CLASE AUXILIAR Y TESTS DE BONDAD DE AJUSTE, INDEPENDENCIA Y ANOVA

HÉCTOR OLIVERO Q. - VÍCTOR RIQUELME F.

1. Un poco de teoría

1.1. Test de Bondad de Ajuste. Recordemos el marco teórico para la distribución multinomial:

- Una muestra de tamaño n.
- $\{y_1, \ldots, y_k\}$ las categorías de donde se pueden extraer los elementos de la población (items).
- X la variable aleatoria que indica la categoría a la que pertenece el item.
- $\vec{p} = (p_1, \ldots, p_k)$ el vector de probabilidades de pertenencia a cada categoría: $\mathbb{P}(X = y_i) = p_i$
- $N = (N_1, \ldots, N_k)$ el vector aleatorio de frecuencias.

Entonces N tiene distribución $Multinomial(n, \vec{p})$.

Teorema 1:
Si N es un vector aleatorio con distribución $Multinomial(n, \vec{p})$, entonces

$$Q = \sum_{j=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

tiene una distribución asintótica χ^2_{k-1} (si $n \to \infty$).

Notemos que np_i es la esperanza de la variable N_i, por lo que se puede escribir el estadístico anterior como

$$Q = \sum_{j=1}^{k} \frac{(N_i - \mathbb{E}(N_i))^2}{\mathbb{E}(N_i)}$$

Supongamos que se tiene el set $(q_i)_{i=1}^k$ donde $\sum_{j=1}^k q_i = 1$, y se quiere contrastar las hipotesis

H_0: $p_i = q_i \ \forall i = 1, \ldots, k$
H_1: $p_i \neq q_i$ para algún $i \in \{1, \ldots, k\}$

Si los valores de q_i están cerca de los valores reales, se tendría que los valores de $\frac{(N_i - nq_i)^2}{nq_i}$ sean pequeños, por lo que si el ajuste es bueno, los valores del estadístico deben ser chicos.

Según el criterio del $p-$valor, si la probabilidad de que el estadístico tome valores mayores que el que tomó es menor a cierto nivel de significación α_0 se rechaza H_0

En el caso que la distribución a la que se quiere hacer el ajuste sea continua, se discretiza, mediante intervalos (un número finito).

1.2. Test de Independencia. Supongamos se tienen dos variables: X e Y (ejemplo: X: Sexo e Y: Comuna de residencia), las que toman valores categóricos en los conjuntos $\{x_1, \ldots, x_n\}$, $\{y_1, \ldots, y_m\}$, (ejemplo: $X \in \{hombre, mujer\}$ e $Y \in \{Las\ \ Condes, Santiago, Maipu\}$) y se realiza una MAS de tamaño N, donde cada item de la muestra presenta alguna característica x_i y alguna característica y_j.

Dada la muestra, se tendrán las frecuencias de pertenencia a la categoría x_i, y_j; denotémosla N_{ij}. Tambien llamemos $p_{ij} = N_{ij}/N$. Definamos, pues $p_i = \sum_{j=1}^m p_{ij}$ la probabilidad de tener la característica x_i; $q_j = \sum_{i=1}^n p_{ij}$ la probabilidad de tener la característica y_j.

Si X e Y fueran independientes, se tendría que la probabilidad de que un objeto tenga las características x_i e y_j fuera igual a $p_i \times q_j$. De esta forma, si hubiera independencia, el valor de $(N_{ij} - Np_iq_j)^2$ sería pequeño.

El test que se usa (para testear la independencia entre X e Y) es

$$Q = \sum_{\substack{i=1,\dots,n \\ j=1,\dots,m}} \frac{(N_{ij} - Np_iq_j)^2}{Np_iq_j} \sim \chi^2_{(n-1)(m-1)}$$

Si el valor del test Q es mayor a cierto valor C, se rechaza la hipótesis de independencia (recordemos también que se puede usar el criterio del $p-$valor).

