Formulario de Estadística I

Tema 1 Consideramos una muestra de tamaño n de una variable X. Sean $x_1, x_2, \ldots, x_k$, $k \leq n$, los diferentes valores que ha tomado esta variable sobre los n individuos de la muestra. Si X es cuantitativa o categórica ordinal, supondremos que tenemos los valores ordenados de manera que $x_1 < x_2 < \ldots < x_k$.

- Se denota por n_i la **frecuencia absoluta** del valor x_i. Se cumple que $\sum_{i=1}^{k} n_i = n$.
- Se define la **frecuencia relativa** del valor x_i como $f_i = n_i/n$. Se verifica que $\sum_{i=1}^{k} f_i = 1$.
- Se define la **frecuencia absoluta acumulada** del valor x_i como $N_i = \sum_{j=1}^{i} n_j$.
- Se define la **frecuencia relativa acumulada** del valor x_i como $F_i = N_i/n$. Se cumple que $F_i = \sum_{j=1}^{i} f_j$.
- Cuando los datos están agrupados en intervalos de clase, llamamos l_{i-1} y l_i a los **extremos inferior** y **superior**, respectivamente, del intervalo i-ésimo.
- La **longitud** o **amplitud** del intervalo i-ésimo es $L_i = l_i - l_{i-1}$.
- La **marca de clase** del intervalo i-ésimo es $x_i = \frac{l_i + l_{i-1}}{2}$.

Tema 2 Las fórmulas de las siguientes medidas numéricas están expresadas considerando la muestra original, es decir, los n valores observados para X son $x_1, x_2, \ldots, x_n$. La muestra ordenada se denota por $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Medidas de centralización o de tendencia central:

- Media aritmética: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
- Mediana:

$$Me = \begin{cases} \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \text{si } n \text{ es par}, \\ x_{(\frac{n+1}{2})}, & \text{si } n \text{ es impar}. \end{cases}$$

- Moda: La moda es el valor x_i que presenta una mayor frecuencia absoluta (o relativa).

Medidas de posición:

- Cuartiles: $Q_k = x_{(k(n+1)/4)}$ para $k = 1, 2, 3$.
- Percentiles: $P_k = x_{(k(n+1)/100)}$ para $k = 1, 2, \ldots, 99$.

Medidas de dispersión o de variabilidad:

- Rango o amplitud: $R = x_{max} - x_{min}$.
- Rango intercuartílico: $RIC = Q_3 - Q_1$.
- Varianza muestral: $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n} \left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]$.
- Desviación típica (o estándar) muestral: $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$.
- Cuasivarianza muestral: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]$.
- Cuasidesviación típica muestral: $s = \sqrt{s^2}$.
- Coeficiente de variación de Pearson: $CV = s/|\overline{x}|$.

Tema 3 Sea (X, Y) una variable bidimensional que puede tomar $k \times m$ pares de valores diferentes (x_i, y_j), $i = 1, \ldots, k$, $j = 1, \ldots, m$, sobre los n individuos de una muestra. Supondremos que los datos están ordenados de manera que $x_1 < x_2 < \ldots < x_k$, $y_1 < y_2 < \ldots < y_m$, a menos que las variables sean categóricas nominales.

Tabla 1: Estructura de la tabla de doble entrada / tabla de contingencias

$X \setminus Y$	y_1	y_2	$\ldots$	y_j	$\ldots$	y_m	Total
x_1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2m}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{im}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\ldots$	n_{kj}	$\ldots$	n_{km}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet j}$	$\ldots$	$n_{\bullet m}$	$n_{\bullet\bullet}$

- Se denota por n_{ij} la **frecuencia absoluta** del par (x_i, y_j) y $n = n_{\bullet\bullet}$.
- Se define la **frecuencia relativa** del par (x_i, y_j) como $f_{ij} = n_{ij}/n$.
- Se define la **frecuencia absoluta marginal** del valor x_i como $n_{i\bullet} = \sum_{j=1}^{m} n_{ij}$.
- Se define la **frecuencia relativa marginal** del valor x_i como $f_{i\bullet} = n_{i\bullet}/n$.
- Se define la **frecuencia absoluta marginal** del valor y_j como $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$.
- Se define la **frecuencia relativa marginal** del valor y_j como $f_{\bullet j} = n_{\bullet j}/n$.

