Universidad de Chile		Rodrigo Assar
FCFM	MA34B	Andrés Iturriaga
DIM		Víctor Riquelme

Análisis de la Varianza (ANOVA) y Correlación

Resumen

El test ANOVA analiza la relación entre una variable numérica y una categórica, y ve si hay relación funcional entre ambas. La correlación nos da una medida de la relación lineal entre dos variables numéricas.

1. ANOVA

1.1. 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 $\{x_1, \ldots, x_p\}$). Entonces queremos analizar la relación entre Y y X (o sea si Y depende de la categoría a la que pertenezca X).

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 al interior del grupo j.
- $b^2 = \frac{1}{n} \sum_{j=1}^{p} n_j (\bar{y}_j - \bar{y})^2$ la varianza 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 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.

1.2. 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 $\frac{nb^2}{\sigma^2} \sim \chi^2_{p-1}$ (suma de las varianzas de los p grupos).

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

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

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. Covarianza y Correlación

2.1. Teóricos

Primero, recordemos la definición de *covarianza* y *correlación*.

Definición 1 *Sean X e Y dos variables aleatorias con alguna distribución conjunta; sean $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y$, $\mathbb{V}ar(X) = {\sigma_X}^2$, $\mathbb{V}ar(Y) = {\sigma_Y}^2$. Definimos la "covarianza" entre X e Y, denotada por $\mathbb{C}ov(X,Y)$, por:*

$$\mathbb{C}ov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$

También, definamos la "correlación" entre X e Y por

$$\rho(X,Y) = \frac{\mathbb{C}ov(X,Y)}{\sigma_X \sigma_Y}$$

Hay una propiedad que permite decir algo importante acerca de la correlación:

Teorema 1 (Desigualdad de Schwarz) *Para cualquier par de variables aleatorias U, V se tiene que*

$$|\mathbb{E}[UV]| \leq \left(\mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right]\right)^{\frac{1}{2}}$$

Dem Si $\mathbb{E}\left[U^2\right] = 0$ entonces $\mathbb{P}(U=0) = 1$. Luego, $\mathbb{P}(UV=0) = 1$, por lo que $\mathbb{E}[UV] = 0$, y la desigualdad se tiene.
Si $\mathbb{E}\left[U^2\right] = \infty$, la desigualdad se tiene trivialmente.
Supongamos que $0 < \mathbb{E}\left[U^2\right] < \infty$, $0 < \mathbb{E}\left[V^2\right] < \infty$. Sabemos que $\forall a, b$

$$\begin{aligned} 0 &\leq \mathbb{E}\left[(aU+bV)^2\right] = a^2\mathbb{E}\left[U^2\right] + b^2\mathbb{E}\left[V^2\right] + 2ab\mathbb{E}[UV] \\ -(a^2\mathbb{E}\left[U^2\right] + b^2\mathbb{E}\left[V^2\right]) &\leq 2ab\mathbb{E}[UV] \end{aligned}$$

y de igual manera

$$\begin{aligned} 0 &\leq \mathbb{E}\left[(aU-bV)^2\right] = a^2\mathbb{E}\left[U^2\right] + b^2\mathbb{E}\left[V^2\right] - 2ab\mathbb{E}[UV] \\ 2ab\mathbb{E}[UV] &\leq (a^2\mathbb{E}\left[U^2\right] + b^2\mathbb{E}\left[V^2\right]) \end{aligned}$$

Elegimos $a = \sqrt{\mathbb{E}\left[V^2\right]}$, $b = \sqrt{\mathbb{E}\left[U^2\right]}$. Entonces

$$\begin{aligned} 2\sqrt{\mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right]}\mathbb{E}[UV] &\leq (\mathbb{E}\left[V^2\right]\mathbb{E}\left[U^2\right] + \mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right]) \\ 2\sqrt{\mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right]}\mathbb{E}[UV] &\leq 2\mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right] \\ \mathbb{E}[UV] &\leq \sqrt{\mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right]} \end{aligned}$$

de igual forma

$$-\sqrt{\mathbb{E}\left[U^2\right]\mathbb{E}\left[V^2\right]} \leq \mathbb{E}\left[UV\right]$$

y eso concluye el resultado. ☯

Ahora, lo que nos interesa es lo siguiente: si llamamos $U = X - \mu_X$, $V = Y - \mu_Y$, obtenemos que

$$\begin{aligned} \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] &\leq \sqrt{\mathbb{E}\left[(X-\mu_X)^2\right]\mathbb{E}\left[(Y-\mu_Y)^2\right]} \\ \mathbb{C}ov\left(X,Y\right) &\leq \sqrt{\mathbb{V}ar\left(X\right)\mathbb{V}ar\left(Y\right)} \end{aligned}$$

Entonces, $\rho\left(X,Y\right) \in [-1,1]$.

