Clase Auxilar 14: Resumen Control 3

Prof. Rodrigo Abt B. - Aux. Héctor Olivero Q.

15 de junio de 2008

La distribución Multinomial:

Consideremos n repeticiones independientes de un experimento con k posibles resultados. Supongamos que la probabilidad de cada resultado está dada por $\{p_i\}_{i=1}^{k}$.
Si denotamos X_i como la cantidad de veces que se repite el resultado i en las n repeticiones, entonces tenemos:

$$\mathbb{P}(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Siempre que $n_1 + \ldots + n_k = n$.

Ejemplo 1:

Si $k = 2$ entonces el modelo multinomial se reduce al modelo binomial. En efecto:
Sea: $p_1 = p$ entonces $p_2 = 1 - p$. Además si $n_1 + n_2 = n$, entonces $n_2 = n - n_1$. Con esto:

$$\mathbb{P}(X_1 = n_1, X_2 = n_2) = \frac{n!}{n_1! n_2!} p_1^{n_1} p_2^{n_2} \quad (1)$$

$$\mathbb{P}(X_1 = n_1, X_2 = n - n_1) = \frac{n!}{n_1!(n - n_1)!} p^{n_1} (1 - p)^{n - n_1} \quad (2)$$

$$= \binom{n}{n_1} p^{n_1} (1 - p)^{n - n_1} \quad (3)$$

Ejemplo 2:

Consideremos el caso de un dado de seis caras. ¿Cuál es la probabilidad de obtener dos unos, dos tres y dos seis, cuando se lanza un dado seis veces?.
Utilizando el resultados anterior tenemos:

$$\mathbb{P}(X_1 = 2, X_2 = 0, X_3 = 2, X_4 = 0, X_5 = 0, X_6 = 2) =$$

$$\frac{6!}{2!0!2!0!0!2!}(\frac{1}{6})^2(\frac{1}{6})^0(\frac{1}{6})^2(\frac{1}{6})^0(\frac{1}{6})^0(\frac{1}{6})^2$$

$$\mathbb{P}(X_1 = 2, X_2 = 0, X_3 = 2, X_4 = 0, X_5 = 0, X_6 = 2) = 0{,}0019$$

Ejemplo 3:

Sea X una v.a. que sigue una distribución $U(0,1)$ y sea:

$$Y = \begin{cases} 1 & X \in [0, \frac{1}{6}] \\ 2 & X \in (\frac{1}{6}, \frac{4}{6}] \\ 3 & X \in (\frac{4}{6}, 1] \end{cases} \tag{4}$$

Notemos que:

$$\mathbb{P}(Y = 1) = \frac{1}{6}, \quad \mathbb{P}(Y = 2) = \frac{1}{2}, \quad \mathbb{P}(Y = 3) = \frac{1}{3}$$

Supongamos que se repite n veces el experimeno Y, entonces:

$$\mathbb{P}(X_1 = n_1, X_2 = n_2, X_3 = n_3) = \frac{n!}{n_1! n_2! n_3!} (\frac{1}{6})^{n_1} (\frac{1}{2})^{n_2} (\frac{1}{3})^{n_3}$$

Si $n_1 + n_2 + n_3 = n$.

Proposición

Sea:

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

Entonces asintóticamente Q sigue una distribución χ^2_{k-1}

Observación: Si para calcular Q es necesario estimar l parámetros, entonces Q sigue una distribución χ^2_{k-l-1}

La idea:

Sea una v.a. X sigue cierta distribución de probabilidad. Si separamos en k subconjuntos los valores que puede tomar X (como en el ejemplo 3) y definimos una nueva v.a. Y que toma valores de 1 a k según el subconjunto en que está el resultado de la realización de X, entonces tenemos que al repetir n veces Y el número de apariciones conjuntas de los posible resultados[1] sigue una distribución multinomial.

Esto lo usamos para realizar el siguiente test de hipótesis:

H_0 : X sigue una distribución F_θ

H_1 : X no sigue una distribución F_θ

Podemos proceder de la siguiente manera: Suponemos que H_0 es cierta y construimos la variable aleatoria Y, con ella construimos el estadístico Q con el cual podemos calcular p-valores o regiones de rechazo.

