ANOVA con un factor

Estimación de los parámetros del modelo

Jhon Jairo Padilla A., PhD.

Estimación de los parámetros del modelo

Media global y medias de los tratamientos

- Suponga que se poseen los valores de las observaciones pero no se conocen las medias de los tratamientos ni la media global.
- Son estimadores razonables de la media global y de las medias de los tratamientos:

$$\hat{\mu} = \bar{y}_{..}$$

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}, \qquad i = 1, 2, \ldots, a$$

- Es decir:
  - La media global se estima con el gran promedio de las observaciones
  - El efecto de cualquier tratamiento es la diferencia entre el promedio del tratamiento y el gran promedio.

Intervalos de confianza

- Un estimador puntual de μ_i sería $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$
- Suponiendo que los errores siguen una distribución normal, cada $\bar{y}_{i.}$ es una $\mathbf{NID}(\mu_i, \sigma^2/n)$
- Si la varianza fuera conocida, podría usarse la distribución normal para definir el intervalo de confianza.
- Como la varianza se puede calcular con el MS_E, el intervalo de confianza se basaría en la distribución t.

Intervalos de confianza

- Por tanto, un intervalo de confianza de 100(1-α) por ciento para la media μ_i del tratamiento i-ésimo es:

$$\bar{y}_{i.} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \leq \mu_i \leq \bar{y}_{i.} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}}$$

- Y un intervalo de confianza del 100(1-α) por ciento para la diferencia de las medias de dos tratamientos i y j es:

$$\bar{y}_{i.} - \bar{y}_{j.} - t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}}$$

Ejemplo

- Utilizando los datos del ejemplo de la fibra de algodón, pueden encontrarse las estimaciones de la media global y de los efectos de los tratamientos como $\hat{\mu} = 376/25 = 15.04$ y

$$\hat{\tau}_1 = \bar{y}_{1.} - \bar{y}_{..} = 9.80 - 15.04 = -5.24$$
$$\hat{\tau}_2 = \bar{y}_{2.} - \bar{y}_{..} = 15.40 - 15.04 = +0.36$$
$$\hat{\tau}_3 = \bar{y}_{3.} - \bar{y}_{..} = 17.60 - 15.04 = -2.56$$
$$\hat{\tau}_4 = \bar{y}_{4.} - \bar{y}_{..} = 21.60 - 15.04 = +6.56$$
$$\hat{\tau}_5 = \bar{y}_{5.} - \bar{y}_{..} = 10.80 - 15.04 = -4.24$$

Ejemplo (continuación)

- Un intervalo de confianza de 95% para la media del tratamiento 4 (30% de algodón) se calcula entonces como:

$$21.60 - 2.086\sqrt{\frac{8.06}{5}} \le \mu_4 \le 21.60 + 2.086\sqrt{\frac{8.06}{5}}$$

$$21.60 - 2.65 \le \mu_4 \le 21.60 + 2.65$$

- Por lo tanto, el intervalo de confianza de 95% deseado es $18.95 \le \mu_4 \le 24.25.$

Intervalos de confianza simultáneos

- Las expresiones estudiadas para los intervalos de confianza son para intervalos de confianza **uno a la vez** (El nivel de confianza $1-\alpha$ sólo se aplica para una estimación particular).
- En muchos problemas el experimentador tal vez quiera calcular varios intervalos de confianza, uno para cada una de varias medias o diferencias entre medias.

Coeficiente de confianza global

- Suponga que tiene r medias a las que les quiere calcular el intervalo de confianza del 100 (1-α) por ciento.
- La probabilidad de que los r intervalos sean correctos simultáneamente es al menos 1-rα. A la probabilidad rα se le llama **índice de error en el modo del experimento** o **coeficiente de confianza global.**
- Cuando crece el número de intervalos r, el índice de confianza empieza a volverse falto de información (se ensanchan demasiado los intervalos de confianza).
- **Ejemplos**: r=5 y α=0,05 (valor típico del índice de confianza). Entonces 1-rα=0,75. Con r=10 y α=0,05, entonces 1-rα=0,5

Método de Bonferroni

- Un enfoque para asegurarse de que el nivel de confianza simultáneo no sea demasiado pequeño es sustituir $\alpha/2$ con $\alpha/(2r)$ en las ecuaciones de los intervalos de confianza uno a la vez. Esto varía $t_{\alpha/2}$
- Esto permite al experimentador construir un conjunto de r intervalos de confianza simultáneos para las medias de los tratamientos o las diferencias en las medias de los tratamientos , para los que el nivel de confianza global es de al menos $100(1-\alpha)$ por ciento.
- Este método es atinado cuando r no es muy grande. Produce intervalos de confianza razonablemente cortos.