Proposición 1 *Para cualquier par de variables aleatorias X e Y, con $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty$, se tiene que $\mathbb{C}ov\left(X,Y\right) = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$*

Dem

$$\begin{aligned} \mathbb{C}ov\left(X,Y\right) &= \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\ &= \mathbb{E}\left[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y\right] \\ &= \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mu_Y - \mathbb{E}\left[Y\right]\mu_X + \mu_X\mu_Y \\ &= \mathbb{E}\left[XY\right] - \mu_X\mu_Y \end{aligned}$$

☯

Proposición 2 *Si X e Y son variables aleatorias independientes con $0 < \sigma_X^2 < \infty$, $0 < \sigma_Y^2 < \infty$, entonces $\mathbb{C}ov\left(X,Y\right) = \rho\left(X,Y\right) = 0$*

Dem Como X e Y son independientes, entonces $\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$. Por la proposición anterior, se concluye. ☯

Proposición 3 *Suponga que X es una variable aleatoria con varianza positiva finita, y que $Y = aX + b$ para algunas constantes a, b, $a \neq 0$. Entonces, si $a > 0$, $\rho\left(X,Y\right) = 1$; si $a > 0$, $\rho\left(X,Y\right) = -1$*

Dem Primero vemos que $\mathbb{E}\left[Y\right] = \mathbb{E}\left[aX+b\right] = a\mathbb{E}\left[X\right] + b = a\mu_X + b$
$\mathbb{V}ar\left(Y\right) = \mathbb{V}ar\left(aX+b\right) = a^2\mathbb{V}ar\left(Y\right) = a^2\sigma_X^2$.

$$\rho\left(X,Y\right) = \frac{\mathbb{C}ov\left(X,Y\right)}{\sqrt{\mathbb{V}ar\left(X\right)\mathbb{V}ar\left(Y\right)}}$$

$$
\begin{aligned}
&= \frac{\mathbb{E}\left[(X-\mu_x)(aX+b-a\mu_X-b)\right]}{\sqrt{\sigma_X^2 a^2 \sigma_X^2}} \\
&= \frac{a\mathbb{E}\left[(X-\mu_X)^2\right]}{|a|\sigma_X^2} \\
&= \frac{a}{|a|}\frac{\mathbb{V}ar\left(X\right)}{\sigma_X^2} \\
&= \frac{a}{|a|}
\end{aligned}
$$

Si $a > 0$, lo anterior vale 1, y si $a < 0$, lo anterior vale -1.

Obs: Para constantes a, b,

$$\mathbb{C}ov\left(aX, bY\right) = ab\mathbb{C}ov\left(X, Y\right)$$

pues

$$\mathbb{C}ov\left(aX, bY\right) = \mathbb{E}\left[(aX)(bY)\right] - \mathbb{E}\left[aX\right]\mathbb{E}\left[bY\right] = ab(\mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right])$$

Teorema 2 *Si $X_1, \ldots, X_n$ son v.a. con varianza finita, entonces*

$$\mathbb{V}ar\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{V}ar\left(X_i\right) + \sum_{\substack{j,k \\ j \neq k}} \mathbb{C}ov\left(X_j, X_k\right) = \sum_{i=1}^n \mathbb{V}ar\left(X_i\right) + 2\sum_{\substack{j,k \\ j < k}} \mathbb{C}ov\left(X_j, X_k\right)$$

Dem Propuesta (para $n = 2$ se ve fácil).

2.2. Empíricos

Ahora consideremos una muestra aleatoria bivariada $\{(x_i, y_i)\}_{i=1}^n$ del par de variables aleatorias (X, Y). Denotemos:

- $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ las medias empíricas.
- $s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2$, $s_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^n y_i^2 - \bar{y}^2$ las varianzas empíricas.
- $cov_{x,y} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$ la covarianza empírica.
- $r_{X,Y} = \frac{cov_{x,y}}{s_x s_y}$ es el coeficiente de correlación empírico.

Calculando el valor $r_{X,Y}$ se puede decir algo acerca de la correlación entre X e Y, de forma análoga a lo dicho para $\rho\left(X, Y\right)$.

3. Problemas

3.1. Problema 1 [Ajuste χ^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?

3.2. Problema 2 [ANOVA]

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.

3.3. Problema 3

(a) Supóngase que X e Y son variables aleatorias que representan las mediciones de la alturas del padre y del hijo respectivamente, donde $\mu_X = \mu_Y = \mu$, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Sean X', Y' v.a. tales que $X = X' + \epsilon$; $Y = Y' + \eta$ donde ϵ, η son los errores de medición, y X',Y' son los valores *reales* de las alturas; supóngase también que η, ϵ son independientes de X' e

Y' e independientes entre si.

Exprese la correlación entre X e Y en términos de la correlación entre X' e Y'. Concluya.

(b) Se obtiene una muestra de 20 familias y se obtienen los siguientes resultados:

- $\bar{x} = 170.92$; $\bar{y} = 170.93$.
- $\frac{1}{20}\sum_{i=1}^{20} x_i^2 = 29239$; $\frac{1}{20}\sum_{i=1}^{20} y_i^2 = 29349$.
- $\frac{1}{20}\sum_{i=1}^{20} x_i y_i = 29300$.

