distintos estados en los que se puede encontrar una característica de un individuo. Estos pueden ser cualitativos (masculino, femenino) o cuantitativos (163 cm):

$$X(\omega) \in Q$$

Observación: Si $\Omega \subset IN$ (población finita o numerable) se anota $X_i \in Q$.

PROBLEMAS TÍPICOS

1. Determinar las Unidades Experimentales o Unidades Muestrales
2. Homogeneidad respecto a otras características que puedan influir
3. Obtención de las medidas: Grandes errores

ORGANIZACIÓN MATRICIAL

En general, los datos a analizar consistirán de un conjunto de *p* variables medidas en *n* unidades muestrales.

Grabados en Hojas de cálculo (EXCEL, LOTUS), Bases de Datos (DBASE, ACCESS) o Programas Estadísticos (STATGRAPHICS, SPSS)

NUMERO	SEXO	EDAD	EJERCICIO	ALCOHOL	TABACO	…
001	H 67		3	35	0	
003	M 76		1	56	10	
004	H 56		2	112	15	
005	M 63		4	67	25	
…						

CLASIFICACIÓN DE VARIABLES

Se puede considerar tres clasificaciones de variables (dependiendo del conjunto Q):

1. Según la escala: Nominal, Ordinal, de Intervalo y de Razón.
2. Cualitativas y Cuantitativas.
3. Discretas y Continuas

Según la Escala:

Una clasificación comúnmente aceptada especifica cuatro tipos de variables: nominal, ordinal, intervalar, de razón.

Variables nominales

Una escala nominal es un sistema de clasificación que sitúa a personas, objetos u otras entidades dentro de categorías mutuamente excluyentes

Podemos usar símbolos (H/M, SI/NO) para representar las dos categorías.

Algunos programas de análisis de datos tratan sólo símbolos numéricos, por lo que es preferible esta representación. Puesto que las categorías pueden considerarse en cualquier orden cualquier conjunto de números será válido para su representación: 0/1, 1/2 (para no confundir ceros con blancos), 1/6 (para evitar errores de grabación).

Variables Ordinales

En este caso se usan categorías, pero existe un orden conocido entre ellas. Por ejemplo una escala de niveles de dureza de minerales, un nivel socioeconómico, etc. Puede usarse cualquier secuencia de números crecientes para su representación. Para definir una variable ordinal la operación básica es determinar si una observación es mayor que otra.

Variables de intervalo

Una variable intervalo es una variable ordinal especial, en la que las diferencias entre dos valores sucesivos es siempre la misma. Por ejemplo, la variable temperatura en grados Fahrenheit.

Variables de razón

Son variables de intervalo en las que además hay un punto natural representando el origen: punto cero. Por ejemplo, la altura.

Cualitativas y Cuantitativas:

Las variables ***cuantitativas*** son aquellas en la que los valores son números. Cuantifican características que unos poseen en mayor **cantidad** que otros:

$$Q \subset IR$$

En las ***cualitativas***, también llamadas categóricas o de clasificación, los diferentes valores representan **grupos** distintos a los que el sujeto puede pertenecer.

Ejemplo 1.3: $Q = \{verde, azul, negro, ...\}$ si X: Color de ojos.

Continuas y Discretas:

Una variable se dice ***continua*** si puede tomar cualquier valor en un rango específico. Por ejemplo, altura, peso, densidad, tiempo, resistencia.

Una variable que no es continua es ***discreta***. Puede tomar sólo ciertos valores específicos. Por ejemplo: número de hijos, sexo, identificación con partido político. A veces a las variables de este tipo se les denomina también *atributos*.

Esta última clasificación lleva, posteriormente, a considerar las posibles **distribuciones de las variables** que se suponen en los análisis. De esta forma una variable discreta puede seguir una distribución Binomial, de Poisson, etc., mientras que la distribución Normal se usa para describir la distribución de las variables continuas.

ENTRADA DE DATOS

NUMERO	SEXO	EDAD	EJERCICIO	ALCOHOL	TABACO	ALIM_GRA	COLEST	ANT_FAM	PROB COR
001	H	67	3	35	0	600	185	1	0
003	M	76	1	56	10	690	210	2	1
004	H	56	2	112	15	-1	195	1	1
005	M	63	4	67	25	650	200	2	0
006	H	55	1	-1	0	750	230	9	-1
Identif	**CUALI**	**CUANT**	**ORDINAL**					**CUALIT**	**CUALIT**

1. Nombres de las variables, nombres de códigos para descodificación

2. Codificar preferiblemente en números

3. Codificación detallada. Ejemplo: Edad, Tabaco con valor exacto.
   No intervalos que se pueden generar posteriormente.

