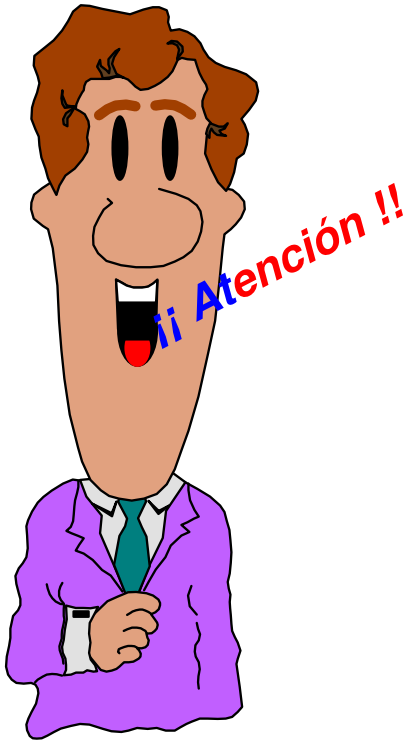
Métodos Estadísticos de la Ingeniería

Tema 8:
Datos categóricos

Contenidos que se tratarán ...

- *Concepto de independencia en contraste de hipótesis*
- *Concepto de dato categórico: experimento multinomial*
- *Tablas unidireccionales*
- *Tablas de contingencia*
- *Concepto de bondad de ajuste*
- *Contraste de modelos*

Dificultades de la unidad didáctica



- Dificultad para analizar el tipo de independencia que se plantea
- Desconocer las distribuciones muestrales de probabilidad
- Proponer el estimador más adecuado a un problema dado, junto con el modelo de probabilidad necesario
- Discriminar el tipo de contraste que se ha de efectuar: determinación de la región crítica
- Obtener la muestra más adecuada a una situación propuesta
- Gestionar el nivel (coeficiente) de significación de la estimación

Un ejemplo de presentación.....

Un ingeniero ha obtenido una correlación para la predicción de la propiedad A de cierta línea de producto a partir de la información que se obtiene cada día de modo rutinario. Propone que se deje de efectuar la medición de la propiedad ya que se puede predecir con exactitud a partir de dicha correlación y los factores en los que se basa dicha correlación están determinados con precisión. El ingeniero de calidad no está convencido y realiza pruebas en 18 muestras de material para comparar los valores observados con los predichos teóricamente, siendo los resultados los de la tabla:

MUESTRA	MEDIDO	PREDICHO
1	78	74
2	59	71
3	56	52
4	94	68
5	84	68
6	81	85
7	66	79
8	78	70
9	59	64
10	56	39
11	88	77
12	88	83
13	75	62
14	75	74
15	72	74
16	81	83
17	84	78
18	73	70

Si fueras el ingeniero de calidad, ¿recomendarías que la toma de medidas, en efecto, fuera cancelada?

Aplicaciones interesantes

(1) **Pruebas de homogeneidad**: pruebas concernientes a proporciones

(a) Probar la igualdad de varias proporciones binomiales
(b) Investigar el cambio de una población multinomial a lo largo del tiempo

(2) **Pruebas de independencia** estadística

(3) **Pruebas de bondad de ajuste**: en cuyo caso el número de grados de libertad ν = k-1-c g.d.l., siendo c el número de parámetros que ha sido necesario estimar

(a) Probar si los datos observados se ajustan a un modelo teórico hipotético
(b) Probar si una población posee una distribución de probabilidad dada

Categorías

Se definen clases o categorías y se registra el número de elementos en cada clase (enumeración o conteo), a consecuencia de un **experimento multinomial**:

(a) n pruebas idénticas e independientes
(b) el resultado de cada prueba cae en alguna de las k clases
(c) la probabilidad de que el resultado de una prueba simple caiga en la clase i es p_i = cte, i = 1, 2, ..., k /

(d) $\sum_{i=1}^{k} n_i = n$ n_i = número de pruebas cuyo resultado cae en la celda i

$$E(n_i) = np_i$$

Estimación de probabilidades de categorías en una tabla unidireccional

Sólo existe una variable cualitativa (es decir, sólo existe un criterio)

1	2	3	.....	k
n_1	n_2	n_3	.....	n_k
p_1	p_2	p_3	.....	p_k

$$\sum_{i=1}^{k} n_i = n$$

$$\sum_{i=1}^{k} p_i = 1$$

La tabla unidireccional se puede convertir en un experimento binomial aislando una categoría ⇒

$$E(\hat{p}_i) = p_i$$

Estimación de probabilidades de categorías en una tabla unidireccional

Para una muestra grande

$$\hat{p}_i = \frac{n_i}{n} \xrightarrow{\text{n} \longrightarrow \infty} N\left[p_i, \sqrt{\frac{p_i(1-p_i)}{n}} \right]$$

$$z = \frac{\hat{p}_i - p_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}} \Rightarrow [L,U] = \hat{p}_i \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}$$

$$\hat{p}_i - \hat{p}_j \Rightarrow$$

$$[L,U] = (\hat{p}_i - \hat{p}_j) \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_i(1-\hat{p}_i) + \hat{p}_j(1-\hat{p}_j) + 2\hat{p}_i\hat{p}_j}{n}}$$

Contraste de hipótesis: tabla unidireccional

H_0: $p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$
H_a: $\exists\, i \,/\, p_i \neq p_{i0}$

