Supuestos del ANOVA: Heterocedasticidad

Clase 10

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2020-19

Outline

Supuestos

Desde la clase 1 estamos asumiendo que: $Y_{ij} \sim_{iid} Normal(\mu_i, \sigma^2)$ donde σ^2 es igual para todos los tratamientos.

¿Que sucede si esto NO es verdad? Es decir, que pasa si:

- La distribución de los datos NO es normal, sino que pertenece a otra familia.
- Los datos NO son independientes, sino que estas asociados entre ellos.
- La varianza de los datos σ^2 NO es constante, sino que cambia entre los tratamientos.

En la clase de hoy os enfocaremos en esta clase en el último caso.

Definiciones

Definiciones

- Homocedasticidad: Decimos que existe homocedasticidad si la varianza es homogénea en los tratamientos, es decir si σ^2 es constante.
- Heterocedasticidad: Decimos que existe heterocedasticidad si la varianza NO es homogénea en los tratamientos, es decir si σ^2 es cambiante (i.e $\sigma^2 = \sigma^2_{ij}$).

¿Por qué decimos que la heterocedasticidad es un problema? Para esto veamos en donde aparece σ^2 en nuestros cálculos

Outline

Implicaciones

- Al definir TODAS las sumas de cuadrados, en ningún momento se mencionó a σ^2, luego las sumas de cuadrados siguen manteniendo la interpretación que les dimos.
- Lo mismo aplica para los grados de libertad.
- Decíamos que dado la descomposición $\frac{SST}{\sigma^2} = \frac{SSA}{\sigma^2} + \frac{SSE}{\sigma^2}$ entonces $\chi^2_{N-1} = \chi^2_{a-1} + \chi^2_{N-a}$ bajo H_0 (Cochran). El resultado distribucional sigue siendo verdad, pero ahora la descomposición está dada por:

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}\frac{(Y_{ij}-\bar{Y}_{..})^2}{\sigma_i^2} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\frac{(\bar{Y}_{i.}-\bar{Y}_{..})^2}{\sigma_i^2} + \sum_{i=1}^{a}\sum_{j=1}^{n_i}\frac{(Y_{ij}-\bar{Y}_{i.})^2}{\sigma_i^2}$$

Implicaciones

- El lío ahora es cuando se quisiera construir el estadístico cuya distribución sea conocida. En palabras simples, ¡el estadístico usual $F = \frac{MSA}{MSE}$ ya no tiene la distribución F que habíamos determinado! Para esto necesitaríamos conocer la estructura de varianzas por nivel, cosa que no tenemos.
- De forma análoga, es claro entonces que el mejor estimador de la varianza ya NO es el MSE debido a que no existe una única varianza.

Implicaciones

- Ahora bien, si el MSE no tiene sentido, entonces nuestros contrastes tampoco

$$EP = \frac{\hat{\theta} - \theta_0}{\sqrt{MSE\left(\sum_{i=1}^{a} \frac{c_i^2}{n_i}\right)}} \sim t_{(mgl_E)}$$

$$IC\left(\theta, 1-\alpha\right) = \left(\sum_{i=1}^{a} c_i \bar{Y}_{i.} + K\right) \pm t_{[1-\frac{\alpha}{2}; gl_E]} \sqrt{MSE\left(\sum_{i=1}^{a} \frac{c_i^2}{n_i}\right)}$$

$$q = \frac{|\bar{Y}_{i.} - \bar{Y}_{j.}|}{\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim q_{m,glE}$$

Implicaciones

En resumen, todo el procedimiento de inferencia estadística se ve comprometido. Esto significa que las pruebas estadísticas NO son creíbles y por tanto, nuestras conclusiones tampoco serán acertadas, o al menos no al nivel de confianza que estamos afirmando.

El objetivo de la estadística es incluir adecuadamente la incertidumbre asociada a nuestras conclusiones, si no se mide bien dicha incertidumbre ¡NO SE ESTÁ HACIENDO NADA!

Outline

Identificación

Para identificar la heterocedasticidad, existen estrategias basadas en estadísticas descriptivas, gráficas y pruebas estadísticas.