4. Chequeo de rangos, máximos, mínimos.

5. Valores Missing: definición, codificación.
- Los análisis multivariantes requieren casos completos
- Valorar la supresión de un caso o una variable con alta proporción de valores missing

6. Copias de seguridad

7. Chequeo inicial: frecuencias, máximos y mínimos, gráficas, detectar valores no admisibles, inconsistentes, errores, etc.

ESTADÍSTICA DESCRIPTIVA

Sea Ω una población finita (o subconjunto finito de una población mayor).

Considere $X : \Omega \to Q$

Como $Card\,\Omega = |\Omega| = n$, podemos asociar Ω con $\{1,2,3,\ldots n\}$.

Así, $\{x_1, x_2, \ldots x_n\}$ son los valores de Q.

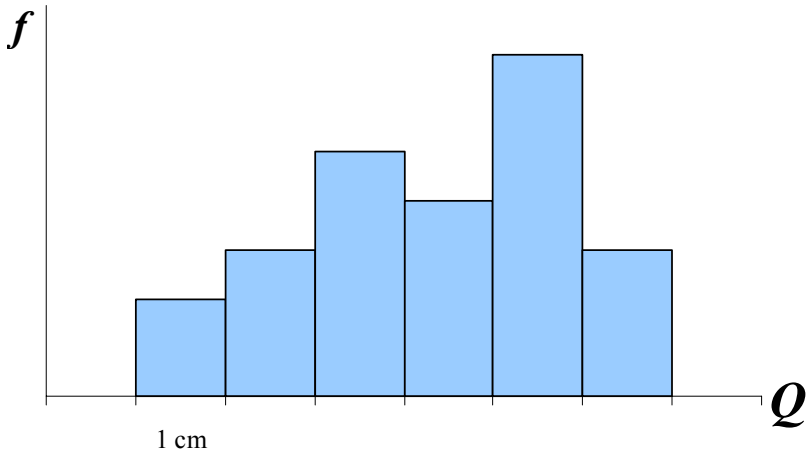
Preguntas:
¿Cómo están repartidos los valores $\{x_1, x_2, \ldots x_n\}$?
¿Dónde se concentran?
¿Cómo se dispersan?

Distribución de Frecuencias:

Sea $f: \begin{array}{l} Q \to IN^{+} \\ q \to f(q) \end{array}$ con $f(q) = Card\{i \in \Omega / x_i = q\}$.

→ f se denomina la distribución de frecuencias de X.

Ejemplo 1.4: $Q = IR$, X: talla de alumnos de IN 540.



Por motivos prácticos Q se particiona en p clases (≡ intervalos):

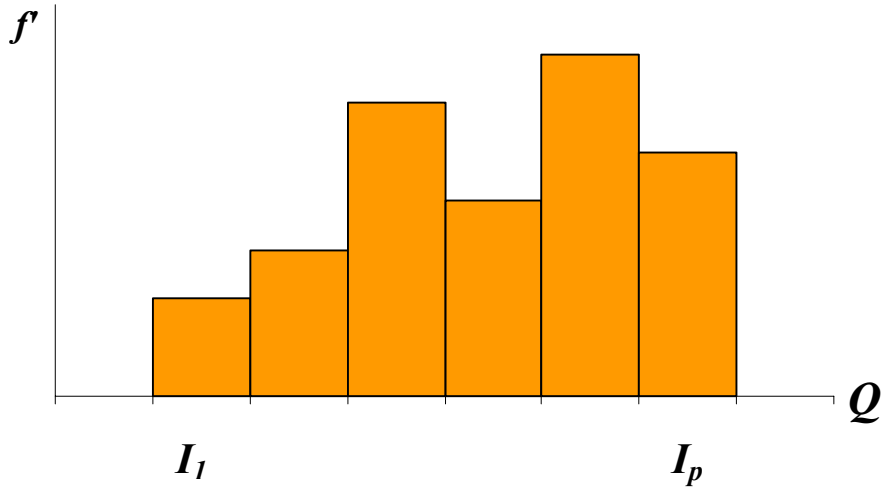
$$Q = \{I_1 \cup I_2 \cup \ldots \cup I_p\}$$