1.3. Razón de Correlación. Supongamos tenemos dos variables aleatorias X e Y, tales que Y toma valores numéricos (digamos continuos), y X toma valores en un conjunto de categorías finito (digamos $\{X_1, \dots, X_p\}$). Entonces queremos analizar la relación entre Y y X (o sea si Y depende de la categoría a la que pertenezca X). Para ello se realiza una MAS de tamaño n donde en cada item i se observa la característica x_i y el valor y_i. Se considera al *grupo* j como el conjunto de items que tienen la característica X_j, y n_j el número de items en el grupo j.

Definamos

- $\bar{y}_j = \frac{1}{n_j}\sum_{k=1}^{n_j} y_{jk}$ media en el grupo j.
- $\bar{y} = \frac{1}{n}\sum_{j=1}^{p} n_j\bar{y}_j$ media total.
- $s_y^2 = \frac{1}{n}\sum_{l=1}^{n}(y_l - \bar{y})^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{n_j}(y_{jk} - \bar{y})^2$ varianza muestral (y_{jk} es el dato del individuo $k-$ésimo que pertenece al grupo j).
- $w_j^2 = \frac{1}{n_j}\sum_{k=1}^{n_j}(y_{jk} - \bar{y}_j)^2$ la varianza muestral al interior del grupo j.
- $b^2 = \frac{1}{n}\sum_{j=1}^{p} n_j(\bar{y}_j - \bar{y})^2$ la varianza muestral ponderada entre-grupos.
- $\omega^2 = \frac{1}{n}\sum_{j=1}^{p} n_j w_j^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{n_j}(y_{jk} - \bar{y}_j)^2$ la varianza muestral intra-grupos.

Obs: Se tiene que $s_y^2 = b^2 + \omega^2$ (esta es una descomposición de la varianza).

Definamos $\eta^2_{Y|X} = \frac{b^2}{s_y^2}$ la razón de correlación.

Si $\omega^2 = 0$ (o sea, $\eta^2_{Y|X} = 1$), entonces $w_j^2 = 0\ \forall j$; o sea, al interior de los grupos no hay diferencia y se concluye que el pertenecer a un grupo define completamente a la variable Y (relación funcional estricta entre X e Y).

Si $b^2 = 0$ (o sea, $\eta^2_{Y|X} = 0$), entonces no hay difenencia entre las medias de los grupos, por lo que pertenecer a algún grupo no define a la variable Y (no hay relación funcional).

En el caso intermedio (o sea, $\eta^2_{Y|X} \in (0,1)$), se tiene cierta tendencia funcional.

Luego, $\eta^2_{Y|X}$ da una medida de la relación funcional entre Y y X.

1.4. ANOVA a un factor. Supongamos que las observaciones en cada grupo j son normales de varianza σ_j^2. Entonces $\frac{n_j w_j^2}{\sigma_j^2} \sim \chi^2_{n_j-1}$. Suponiendo que $\sigma_j^2 = \sigma^2\ \forall j$, se tiene que $\frac{n\omega^2}{\sigma^2} \sim \chi^2_{n-p}$. También, que $\frac{nb^2}{\sigma^2} \sim \chi^2_{p-1}$.

Consideremos el estadístico

$$F = \frac{\frac{nb^2/\sigma^2}{(p-1)}}{\frac{n\omega^2/\sigma^2}{n-p}} = \frac{b^2/(p-1)}{\omega^2/(n-p)} \sim F_{p-1,n-p}$$

Nuestra hipótesis nula será

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$

Como ya vimos, si no hay relación funcional entre Y y X, b^2 es chico y las medias serían muy parecidas, por lo que si el $p-$valor es chico (menor a cierto nivel de significación) se rechaza la hipótesis (recordemos que el $p-$valor es $\mathbb{P}\left(F \geq \bar{F}\right)$). Esto significa que el factor es significativo.

Podemos escribir nuestra información en una tabla, como sigue:

	S.C.	G.L.	C.M.	F	$p-$valor
Factor	nb^2	$p-1$	$nb^2/(p-1)$	$\frac{nb^2/(p-1)}{n\omega^2/(n-p)}$	$\mathbb{P}\left(F > \bar{F}\right)$
Errores	$n\omega^2$	$n-p$	$n\omega^2/(n-p)$		
Total	ns_y^2	$n-1$			

2. Problemas

Problema 1:
Supóngase que la distribución de las estaturas de los hombres que residen en cierta gran ciudad es una normal de media 68 pulgadas y varianza 1 pulgada^2. Supóngase ademas que cuando se midieron las estaturas de 50 hombres que residen en cierto barrio de la ciudad se obtuvo la siguiente distribución:

Estaturas	Número de hombres
Menos de 66 pulgadas	18
Entre 66 y 67.5 pulgadas	177
Entre 67.5 y 68.5 pulgadas	198
Entre 68.5 y 70 pulgadas	102
Más de 70 pulgadas	5

Contrástese la hiótesis de que, en lo que se refiere a la estatura, estos 500 hombres constituyen una MAS de todos los hombres que residen en la ciudad.