[1] Ver ejemplo 3

La idea:

Sea una v.a. X sigue cierta distribución de probabilidad. Si separamos en k subconjuntos los valores que puede tomar X (como en el ejemplo 3) y definimos una nueva v.a. Y que toma valores de 1 a k según el subconjunto en que está el resultado de la realización de X, entonces tenemos que al repetir n veces Y el número de apariciones conjuntas de los posible resultados[1] sigue una distribución multinomial.
Esto lo usamos para realizar el siguiente test de hipótesis:
H_0 : X sigue una distribución F_θ
H_1 : X no sigue una distribución F_θ
Podemos proceder de la siguiente manera: Suponemos que H_0 es cierta y construimos la variable aleatoria Y, con ella construimos el estadístico Q con el cual podemos calcular p-valores o regiones de rechazo.

[1] Ver ejemplo 3

La idea:

Sea una v.a. X sigue cierta distribución de probabilidad. Si separamos en k subconjuntos los valores que puede tomar X (como en el ejemplo 3) y definimos una nueva v.a. Y que toma valores de 1 a k según el subconjunto en que está el resultado de la realización de X, entonces tenemos que al repetir n veces Y el número de apariciones conjuntas de los posible resultados[1] sigue una distribución multinomial.
Esto lo usamos para realizar el siguiente test de hipótesis:
H_0 : X sigue una distribución F_θ
H_1 : X no sigue una distribución F_θ
Podemos proceder de la siguiente manera: Suponemos que H_0 es cierta y construimos la variable aleatoria Y, con ella construimos el estadístico Q con el cual podemos calcular p-valores o regiones de rechazo.

[1] Ver ejemplo 3

Independencia

El test para detectar independencia entre dos variables X e Y es un caso particular de lo anterior. En este caso, lo que testeamos es que la distruibución conjunta de las dos variables sea el producto de las distribuciones marginales. El estadístico que usamos es:

$$Q = \sum_{i,j} \frac{(M_{i,j} - np_ip_j)^2}{np_ip_j}$$

Con $p_i = \mathbb{P}(X = i)$, $p_j = \mathbb{P}(Y = j)$ y $M_{i,j} = frec(X = i, Y = j)$, que asintóticamente sigue una distribución χ^2_{pq-1}. Si no conocemos p_i y p_j usamos estimaciones de los mismos. Y en este caso tenemos que Q sigue una distribución $\chi^2_{(p-1)(q-1)}$.

El coeficiente de Correlación Lineal

Definición:

Sea $\{(x_i, y_i)\}_{i=1}^n$ una muestra aleatoria bivariada del par de variables (X, Y). se define el coeficiente de correlación lineal como sigue:

$$r_{X,Y} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y}_n)^2}} \tag{5}$$

Este coeficiente mide el grado de relación de tipo lineal que existe entre X e Y.

Interpretación del coeficiente de Correlación

En este punto hay que ser cuidadoso, ya que el coeficiente de correlación no entrega información suficiente, como se vé en los graficos que se encuentran en el archivo Excel correspondiente a esta clase. De todas formas se tiene lo siguiente.

$r_{X,Y} = -1$	Relación estrictamente lineal negativa.
$-1 < r_{X,Y} < 0$	Tendencia lineal con pendiente negativa.
$r_{X,Y} = 0$	Ausencia de tendencia lineal.
$0 < r_{X,Y} < 1$	Tendencia lineal con pendiente negativa.
$r_{X,Y} = 1$	Relación estrictamente lineal positiva.

Cuadro: Interpretación Coeficiente de Correlación

Razón de Correlación

Seguimos en el problema de detectar una relación funcional entre dos variables. Con el coeficiente de correlación encontramos una respuesta para el caso de dos variables cuantitativas. Abordaremos ahora el caso en que una de las variables es cualitativa.

Consideremos dos variables aleatorias X e Y donde la variable X es cualitativa y puede tomar q modalidades o categorías (e.g. Color de Pelo, Raza de un perro, etc.). Supongamos ahora que tenemos n_j observaciones de la variable Y en la categoría j (e.g. Tomo diez perros de la misma raza y mido su peso). Sean $\{y_{ij}\}_{i=1}^{n_j}$ los valores de estas observaciones.