Datos no balanceados

- Algunos experimentos con un solo factor podrían no tener el mismo número de observaciones dentro de cada tratamiento. En este caso, se dice que **el diseño es no balanceado.**
- Sigue pudiéndose aplicar el análisis de varianza descrito antes, pero deben hacerse ligeras modificaciones en las fórmulas de las sumas de cuadrados.

Diseño balanceado

- Sea que se hagan n_i observaciones bajo el tratamiento *i (i=1, 2, …, a)*
- Y sea $N = \sum_{i=1}^{a} n_i$
- Las fórmulas para los cuadrados serán

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SS_{\text{Tratamientos}} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

- No se requieren más cambios en el análisis de varianza

Recomendación

- Hay dos ventajas al elegir un diseño balanceado:
  - El estadístico de prueba es relativamente insensible a las desviaciones pequeñas del supuesto de la igualdad de las varianzas de los *a* tratamientos cuando los tamaños de las muestras son iguales. (No sucede así con tamaños de muestras diferentes)
  - La potencia de la prueba se maximiza cuando las muestras tienen el mismo tamaño.

Verificación de la adecuación del modelo

- Recordemos que para poder usar ANOVA se requieren ciertos supuestos:
  - Que el modelo $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ describe bien las observaciones
  - Que los errores siguen una distribución normal e independiente con media cero y varianza σ^2 constante pero desconocida
- En la práctica, es común que estos supuestos no se satisfagan exactamente.
- En general, no es prudente confiar en el análisis de varianza hasta haber verificado estos supuestos.

El examen de los residuales

- Las violaciones a los supuestos básicos y la adecuación del modelo pueden investigarse con facilidad mediante el **examen de los residuales**.
- El residual de la observación j-ésima en el tratamiento i-ésimo se define como: $e_{ij} = y_{ij} - \hat{y}_{ij}$
- Donde $\hat{y}_{ij}$ es una estimación de la observación y_{ij} correspondiente.
- La estimación de la observación se obtiene de la siguiente manera:

$$\begin{aligned} \hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i \\ &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) \\ &= \bar{y}_{i.} \end{aligned}$$

- Es decir, la estimación de la observación coincide con la estimación más intuitiva que consiste en calcular la media de el tratamiento i-ésimo.

El examen de los residuales

- Debe ser una parte automática de cualquier análisis de varianza
- Si el modelo es adecuado, los residuales deberán estar sin estructura (no deberán contener patrones obvios)
- Pueden descubrirse muchos tipos de inadecuaciones del modelo y violaciones a los supuestos subyacentes.
- Ahora estudiaremos cómo verificar el modelo mediante el análisis gráfico de los residuales y cómo resolver varias anormalidades que ocurren comúnmente.

Verificación del histograma de los residuales

- La verificación de la normalidad se puede realizar graficando un histograma de los residuales.
- Si se satisface el supuesto de media cero y varianza σ^2 para los errores, la gráfica deberá tener una distribución normal con media cero.
- El problema es que cuando se trabaja con muestras pequeñas, las fluctuaciones suelen ser significativas:
  - La aparición de una desviación moderada de la normalidad no implica necesariamente una violación seria de los supuestos.
  - Las desviaciones marcadas de la normalidad son potencialmente serias y requieren un análisis adicional.

Histograma de los residuales

- Se suele utilizar una gráfica de probabilidad normal de los residuales (ya habíamos utilizado antes esta gráfica para la prueba t)
- En el Análisis de Varianza es más útil hacer esto con los residuales.
- Si la distribución fundamental de los errores es normal, esta gráfica tendrá la apariencia de una línea recta.
- Para visualizar la línea recta, deberá prestarse más atención a los valores centrales de la gráfica que a los valores extremos.