Problema 2:
Se ha tomado una muestra de 90 motores de cierta marca y se ha medido el tiempo de funcionamiento en miles de horas, hasta que fallan por primera vez, obteniendose los siguientes resultados:

Tiempo	Frecuencia
$(0, 1]$	35
$(1, 2]$	26
$(2, 3]$	12
$(3, 4]$	6
Más de 4	11

¿Se puede aceptar que el tiempo hasta el fallo de estos motores sigue una distribución exponencial?

Problema 3:
Supongase que se seleccionan 300 personas al azar de una gran población y que cada persona de la muestra de clasifica según su tipo de sangre: 0, A, B o AB, también si su Rh es positivo o negativo. Los números observados son los de la tabla siguiente:

	0	A	B	AB	Total
Rh positivo	82	89	54	19	244
Rh negativo	13	27	7	9	56
Total	95	116	61	28	300

Testée la hipótesis de que las dos clasificaciones de tipo de sangre son independientes.

Problema 4:
Un agricultor quiere analizar la influencia de 4 grupos F1S1,F2S2,F1S2,F2S1 del factor "fertilizante-suelo" con la producción de choclos. Se obtiene la siguiente tabla:

Fertilizante-Suelo	Frecuencia	Media producción	Desviación Standard
F1S1	20	16.23	1.710
F2S1	20	13.38	1.940
F1S2	30	10.94	1.856
F2S2	30	8.97	1.394
Total	100	11.9	3.179

Haga un test que indique o que le permita deducir la igualdad de las medias para los 4 grupos.

3. Resolución

Solución (Problema 1)**:**
Lo que se pide es ver si la distribución de las categorías de los hombres del barrio se ajusta a la distribución de la estatura de los hombres de toda la ciudad.

Definamos los intervalos $I_1 = (-\infty, 66)$, $I_2 = (66, 67.5)$, $I_3 = (67.5, 68.5)$, $I_4 = (68.5, 70)$, $I_5 = (70, \infty)$; las probabilidades de que un hombre *de la ciudad* pertenezca a estos intervalos son $p_1 = 0.0227$, $p_2 = 0.2858$, $p_3 = 0.383$, $p_4 = 0.2858$, $p_5 = 0.0227$ (viene de normalizar la distribución, y usar que $\mathbb{P}(Z < 0) = 0.5$, $\mathbb{P}(Z < 0.5) = 0.6915$, $\mathbb{P}(Z < 2) = 0.9773$). Si suponemos que la distribución de los hombres del barrio es representativa de los hombres de la ciudad completa, la probabilidad de que la altura un hombre del barrio pertenezca al intervalo I_j sería p_j, $j = 1, \ldots, 5$.

El valor del estadístico (que se distribuye como una χ^2_4) es

$$\begin{aligned}\bar{Q} &= \frac{(18-11.35)^2}{11.35} + \frac{(177-142.9)^2}{142.9} + \frac{(198-191.5)^2}{191.5} + \frac{(102-142.9)^2}{142.9} + \frac{(5-11.35)^2}{11.35} \\ &= 3.9 + 8.14 + 0.22 + 11.71 + 3.55 \\ &= 27.52\end{aligned}$$

La probabilidad $\mathbb{P}\left(Q > \bar{Q}\right) = \mathbb{P}(Q > 27.52) < \mathbb{P}(Q > 14.86) = 0.005 < 0.05$, por lo que se rechaza H_0, o sea, los hombres del barrio no son representativos del total de la ciudad. ☯

Observación: si no se conocieran los valores de la media de la normal y de la varianza, se pueden estimar, pero la distribución χ^2 del estadístico pierde grados de libertad por los datos estimados.

