Universidad Maimónides

Estadística Aplicada

Universidad Maimónides
2014

Clase 8 – Análisis de Varianza

Pedro Elosegui

Características de la distribución *F*

- Existe una "familia" de distribuciones *F*.
- Cada miembro de la familia está determinado por dos parámetros: los grados de libertad (*gl*) en el numerador y los grados de libertad en el denominador.
- El valor de *F* no puede ser negativo y es una distribución continua.
- La distribución *F* tiene sesgo positivo.
- Sus valores varían de 0 a ∞ . Conforme $F \rightarrow \infty$ la curva se aproxima al eje *X*.

Prueba para varianzas iguales

- Para prueba de dos colas, el estadístico de prueba está dado por:

$$F = \frac{S_1^2}{S_2^2}$$

- S_1^2 y S_2^2 son las varianzas muestrales para las dos muestras. La hipótesis nula se rechaza si el cálculo del estadístico de prueba es más grande que el valor crítico (de tablas) con nivel de confianza $\alpha / 2$ y grados de libertad para el numerador y el denominador.

EJEMPLO 1

- Colin, agente de bolsa, reportó que la tasa media de retorno en una muestra de 10 acciones de software fue 12.6% con una desviación estándar de 3.9%. La tasa media de retorno en una muestra de 8 acciones de compañías de servicios fue 10.9% con desviación estándar de 3.5%. Para .05 de nivel de significancia, ¿puede Colin concluir que hay mayor variación en las acciones de software?

EJEMPLO 1 *continuación*

- Paso 1: $H_0: \sigma_s \leq \sigma_u \qquad H_1: \sigma_s > \sigma_u$
- Paso 2: H_0 se rechaza si $F > 3.68$,

  $gl = (9, 7), \alpha = .05$
- Paso 3: $F = (3.9)^2 / (3.5)^2 = 1.2416$
- Paso 4: H_0 no se rechaza. No hay evidencia suficiente para asegurar que hay mayor variación en las acciones de software.

Suposiciones de ANOVA

- La distribución *F* también se usa para probar la igualdad de más de dos medias con una técnica llamada análisis de variancia (ANOVA).
- ANOVA requiere las siguientes condiciones:
  - la población que se muestrea tiene una distribución normal
  - las poblaciones tienen desviaciones estándar iguales
  - las muestras se seleccionan al azar y son independientes

Procedimiento de análisis de variancia

- Hipótesis nula: las medias de las poblaciones son iguales.
- Hipótesis alterna: al menos una de las medias es diferente.
- Estadístico de prueba: *F* = (variancia *entre* muestras)/(variacia *dentro* de muestras).
- Regla de decisión: para un nivel de significancia α, la hipótesis nula se rechaza si *F* (calculada) es mayor que *F* (en tablas) con grados de libertad en el numerador y en el denominator.

NOTA

- Si se muestrean k poblaciones, entonces los gl (numerador) = k - 1
- Si hay un total de N puntos en la muestra, entonces los gl (denominador) = N - k
- El estadístico de prueba se calcula con: F = [(SST) /(k - 1)] /[(SSE) /(N - k)].
- SST es la suma de cuadrados de los tratamientos.
- SSE es la suma de cuadrados del error.
- Sea T_c el total de la columna, n_c el número de observaciones en cada columna, y ΣX la suma de todas las observaciones.

Fórmulas

$$SS(total) = \Sigma\left(X^2\right) - \frac{(\Sigma X)^2}{n}$$

$$SST = \Sigma\left(\frac{T_c^2}{n_c}\right) - \frac{(\Sigma X)^2}{n}$$

$$SSE = SS\ (total) - SST$$

EJEMPLO 2

- Los restaurantes Rosenbaum se especializan en comidas para retirados y familias. Su presidenta Katy Polsby acaba de desarrollar un nuevo platillo de pastel de carne. Antes de hacerlo parte del menú normal decidió probarlo en varios de sus restaurantes. Quiere saber si hay diferencia en el número medio de comidas vendidas por día en los restaurantes Sylvania, Perrysburg y Point Place para una muestra de cinco días. Con .05 de nivel de significancia, ¿puede Katy concluir que hay una diferencia en el número medio de comidas de carne vendidas por día en los tres restaurantes?