Descriptivas

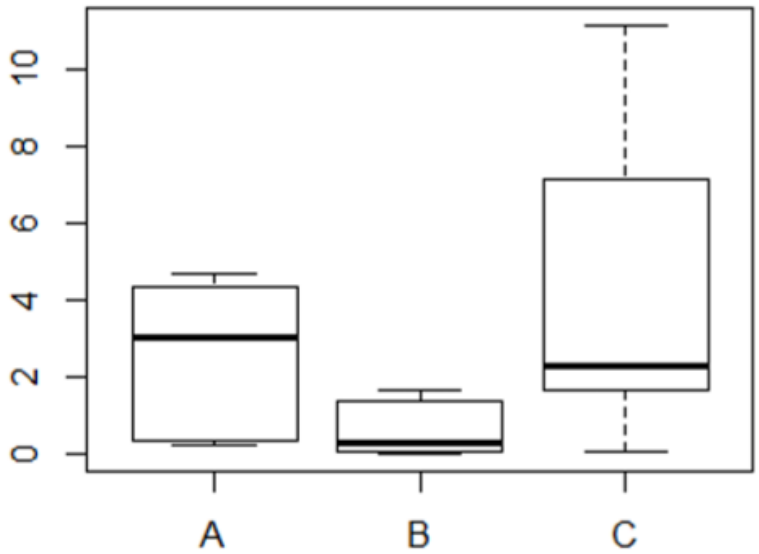
Dado que queremos verificar la homogeneidad de la varianza entre los niveles de un factor, podemos examinar la varianza muestral en cada uno de ellos y verificar si son extremadamente diferentes.

	Diseño 1	**Diseño 2**	**Diseño 3**	**Diseño 4**
Tamaño Muestra	10	6	12	8
Promedio	3.44	3.81	3.95	4.10
Varianza	0.18	0.12	0.06	0.09

Identificación

Gráficas

Si el largo de los Box Plots difiere bastante, hay indicios de heterocedasticidad.



Identificación

Pruebas de Hipótesis

La idea es realizar la prueba

$$H_0: \quad \text{NO hay Heterocedasticidad}$$
$$H_1: \quad \text{SI hay Heterocedasticidad}$$

Para contrastar estas hipótesis existen varios estadísticos, entre ellos el de Levene y Brown-Forsythe.

Identificación

Pruebas de Hipótesis

La idea es realizar la prueba

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2 \\ H_1: &\quad \text{Algún par } \sigma_i^2 \neq \sigma_j^2 \end{aligned}$$

La prueba de hipótesis es bastante parecida a la que realizamos en el ANOVA pero con las varianzas en lugar de las medias.

Identificación

Prueba de Levene

La intuición de la prueba es hacer el mismo análisis de varianza que se realiza para las medias, pero para una medida de dispersión en lugar de una medida de localización.

Para esto podemos usar una transformación que esté relacionada con la varianza, como esta:

$$Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$$

En este caso se tiene que:

$$\mu_i^Z = E(Z_{ij}) \propto \sigma_i$$

Identificación

Prueba de Levene

Por lo tanto la prueba de hipótesis:

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2 \\ H_1: &\quad \text{Algún par } \sigma_i^2 \neq \sigma_j^2 \end{aligned}$$

Es equivalente a la prueba:

$$\begin{aligned} H_0: &\quad \mu_1^Z = \mu_2^Z = \cdots = \mu_a^Z \\ H_1: &\quad \text{Algún par } \mu_i^Z \neq \mu_j^Z \end{aligned}$$

Esta última prueba ya la sabemos probar con el ANOVA

Identificación

Prueba de Brown-Forsythe

La prueba de Brown-Forsythe funciona exactamente igual que la prueba de Levene, la única diferencia es que la transformación de las variables que se utiliza es la siguiente

$$\tilde{Z}_{ij} = |Y_{ij} - \tilde{Y}_{i.}|$$

donde $\tilde{Y}_{i.} =$ mediana $\{Y_{ij}, j \in \{1, \cdots, n_i\}\}$ La mediana es una medida de localización más robusta que la media, luego suele ser preferible utilizarla en comparación a la media simple.