Estadística de prueba: **prueba ji cuadrado de Pearson**

$$\chi^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)} = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n$$

que tiene una distribución χ^2 con $\nu = k - 1$ g.d.l. si se trata de una muestra grande (es decir, $E(n_i) = np_i \geq 5\ \forall\ i$)

Región de rechazo:

$$\chi^2 > \chi^2_{\alpha, k-1}$$

Tablas de contingencia

Supongamos que hay dos criterios (direcciones) de selección (o sea, dos variables cualitativas)

Primer criterio de selección

Segundo criterio de selección

	1	2	3	.....	c	
1	n_{11}	n_{12}	n_{13}	.....	n_{1c}	$n_{1\cdot}$
2	n_{21}	n_{22}	n_{23}	.....	n_{2c}	$n_{2\cdot}$
3	n_{31}	n_{32}	n_{33}	.....	n_{3c}	$n_{3\cdot}$
.....	.....	.....	.....	.....	.....	.....
r	n_{r1}	n_{r2}	n_{r3}	.....	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	.....	$n_{\cdot c}$	

Total de filas

Total de columnas

H_0: ¿¿¿¿ Los dos criterios de selección son independientes ????

Tablas de contingencia

Es decir, se trata de analizar la posibilidad de que la probabilidad de que cualquier celda de la tabla sea igual al producto de las probabilidades marginales de fila y columna

Primer criterio de selección

Segundo criterio de selección	1	2	3	.....	c	
1	p_{11}	p_{12}	p_{13}	.....	p_{1c}	$p_{1\bullet}$
2	p_{21}	p_{22}	p_{23}	.....	p_{2c}	$p_{2\bullet}$
3	p_{31}	p_{32}	p_{33}	.....	p_{3c}	$p_{3\bullet}$
.....	.....	.....	.....	.....	.....	.....
r	p_{r1}	p_{r2}	p_{r3}	.....	p_{rc}	$p_{r\bullet}$
	$p_{\bullet 1}$	$p_{\bullet 2}$	$p_{\bullet 3}$	.....	$p_{\bullet c}$	1

Probabilidad marginal de filas

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n} = \frac{Total\ \ fila\ i}{n}$$

$$\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n} = \frac{Total\ columna\ \ j}{n}$$

Probabilidad marginal de columnas

$$p_{ij} = \hat{p}_{i\bullet}\hat{p}_{\bullet j}$$

Tablas de contingencia

H_0: Las dos clasificaciones son independientes

H_a: Las dos clasificaciones son dependientes

$$\hat{E}(n_{ij}) = np_i p_j = n\frac{n_i}{n}\frac{n_j}{n} = \frac{n_i n_j}{n}$$

Estadística de prueba:

$$\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{\left[n_{ij} - \hat{E}(n_{ij})\right]^2}{\hat{E}(n_{ij})} = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{\left[n_{ij} - \frac{n_i n_j}{n}\right]^2}{\frac{n_i n_j}{n}}$$

Si $\hat{E}(n_{ij}) \geq 5$ entonces tiene de forma aproximada una distribución χ^2 con $\nu = (r-1)(c-1)$ g.d.l.

Región de rechazo:

$$\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$$

Tablas de contingencia con totales marginales fijos

Si se desea aumentar la probabilidad de que los conteos esperados (estimados) de celda tengan el tamaño suficiente $\hat{E}(n_{ij}) \geq 5$ se pueden fijar los totales de las filas o de las columnas. Entonces:

H_0: Las proporciones de fila de cada celda no dependen de la fila (≡ las distribuciones de las observaciones en las categorías por columnas son iguales para todas las filas)

H_a: Las proporciones de algunas de las celdas (o todas) dependen de la fila (≡ las distribuciones de las observaciones en las categorías por columna difieren para cuando menos dos de las filas)

Estadística de prueba: $\chi^2 = \sum_{j=1}^{c}\sum_{i=1}^{r} \frac{\left[n_{ij} - \hat{E}(n_{ij})\right]^2}{\hat{E}(n_{ij})}$ $/\ \hat{E}(n_{ij}) = \frac{n_i n_j}{n}$ con $\nu = (r-1)(c-1)$ g.d.l

Región de rechazo: $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$

Simulación de modelos estadísticos

Al realizar una prueba de bondad de ajuste se han de seguir la siguiente estrategia de trabajo

- **Establecer el modelo que se desea contrastar, lo que implicará estimar un número dado "c" de parámetros, a partir de la muestra utilizada**

- **Obtener las frecuencias teóricas explotando de las características propias del modelo estadístico que se piensa satisface la población de la que proviene la serie estadística**

- **Calcular el estadístico del contraste realizando la inferencia correspondiente, a partir del nivel de significación previsto y del número de grados de libertad** $\nu = k - 1 - c$

- **Obtener las conclusiones correspondientes**

Conclusiones

1. Establecer la naturaleza de problema.
2. Dar la hipótesis nula.
3. Proporcionar la hipótesis alternativa.
4. Obtener los estadísticos que sean pertinentes a partir de la muestra proporcionada
5. Calcular el valor del estadístico del contraste.
6. Obtener el valor teórico del estadístico del contraste según el nivel de significación exigido a la prueba y el número de grados de libertad involucrados.
7. Fijar la región crítica.
8. Establecer el contraste: aceptar/rechazar la hipótesis nula.
9. Aplicar los resultados obtenidos al modelo concreto del problema planteado.