Deduzca el coeficiente de correlación entre las tallas observadas del padre y del hijo.

Si se supone que la varianza de los errores de medición es 9, de una estimación de la correlación $\rho(X', Y')$ entre las tallas verdaderas X', Y'. Comente

4. Resolución de los problemas

4.1. Problema 1

Primero, para postular la distribución a ser testeada, estimamos el parámetro de la exponencial.

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

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

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. ☯

4.2. Problema 2

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. ☯

4.3. Problema 3

Parte (a)

- Primero calculemos la varianza de X:

$$\mathbb{V}ar\,(X) = \mathbb{V}ar\,(X' + \epsilon) = \mathbb{V}ar\,(X') + \mathbb{V}ar\,(\epsilon) = \sigma^2_{X'} + \mathbb{V}ar\,(\epsilon) \geq \sigma^2_{X'}$$

$$\sigma^2_X \geq \sigma^2_{X'}$$

- De igual manera, para Y:

$$\mathbb{V}ar\,(Y) = \mathbb{V}ar\,(Y' + \eta) = \mathbb{V}ar\,(Y') + \mathbb{V}ar\,(\eta) = \sigma^2_{Y'} + \mathbb{V}ar\,(\eta) \geq \sigma^2_{Y'}$$

$$\sigma^2_Y \geq \sigma^2_{Y'}$$

- Ahora, la covarianza entre X e Y:

$$\mathbb{C}ov\,(X,Y) = \mathbb{C}ov\,(X'+\epsilon, Y'+\eta) = \mathbb{C}ov\,(X',Y') + \mathbb{C}ov\,(X',\eta) + \mathbb{C}ov\,(\epsilon,Y') + \mathbb{C}ov\,(\epsilon,\eta)$$

$$\mathbb{C}ov\,(X,Y) = \mathbb{C}ov\,(X',Y')$$

Luego,

$$\rho\,(X,Y) = \frac{\mathbb{C}ov\,(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathbb{C}ov\,(X',Y')}{\sqrt{(\sigma^2_{X'} + \mathbb{V}ar\,(\epsilon))(\sigma^2_{Y'} + \mathbb{V}ar\,(\eta))}}$$

$$\rho\,(X,Y) \leq \rho\,(X',Y')$$

Podemos concluir que la correlación de los datos de la muestra subestima la correlación de los datos reales. ☯

Parte (b)

Como estimador de la correlación entre X e Y usaremos el coeficiente de correlación empírico $r_{X,Y}$

$$\begin{aligned}
r_{X,Y} &= \frac{\frac{1}{20}\sum_{i=1}^{20}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{20}\sum_{i=1}^{20}(x_i-\bar{x})^2}\sqrt{\frac{1}{20}\sum_{i=1}^{20}(y_i-\bar{y})^2}} \\
&= \frac{\frac{1}{20}\sum_{i=1}^{20}x_iy_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{20}\sum_{i=1}^{20}x_i^2 - \bar{x}^2}\sqrt{\frac{1}{20}\sum_{i=1}^{20}y_i^2 - \bar{y}^2}} \\
&= \frac{29300 - 170.92 \times 170.93}{\sqrt{29293 - 170.92^2}\sqrt{29349 - 170.93^2}} \\
&= \frac{84.64}{8.91 \times 11.47} \\
&= \frac{84.64}{102.2} \\
&= 0.828
\end{aligned}$$

Para estimar la correlación entre los valores reales, consideremos

$$r_{X',Y'} = \frac{\sigma_X\sigma_Y}{\sigma_{X'}\sigma_{Y'}} r_{X,Y}$$

donde

$$\begin{aligned}
\sigma^2_X &= \sigma^2_{X'} + \mathbb{V}ar\,(\epsilon) \\
79.35 &= \sigma^2_{X'} + 9 \\
\sigma_{X'} &= 8.39
\end{aligned}$$

$$
\begin{aligned}
\sigma_Y^2 &= \sigma_{Y'}^2 + \mathbb{V}ar\,(\eta) \\
131.94 &= \sigma_{Y'}^2 + 9 \\
\sigma_{Y'} &= 11.09
\end{aligned}
$$

Luego,

$$
\begin{aligned}
r_{X',Y'} &= \frac{8.91 \times 11.47}{8.39 \times 11.09} 0.828 \\
&= \frac{84.64}{93.05} \\
&= 0.91
\end{aligned}
$$

Aquí se nota cuanto se subestima la correlación de los valores reales con el uso de los datos muestrales ☯

Referencias

[DeGroot;1986] DeGroot, M.H.; *"Probability and Statistics, Second Edition"*; Addison-Wesley S.A.; pp. 213-219; 1986.

[Lacourly;2004] Lacourly, N.; *"Estadística"*; Depto. de Publicaciones DIM ; pp. 90-104; 2004.