Razón de Correlación

Seguimos en el problema de detectar una relación funcional entre dos variables. Con el coeficiente de correlación encontramos una respuesta para el caso de dos variables cuantitativas. Abordaremos ahora el caso en que una de las variables es cualitativa.
Consideremos dos variables aleatorias X e Y donde la variable X es cualitativa y puede tomar q modalidades o categorías (e.g. Color de Pelo, Raza de un perro, etc.). Supongamos ahora que tenemos n_j observaciones de la variable Y en la categoría j (e.g. Tomo diez perros de la misma raza y mido su peso). Sean $\{y_{ij}\}_{i=1}^{n_j}$ los valores de estas observaciones.

Razón de Correlación

La idea es separar la variabilidad total de la variable Y sobre la muestra en dos partes: una que represente la variabilidad entre categorías y otra que represente la variabilidad dentro de las categorías.

De esta manera si tuvieramos que la variabilidad entre categorias es nula, podríamos concluir que el conocimiento de la categoría para X en que se encuentra una observación de Y no nos sirve para determinar su valor. Es decir, concluiríamos que no hay indicios de una relación funcional entre las variables X e Y.

Si por el contrario encontramos que la variabilidad dentro de cada categoría es nula, podríamos concluir que saber en que categoría de la variable X esta la observación de la variable Y determina el valor de esta.

Razón de Correlación

La idea es separar la variabilidad total de la variable Y sobre la muestra en dos partes: una que represente la variabilidad entre categorías y otra que represente la variabilidad dentro de las categorías.
De esta manera si tuvieramos que la variabilidad entre categorias es nula, podríamos concluir que el conocimiento de la categoría para X en que se encuentra una observación de Y no nos sirve para determinar su valor. Es decir, concluiríamos que no hay indicios de una relación funcional entre las variables X e Y.
Si por el contrario encontramos que la variabilidad dentro de cada categoría es nula, podríamos concluir que saber en que categoría de la variable X esta la observación de la variable Y determina el valor de esta.

Razón de Correlación

La idea es separar la variabilidad total de la variable Y sobre la muestra en dos partes: una que represente la variabilidad entre categorías y otra que represente la variabilidad dentro de las categorías.
De esta manera si tuvieramos que la variabilidad entre categorias es nula, podríamos concluir que el conocimiento de la categoría para X en que se encuentra una observación de Y no nos sirve para determinar su valor. Es decir, concluiríamos que no hay indicios de una relación funcional entre las variables X e Y.
Si por el contrario encontramos que la variabilidad dentro de cada categoría es nula, podríamos concluir que saber en que categoría de la variable X esta la observación de la variable Y determina el valor de esta.

Razón de Correlación

Notemos que si definimos:

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \tag{6}$$

$$w_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \tag{7}$$

$$w^2 = \sum_{j=1}^{q} \frac{n_j}{n} w_j^2 \tag{8}$$

$$b^2 = \sum_{j=1}^{q} \frac{n_j}{n} (\bar{y}_j - \bar{y}_n)^2 \tag{9}$$

Razón de Correlación

Entonces:

$$S_n^2(Y) = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_n)^2 \quad (10)$$

$$S_n^2(Y) = w^2 + b^2 \quad (11)$$

Donde b^2 captura la variabilidad entre los grupos y w^2 captura la variabilidad dentro de los grupos o categorías. Notemos que si $b^2 = 0$ entonces las medias de todas las categorías son iguales a la media de la muestra completa y por lo tanto no hay relación funcional entre X e Y. Si por el contrario $w^2 = 0$ tenemos que $w_j^2 = 0$ para todo $j = 1 \ldots q$ y por lo tanto dentro de cada categoría todas las observaciones son iguales. De lo que concluimos que existe relación funcional entre X e Y.

Razón de Correlación

Entonces:

$$S_n^2(Y) = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_n)^2 \quad (10)$$

$$S_n^2(Y) = w^2 + b^2 \quad (11)$$

Donde b^2 captura la variabilidad entre los grupos y w^2 captura la variabilidad dentro de los grupos o categorías. Notemos que si $b^2 = 0$ entonces las medias de todas las categorías son iguales a la media de la muestra completa y por lo tanto no hay relación funcional entre X e Y. Si por el contrario $w^2 = 0$ tenemos que $w_j^2 = 0$ para todo $j = 1 \ldots q$ y por lo tanto dentro de cada categoría todas las observaciones son iguales. De lo que concluimos que existe relación funcional entre X e Y.