EJEMPLO 2 *continuación*

	Sylvania	Perrysburg	Point Place	
	13	10	18	
	12	12	16	
	14	13	17	
	12	11	17	
			17	total
T_c	51	46	85	182
n_c	4	4	5	13
	653	534	1447	2634

EJEMPLO 2 *continuación*

- De la tabla, Katy determina SST = 76.25, SSE = 9.75, y el estadístico de prueba:

  F = [76.25 /2] /[9.75 /10] = 39.1026
- Paso 1: H_0: $\mu_1 = \mu_2 = \mu_3$ H_1: no todas las medias son iguales
- Paso 2: H_0 se rechaza si $F > 4.10$
- Paso 3: F = 39.10
- Paso 4: H_0 se rechaza. Existe una diferencia en el número medio de comidas vendidas.

Inferencias acerca de las medias de tratamiento

- Cuando se rechaza la hipótesis nula de que las medias son iguales, quizá sea bueno saber qué medias de tratamiento difieren.
- Uno de los procedimientos más sencillo es el uso de los intervalos de confianza.

Intervalos de confianza para la diferencia entre dos medias

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t\sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- donde *t* se obtiene de la tabla con (*N* - *k*) grados de libertad.
- MSE = [SSE /(*N* - *k*)]

EJEMPLO 3

- Del EJEMPLO 2 desarrolle un intervalo de confianza de 95% para la diferencia en el número medio de comidas de pastel de carne vendidas en Point Place (pob # 1) y Sylvania (pob # 2). ¿Puede Katy concluir que existe diferencia entre los dos restaurantes?

$$(17-12.75)\pm 2.228\sqrt{.975\left(\frac{1}{4}+\frac{1}{5}\right)}$$

$$4.25\pm 1.48 \Rightarrow (2.77,5.73)$$

Dos factores ANOVA

- Para ANOVA de dos factores se prueba si existe una diferencia signifcativa entre el *efecto de tratamiento* y si existe una diference en la *variable de bloqueo*.
- Sea B_r el total de bloque (*r* según las filas)
- SSB representa la suma de los cuadrados de los bloques, donde:

$$SSB = \Sigma\left[\frac{B_r^2}{k}\right] - \frac{(\Sigma X)^2}{n}$$

EJEMPLO 4

- La Bieber Manufacturing Co. opera 24 horas al día, cinco días a la semana. Los trabajadores rotan su turno cada semana. Todd Bieber, el propietario, se interesa en saber si hay una diferencia en el número de unidades producidas cuando los empleados trabajan diferentes turnos. Se seleccionó una muestra de cinco trabajadores y se registró su producción en cada turno. Con .05 de nivel de significancia, ¿se puede concluir que existe una diferencia en la producción media por turno y por empleado?

EJEMPLO 4 *continuación*

Empleado	Producción en el día	Producción en la tarde	Producción en la noche
McCartney	31	25	35
Neary	33	26	33
Schoen	28	24	30
Thompson	30	29	28
Wagner	28	26	27

EJEMPLO 4 *continuación*

- Variable de tratamiento
  - Paso 1: H_0: $\mu_1 = \mu_2 = \mu_3$ H_1: no todas las medias son iguales.
  - Paso 2: H_0 se rechaza si $F > 4.46$, $gl = (2, 8)$.
  - Calcule la varible de suma de cuadrados: SS(total) = 139.73, SST = 62.53, SSB = 33.73, SSE = 43.47. *gl*(bloque) = 4, *gl*(tratamiento) = 2, *gl*(error) = 8.
  - Paso 3: F = [62.53 /2] /[43.47 /8] = 5.75

EJEMPLO 4 *continuación*

- Paso 4: H_0 se rechaza. Existe una diferencia en el número promedio de unidades producidas para los distintos periodos o turnos.

- Variable de bloqueo:
  - Paso 1: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ H_1: no todas las medias son iguales.
  - Paso 2: H_0 se rechaza si $F > 3.84$, $gl = (4,8)$
  - Paso 3: $F = [33.73 / 4] / [43.47 / 8] = 1.55$
  - Paso 4: H_0 no se rechaza ya que no existe una diferencia significativa en el número promedio de unidades producidas para los distintos trabajadores.