Si llamamos $Q' = \{I_1, I_2, \ldots, I_p\}$ y definimos $f': \begin{array}{l} Q' \to IN^+ \\ I_j \to f(I_j) \end{array}$ con

$$f'(I_j) = Card\{i \in \Omega / x_i \in I_j\}.$$

Observaciones:

- $f'(I_j) = \sum_{q \in I_j} f(q)$
- Al pasar de *f* a *f'* se pierde información. Se tiene aquí el compromiso entre perder información e interpretar mejor o tener mayor información pero poca claridad.



Características de la Distribución de Frecuencias:

a) Características de Posición Central

i. Media Aritmética: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

ii. Media Geométrica: $$g = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

Obs: $$ln(g) = \frac{1}{n}\sum_{i=1}^{n} ln(x_i)$$

iii. Media Armónica: $$h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Estas tres medias se definen para variables cuantitativas.

iv. Mediana: Es el valor M en Q t.q. 50% de Ω toma valores menores que M y 50% toma valores mayores.

Obs: No siempre es única (puede ser un intervalo)

v. Moda: Es el valor M_d en Q t.q. $f(M_d)$ es mayor.

Obs: No siempre es única.

Ambas características son aplicables a toda clase de variables.

b) Valores Extremos

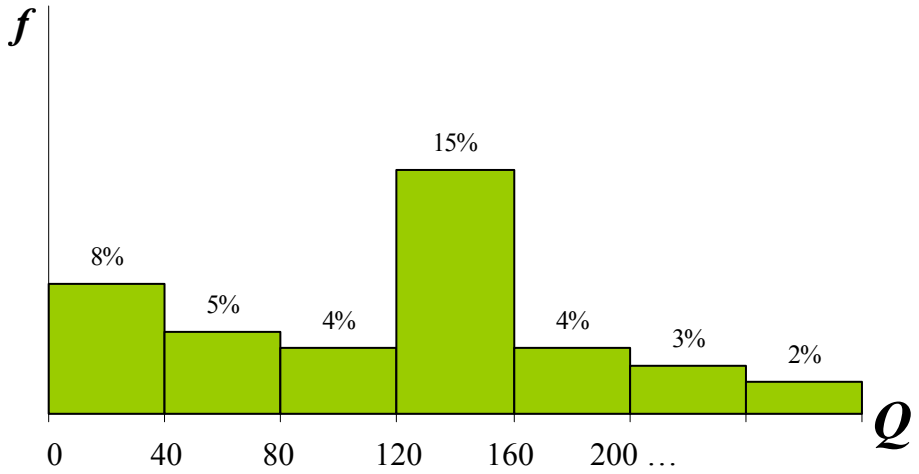
i. Mínimo: $x_m = Min\{x_i / i \in \Omega\}$

ii. Máximo: $x_M = Max\{x_i / i \in \Omega\}$

iii. Cuantila de orden α%: Es el valor C_α t.q. α% de Ω toma valores menores que C_α.

Obs: No siempre es única.
Si α=50 entonces $C_\alpha = M$

Ejemplo 1.5: $Q = [0,400]$ (US$)



$M_d \in [120,160]$

$C_{10} \approx 60$

c) Características de Dispersión

i. Varianza: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

ii. Desviación Típica: S

iii. Rango: $x_M - x_m$

iv. Intervalo Intercuantil: $C_\beta - C_\alpha$, $\alpha<\beta$.

d) Momentos

Se define el *Momento de orden k*: $m_k = \sqrt[k]{\frac{1}{n}\sum_{i=1}^{n} x_i^k}$

<u>Ejercicio</u>:

Muestre que

- $m_0 = g$
- $m_1 = \bar{x}$
- $m_{-1} = h$
- $m_k \xrightarrow{k \to -\infty} x_m$
- $m_k \xrightarrow{k \to \infty} x_M$

Se define el *Momento de orden k centrado en a*: $m_k(a) = \sqrt[k]{\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^k}$

Observación:

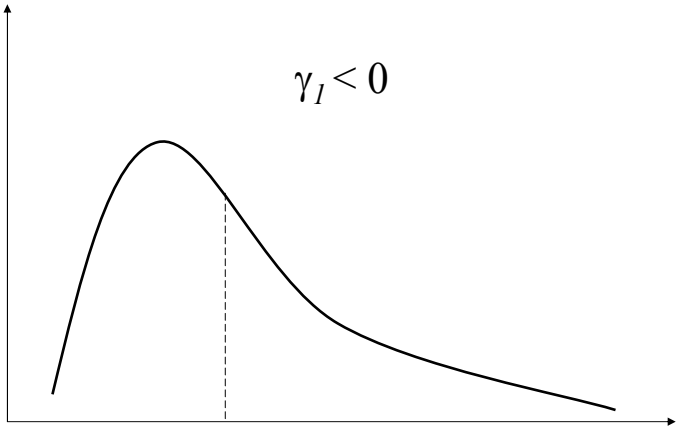
- $m_2(\bar{x}) = S$
- Se anota $\mu_k = m_k(\bar{x})$ al momento de orden k centrado en la media.

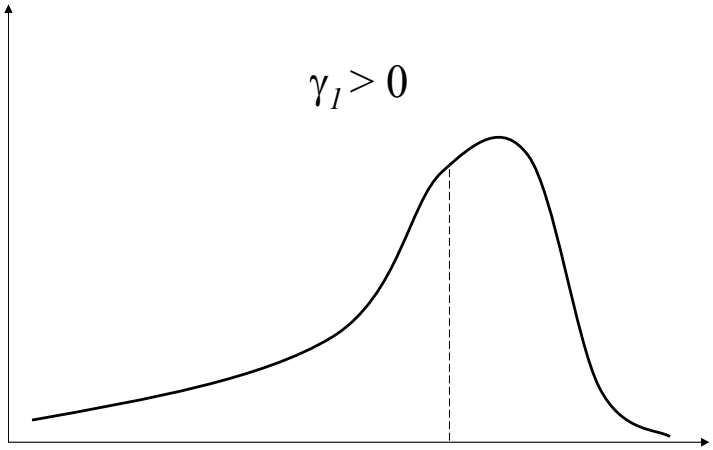
e) Características de Forma:

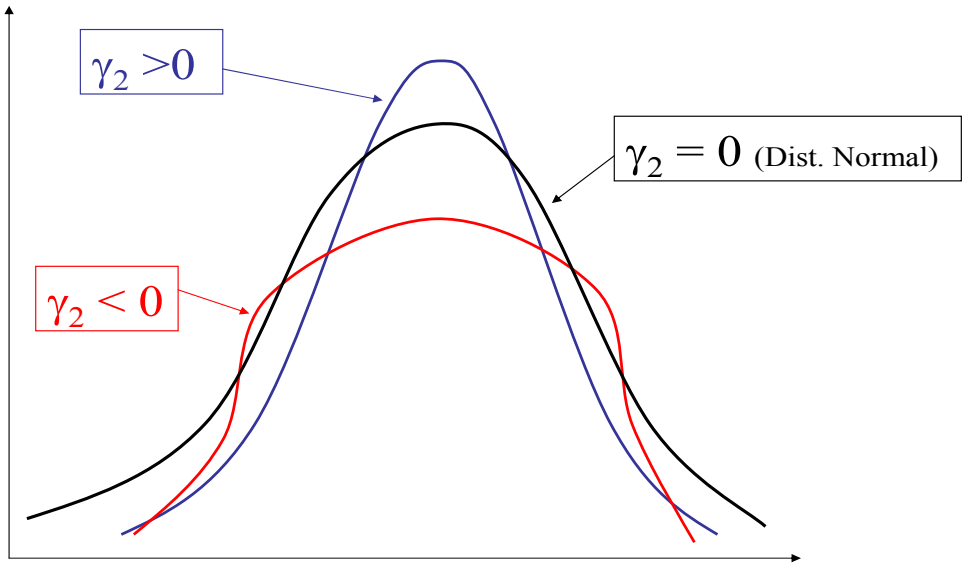
i. Coeficiente de Variación: $$CV = \frac{\mu_2}{\bar{x}} = \frac{S}{\bar{x}}$$

ii. Coeficiente de Asimetría: $$\gamma_1 = \frac{\mu_3}{(\mu_2)^{3/2}}$$

iii. Coeficiente de Achatamiento: $$\gamma_2 = \frac{\mu_4}{(\mu_2)^2} - 3$$


$\gamma_1 < 0$


$\gamma_1 > 0$


$\gamma_2 > 0$
$\gamma_2 = 0$ (Dist. Normal)
$\gamma_2 < 0$

Ejercicios:

Muestre que

- Si $x_1 = x_2 = ... = x_n$ entonces $CV = 0$
- Si la distribución de frecuencias es simétrica con respecto a $\bar{x}$ entonces $\gamma_1 = 0$
- Si los valores de X siguen una distribución normal entonces $\gamma_2 = 0$

VARIABLES ALEATORIAS Y DISTRIBUCIONES

Consideremos $X : \Omega \to IR$ y anotemos $[X \in B] = \{\omega \in \Omega / X(\omega) \in B\}$ donde $B \subset IR$.

Observación:

- $[X \leq x] = \{\omega \in \Omega / X(\omega) \leq x\}$
- $[X = x] = \{\omega \in \Omega / X(\omega) = x\}$

Variable Aleatoria: una v.a. X es una función real definida en Ω (es decir $X : \Omega \to IR$) tal que $[X \leq x]$ es un evento aleatorio $\forall x \in IR$.

X es una v.a. si se puede asignar una probabilidad al evento $[X \leq x]$, $\forall x \in IR$.

Función Distribución: Se llama función de distribución de una v.a. X a la función real $F_X(x) = IP(X \leq x) \ \forall x \in IR$.

Observación:

- F_X también es llamada *función de distribución acumulada de* X.
- F_X es continua sii $IP(X = x) = P_X(x) = 0, \forall x \in IR$.

Tipos de Variables Aleatorias:

a. ***Discreta:*** una v.a. X es *discreta* si toma un número finito o numerable de valores, es decir si existe $\{x_1, x_2, ...\} \subset IR$ tal que $X(\omega) \in \{x_1, x_2, ...\}$, $\forall \omega \in \Omega$.

Notemos que si X es una v.a. discreta $F_X(x) = \sum_{i / x_i \leq x} P_X(x_i)$, $\forall x \in IR$.

b. ***Absolutamente Continua:*** una v.a. X es *absolutamente continua* si existe una función real f_X tal que

- $f_X(x) \geq 0, \ \forall x \in IR.$
- $F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \ \forall x \in IR$

Se llama a f_X *función densidad* (o densidad) de X.

Distribución (o Ley) de una v.a.: Se llama *distribución* de una v.a. X a la probabilidad definida por $P_X(B) = IP(X \in B)$, $\forall B$ evento aleatorio en IR

- Si X es una v.a. discreta $P_X(x) = \sum_{i\,/\,x_i \in B} P_X(x_i)$.
- Si X es una v.a. abs. continua $P_X(B) = \int_B f_X(t)dt$.

- La distribución de una v.a. X está determinada por cualquiera de las siguientes funciones, denominadas *representaciones* de la ley de X:

i. La función de distribución F_X
ii. La densidad f_X, si X es abs. continua
iii. La función de probabilidad P_X, si X es discreta
iv. La función característica φ_X

Ejercicio: ¿Cuál es la definición de φ_X?

Ejemplo 1.6:

1) $X \sim N(0,1)$ sii $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $x \in IR$

2) $X \sim Bernoulli(p)$ sii $IP(X = 1) = p; IP(X = 0) = 1 - p.$

3) $X \sim Binomial(n, p)$ sii $P_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0,1,...,n.$

Ejercicio:

Muestre que si $X_1, X_2, ..., X_n$ son v.a. i.i.d. (variables aleatorias independientes e idénticamente distribuidas) *Bernoulli(p),* entonces $X = \sum_{i=1}^{n} X_i \sim Binomial(n, p).$

Valor Esperado de una v.a.: Se llama valor esperado (o *esperanza)* de una v.a. X que toma valores en Q, al número

- $IE(X) = \sum_{x_i \in Q} x_i P_X(x_i)$, si X es discreta (Q finito o numerable).
- $IE(X) = \int_{-\infty}^{\infty} x f_X(x) dx$, si X es abs. continua.