Nuevamente, se tiene que

$$\mu_i^{\tilde{Z}} = E(\tilde{Z}_{ij}) \propto \sigma_i$$

Identificación

Prueba de Brown-Forsythe

Por lo tanto la prueba de hipótesis:

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2 \\ H_1: &\quad \text{Algún par } \sigma_i^2 \neq \sigma_j^2 \end{aligned}$$

Es equivalente a la prueba:

$$\begin{aligned} H_0: &\quad \mu_1^{\tilde{Z}} = \mu_2^{\tilde{Z}} = \cdots = \mu_a^{\tilde{Z}} \\ H_1: &\quad \text{Algún par } \mu_i^{\tilde{Z}} \neq \mu_j^{\tilde{Z}} \end{aligned}$$

Esta última prueba ya la sabemos probar con el ANOVA

Identificación

En resumen, el procedimiento sería:

Calcule las variables: $Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$ ó $\tilde{Z}_{ij} = |Y_{ij} - \tilde{Y}_{i.}|$ y pruebe la hipótesis

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2 \Leftrightarrow \mu_1^Z = \mu_2^Z = \cdots = \mu_a^Z \\ H_1: &\quad \text{Algún par } \sigma_i^2 \neq \sigma_j^2 \Leftrightarrow \text{Algún par } \mu_i^Z \neq \mu_j^Z \end{aligned}$$

por medio de la prueba ANOVA para los datos Z_{ij}, es decir, calcule el estadístico F respectivo

$$F = \frac{MSA_Z}{MSE_Z} \sim F_{a-1,N-a}$$

y finalmente decrete la presencia de heterocedasticidad si:

$$RHO \Leftrightarrow F \geq F_{1-\alpha,a-1,N-a}$$

Outline

Corrección

Una vez identificada la presencia de heterocedasticidad, el siguiente paso es corregirla para que nuestras conclusiones sean acertadas.

Un enfoque consiste en utilizar una transformación que logre homogeneizar la varianza entre los tratamientos. Las transformaciones más populares son las de la familia de Box-Cox, que en general se resumen en:

$$f_1(y) = \sqrt{y},\ f_2(y) = log(y),\ f_3(y) = \frac{1}{y},\ f_4(y)\frac{1}{\sqrt{y}}$$

Si bien existe una teoría que las justifica y explicita cual utilizar, en nuestro caso es suficiente utilizar un enfoque de prueba y error: Si una trasformación no sirve, pruebe otra. Si ninguna sirve, quédese con la que más funcione.

Realmente puede utilizar cualquier transformación que se le ocurra, siempre y cuando sea monotona, y tenga algo de sentido.

Outline

Resumen

En ese orden de ideas, el procedimiento se resume a:

1. Verifique si hay presencia de heterocedasticidad (Prueba de Levene o prueba de Brown-Forsythe).
2. Si hay heterocedasticidad, realice una de las transformaciones mencionadas para sus datos (i.e halle $f(Y_{ij})$).
3. Tome sus datos transformados, y verifique si aún se tiene un problema de heterocedasticidad. Si el problema de heterocedasticidad continua, utilice otra transformación y repita el proceso.
4. Si el problema de heterocedasticidad se soluciona, utilice dichos datos transformados para el resto del análisis del ANOVA y los contrastes

Outline

Enunciado

Un profesor de estadística quiere analizar la diferencia entre cuatro diferentes estrategias para jugar al "Blacjack" (veintiuno). Las estrategias fueron: 1. La del banquero, 2. Cuenta de cinco, 3. Cuenta de diez, 4. Cuenta de más de diez. Se utilizó una calculadora que jugara "blackjack" y con ella se recopiló información de algunas sesiones para cada estrategia. Las ganancia (ó pérdidas) de cada sesión fueron las siguientes:

Banquero	**Cuenta de Cinco**	**Cuenta de Diez**	**Cuenta de más de Diez**
-56	-26	16	60
-78	-12	20	40
-20	18	-14	-16
-46	-8		12
	-16		

¿Se cumple el supuesto de homocedasticidad del ANOVA?