Solución (Problema 2)**:**
Primero, para postular la distribución a ser testeada, estimamos el parámetro de la exponencial (el EMV de λ es $\hat{\lambda} = \frac{1}{\bar{X}_n}$).

$$H_0 : T \sim Exp(\hat{\lambda})$$

$$\begin{aligned}\hat{\lambda} &= \frac{1}{\bar{T}} = \frac{90}{0.5 \times 35 + 1.5 \times 26 + 2.5 \times 12 + 3.5 \times 6 + 4.5 \times 11} \\ &= \frac{90}{157} \\ &= 0.57\end{aligned}$$

Ahora necesitamos el vector de probabilidades a testear. En este caso, será el vector $\vec{p} \in \mathbb{R}^5$ tal que $p_i = \mathbb{P}(T \in I_i)$ ($I_1 = (0, 1]$, $I_2 = (1, 2]$, $I_3 = (2, 3]$, $I_4 = (3, 4]$, $I_5 = (4, \infty]$, y una distribución exponencial de parametro $\hat{\lambda}$).

$$\begin{aligned}\mathbb{P}(\alpha < T \leq \beta) &= \mathbb{P}(\alpha \leq T) - \mathbb{P}(\beta \leq T) \\ &= e^{-\hat{\lambda}\alpha} - e^{-\hat{\lambda}\beta}\end{aligned}$$

Luego, $\vec{p} = (0.43; 0.25; 0.14; 0.08; 0.1)$. Entonces el estadístico (distribuido como χ^2_{5-1-1}) (se resta uno más por la estimación del parámetro) toma el valor $\bar{Q} = 1.57$.

El $p-$valor es $\mathbb{P}(Q > 1.57) \in (0.6, 0.7) > 0.05$, con lo que se concluye que no se rechaza la hipótesis de que los tiempos de falla sean exponenciales. ☯

Solución (Problema 3)**:**
Llamemos $X = Rh$, $Y =$Grupo sanguíneo. Las probabilidades marginales son $p_{Rh^+} = 0.81$, $p_{Rh^-} = 0.19$; $q_0 = 0.32$, $q_A = 0.39$, $q_B = 0.2$, $q_{AB} = 0.09$.

La tabla de frecuencias teóricas (las de la forma Np_iq_j) es:

	0	A	B	AB
Rh positivo	77	94	50	23
Rh negativo	18	22	11	5

Entonces, el valor del estadístico es $\bar{Q} = 8.8$. La probabilidad de que el estadístico tome valores mayores que 8.8 es (recordando que $Q \sim \chi^2_3$) $\mathbb{P}(Q > 8.8) \in (0.025, 0.05)$, por lo que con un nivel de significancia de 0.05 se rechaza la hipótesis de independencia. ☯

Solución (Problema 4)**:**
Calculemos las varianzas entregrupo e intragrupo:

$$b^2 = \sum_{j=1}^{4} \frac{n_j}{n}(\bar{y}_j - \bar{y})^2 = 7.04$$

$$\omega^2 = \sum_{j=1}^{4} \frac{n_j}{n} w_j^2 = 2.954$$

Entonces, haciendo la tabla

	S.C.	G.L.	C.M.	F	$p-$valor
Factor	704	4-1=3	234.7	76.2	0.00
Errores	295.4	100-4=96	3.08		
Total	999.4	99			

Analizando el $p-$valor, rechazamos la hipótesis, y por lo tanto concluimos que el tipo de suelo es relevante para la producción de choclos.

Ahora, la razón de correlación es

$$\eta^2_{Y|X} = \frac{b^2}{s_y^2} = \frac{nb^2}{ns_y^2} = \frac{704}{999.4} = 0.7$$

Podemos concluir que existe una tendencia funcional entre la producción de choclos y el tipo de fertilizante. ☯

Observación: En la tabla que es dada en el enunciado de esta pregunta son dadas las desviaciones estandar por grupos. Cada una de estas es la raiz de la varianza por cada grupo. Estas últimas son las w_j. Además, la desviación estandar total dada en la tabla es la raiz de la varianza muestral. Esta última es $s_y^2 = b^2 + \omega^2$.