Ejercicio:

- $X \sim Exp(\lambda)$ sii $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si no} \end{cases}$. Pruebe que $IE(X) = 1/\lambda$.
- $X \sim Cauchy$ sii $f_X(x) = \dfrac{1}{\pi(1+x^2)}$, $x \in IR$. Pruebe que $IE(X)$ no existe.

Varianza una v.a.: Se llama *varianza* de una v.a. X que toma valores en $Q,$ al número

$$V(X) = IE(X - IE(X))^2 = IE(X^2) - IE(X)^2$$

La varianza corresponde a una medida de dispersión de la distribución de X con respecto a su esperanza, por ello se denomina también *desviación cuadrática media.*

Observaciones:

- Se nota $V(X) = Var(X) = \sigma_X^2 = \sigma^2(X)$
- A la raíz cuadrada de la varianza se llama *desviación estándar* $\sigma_X = \sqrt{V(X)}$

Ejercicio:

- Pruebe que $V(X) = 0$ sii existe una constante c tal que $IP(X = c) = 1.$
- Pruebe que cualquiera sean a y b constantes, $V(aX + b) = a^2 V(X).$

TEOREMA CENTRAL DEL LÍMITE, LEY DE LOS GRANDES NÚMEROS Y DESIGUALDAD DE TCHEBYCHEFF

Consideremos $X_1, X_2, ..., X_n$ son v.a. i.i.d. (una *muestra aleatoria*).

Desigualdad de Tchebycheff: Sea $T : IR^n \to IR$. Entonces $\forall \varepsilon > 0$:

$$IP(\left|T(X_1, X_2, ..., X_n) - IE(T(X_1, X_2, ..., X_n))\right| \geq \varepsilon) \leq \frac{V(T(X_1, X_2, ..., X_n))}{\varepsilon^2}$$

En particular si $T(X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \equiv \overline{X_n}$ (la *media muestral*),

entonces $IP(\left|\overline{X_n} - IE(X)\right| \geq \varepsilon) \leq \frac{V(X)}{n\varepsilon^2}$.

Ley de los Grandes Números: $\overline{X}_n \xrightarrow[n\to\infty]{c.s.} IE(X).$

La media muestral converge casi seguramente a la media poblacional.

Teorema Central de Límite:

$$\frac{\overline{X}_n - IE(\overline{X}_n)}{\sqrt{V(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - IE(X))}{\sqrt{V(X)}} \xrightarrow[n\to\infty]{d} N(0,1)$$

La media muestral (estandarizada) converge en distribución a una normal.

PRUEBAS DE HIPÓTESIS

Se formulan hipótesis acerca de leyes o fenómenos físicos o naturales, que es necesario demostrar o rechazar por medio de "contrastes" (tests) o "pruebas". La prueba de la hipótesis es el Contraste de Hipótesis que nos llevará a su aceptación o rechazo.

El procedimiento estándar consiste en recopilar información en forma de observaciones numéricas que serán la base de nuestra decisión.

Por ejemplo si tiramos una moneda 100 veces y obtenemos siempre cara podemos percibir que la hipótesis de que la moneda no está trucada no es aceptable. Sin embargo es posible obtener este resultado con una moneda no trucada, por consiguiente no podremos estar completamente seguros de nuestra decisión.

Los procedimientos de Inferencia Estadística nos posibilitan, bajo ciertas condiciones, establecer la probabilidad de aceptar hipótesis falsas o rechazar hipótesis verdaderas. Es decir permiten calcular la probabilidad de cometer error con nuestra decisión.

El objetivo de un contraste de hipótesis es comprobar si los datos muestrales apoyan la hipótesis nula, o por el contrario rechazan H_0, lo cual nos llevaría a aceptar H_1.

En un enfoque totalmente práctico hay que tener en cuenta dos cosas:

a) La hipótesis nula que se contrasta

b) El *p*-valor obtenido.

Se puede interpretar el *p*-valor de dos formas:

i. La probabilidad de error (o sea de equivocarse) si se rechaza la hipótesis nula cuando realmente es cierta. Es el error llamado de Tipo I.

ii. La probabilidad de que las diferencias observadas sean debidas al azar.

Por ese motivo se rechaza la hipótesis nula cuando el *p*-valor es pequeño. El valor fijo a partir del cual el p-valor se considera pequeño es el nivel de significación a (0.10, 0.05, 0.01, 0.001).