Definición:

Se define la razón de correlación entre X e Y como:

$$\eta^2_{Y|X} = \frac{b^2}{S^2_n(Y)} \tag{12}$$

Se tiene lo siguente:

- Este coeficiente toma valores entre 0 y 1.
- Si es igual a cero no hay relación funcional entre X e Y.
- Si es igual a uno hay relación funcional estricta.

Definición:

Se define la razón de correlación entre X e Y como:

$$\eta^2_{Y|X} = \frac{b^2}{S^2_n(Y)} \tag{12}$$

Se tiene lo siguente:

- Este coeficiente toma valores entre 0 y 1.
- Si es igual a cero no hay relación funcional entre X e Y.
- Si es igual a uno hay relación funcional estricta.

Definición:

Se define la razón de correlación entre X e Y como:

$$\eta^2_{Y|X} = \frac{b^2}{S^2_n(Y)} \tag{12}$$

Se tiene lo siguente:

- Este coeficiente toma valores entre 0 y 1.
- Si es igual a cero no hay relación funcional entre X e Y.
- Si es igual a uno hay relación funcional estricta.

Definición:

Se define la razón de correlación entre X e Y como:

$$\eta^2_{Y|X} = \frac{b^2}{S^2_n(Y)} \tag{12}$$

Se tiene lo siguente:

- Este coeficiente toma valores entre 0 y 1.
- Si es igual a cero no hay relación funcional entre X e Y.
- Si es igual a uno hay relación funcional estricta.

Definición:

Se define la razón de correlación entre X e Y como:

$$\eta^2_{Y|X} = \frac{b^2}{S^2_n(Y)} \tag{12}$$

Se tiene lo siguente:

- Este coeficiente toma valores entre 0 y 1.
- Si es igual a cero no hay relación funcional entre X e Y.
- Si es igual a uno hay relación funcional estricta.

Anova

Aunque estemos en el caso de que exista una relación funcional estricta entre dos variables que estamos estudiando, en una muestra dada es posible que la razón de correlación no nos dé exactamente 1. El test de anova nos da un criterio para decidir cuando la variabilidad que observamos se debe a verdaderas diferencias entre categorías y cuando se debe a la aleatoriedad propia del proceso.

Anova

Debemos tener en cuenta algunos puntos con respecto a este test:

- Se asume que la variable cuantitativa estudiada sigue una distribución normal
- Se asume que dentro de todas las categorías se tiene la misma varianza.
- La hipótesis nula es $\mu_i = \mu_j \forall i, j$.
- El test es robusto frente a pequeñas variaciones en el tamaño de las categorías, aunque para las deducciones se suponen todas del mismo tamaño.

Anova

Debemos tener en cuenta algunos puntos con respecto a este test:

- Se asume que la variable cuantitativa estudiada sigue una distribución normal
- Se asume que dentro de todas las categorías se tiene la misma varianza.
- La hipótesis nula es $\mu_i = \mu_j \forall i, j$.
- El test es robusto frente a pequeñas variaciones en el tamaño de las categorías, aunque para las deducciones se suponen todas del mismo tamaño.

Anova

Debemos tener en cuenta algunos puntos con respecto a este test:

- Se asume que la variable cuantitativa estudiada sigue una distribución normal
- Se asume que dentro de todas las categorías se tiene la misma varianza.
- La hipótesis nula es $\mu_i = \mu_j \forall i, j$.
- El test es robusto frente a pequeñas variaciones en el tamaño de las categorías, aunque para las deducciones se suponen todas del mismo tamaño.

Anova

Debemos tener en cuenta algunos puntos con respecto a este test:

- Se asume que la variable cuantitativa estudiada sigue una distribución normal
- Se asume que dentro de todas las categorías se tiene la misma varianza.
- La hipótesis nula es $\mu_i = \mu_j \forall i, j$.
- El test es robusto frente a pequeñas variaciones en el tamaño de las categorías, aunque para las deducciones se suponen todas del mismo tamaño.

Anova

Debemos tener en cuenta algunos puntos con respecto a este test:

- Se asume que la variable cuantitativa estudiada sigue una distribución normal
- Se asume que dentro de todas las categorías se tiene la misma varianza.
- La hipótesis nula es $\mu_i = \mu_j \forall i, j$.
- El test es robusto frente a pequeñas variaciones en el tamaño de las categorías, aunque para las deducciones se suponen todas del mismo tamaño.