Características numéricas conjuntas para tablas de doble entrada.

Sean $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ los n pares de valores observados para (X, Y). Entonces:

- Covarianza:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}) = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i\, y_i - n \overline{x}\, \overline{y} \right]$$

- Coeficiente de correlación lineal de Pearson:

$$r_{(x,y)} = \frac{s_{xy}}{s_x\, s_y}$$

La ecuación de la recta de regresión de Y sobre X por el método de los mínimos cuadrados es:

$$\hat{y} = a + b\,x, \quad \text{donde} \quad b = \frac{s_{xy}}{s_x^2}, \;\; a = \overline{y} - b\,\overline{x}.$$

Se denominan **valores predichos (o ajustados)** a los valores $\hat{y}_i = a + b\,x_i$, para $i = 1, \ldots, n$, y se llaman **residuos** a $e_i = y_i - \hat{y}_i$, para $i = 1, \ldots, n$. Se cumple que $\overline{e} = 0$.

Tema 4 Sea Ω el espacio muestral asociado a cierto experimento aleatorio, $B_1, \ldots, B_k$ una partición de Ω, tal que $P(B_i) \neq 0$ para $i = 1, \ldots, k$ y A un suceso cualquiera.

- **Ley de la probabilidad total**:

$$P(A) = P(A \cap B_1) + \ldots + P(A \cap B_k) = P(A|B_1)\, P(B_1) + \ldots + P(A|B_k)\, P(B_k).$$

- **Teorema de Bayes**:

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j)\, P(B_j)}{P(A|B_1)\, P(B_1) + \ldots + P(A|B_k)\, P(B_k)}, \quad \text{para } j = 1, \ldots, k.$$

Esperanza y varianza de una variable aleatoria.

Sea X una v.a. que toma valores en un conjunto S. La esperanza y varianza de X se definen como:

$$\mathsf{E}(X) = \begin{cases} \displaystyle\sum_{x \in S} x\, P(X = x), & \text{si } X \text{ es una v.a. discreta,} \\ \displaystyle\int_S x\, f(x)\, dx, & \text{si } X \text{ es una v.a. continua.} \end{cases}$$

$$\mathsf{var}(X) = \begin{cases} \displaystyle\sum_{x \in S} (x - \mathsf{E}(X))^2\, P(X = x), & \text{si } X \text{ es una v.a. discreta,} \\ \displaystyle\int_S (x - \mathsf{E}(X))^2\, f(x)\, dx, & \text{si } X \text{ es una v.a. continua.} \end{cases}$$

Modelos de probabilidad.

Ley de X	Conjunto S	Función de probabilidad / densidad	$\mathsf{E}(X)$	$\mathsf{var}(X)$
$Ber(p)$	$\{0, 1\}$	$P(X = 1) = p$, $P(X = 0) = 1 - p$	p	$p\,(1 - p)$
$B(n, p)$	$\{0, 1, \ldots, n\}$	$P(X = x) = \binom{n}{x}\, p^x\, (1 - p)^{n - x}$	$n\,p$	$n\,p\,(1 - p)$
$Pois(\lambda)$	$\mathbb{N} \cup \{0\}$	$P(X = x) = e^{-\lambda}\, \frac{\lambda^x}{x!}$	λ	λ
$U(a, b)$	(a, b)	$f(x) = \frac{1}{b - a}$	$(a + b)/2$	$(b - a)^2/12$
$exp(\lambda)$	$\mathbb{R}^+$	$f(x) = \lambda\, e^{-\lambda\, x}$	$1/\lambda$	$1/\lambda^2$
$N(\mu, \sigma)$	$\mathbb{R}$	$f(x) = \frac{1}{\sigma\,\sqrt{2\,\pi}}\, \exp\left\{-\frac{1}{2\,\sigma^2}\,(x - \mu)^2\right\}$	μ	σ^2

Tema 5 Intervalos de confianza al $(1 - \alpha) \times 100\%$

- para la media poblacional (suponiendo varianza conocida) $\left[\overline{x} \mp z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}\right]$,
- para la proporción poblacional $\left[\hat{p} \mp z_{\alpha/2} \sqrt{\dfrac{\hat{p}(1 - \hat{p})}{n}}\right]$,

donde $z_{\alpha/2}$ es el percentil $(1 - \alpha/2) \times 100$ de la ley $N(0, 1)$.