Verificación de la normalidad: Ejemplo

Tabla 3-6 Datos y residuales del ejemplo 3-1[a]

Peso porcentual del algodón	Observaciones (j)										$y_{ij} = \bar{y}_i$
	1		2		3		4		5		
15	7	−2.8 (15)	7	−2.8 (19)	15	5.2 (25)	11	1.2 (12)	9	−0.8 (6)	9.8
20	12	−3.4 (8)	17	1.6 (14)	12	−3.4 (1)	18	2.6 (11)	19	2.6 (3)	15.4
25	14	−3.6 (18)	18	0.4 (13)	18	0.4 (20)	19	1.4 (7)	19	1.4 (9)	17.6
30	19	−2.6 (22)	25	3.4 (5)	22	0.4 (2)	19	−2.6 (24)	23	1.4 (10)	21.6
35	7	−3.8 (17)	10	−0.8 (21)	11	0.2 (4)	15	4.2 (16)	11	0.2 (23)	10.8

[a]Los residuales se indican en el recuadro de cada celda. Los números entre paréntesis indican el orden en que se recolectaron los datos.

Verificación de la normalidad: Ejemplo

- La distribución de los errores tiene un ligero sesgo: cola derecha más larga que la cola izquierda
- Tendencia a curvarse hacia abajo del lado izquierdo: La cola izquierda de los errores es más delgada de lo que debería ser en una distribución normal. Los residuales negativos no son tan grandes como se esperaba.
- A pesar de ello, la gráfica no muestra una desviación marcada de la distribución normal.

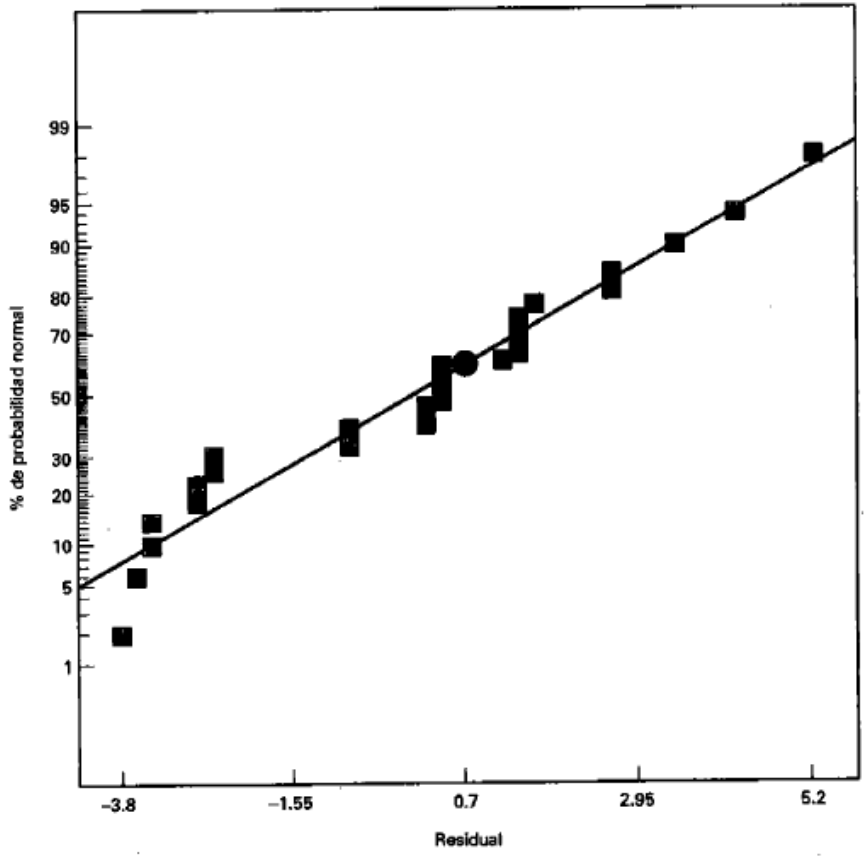

Figura 3-4 Gráfica de probabilidad normal de los residuales del ejemplo 3-1.

Verificación de la normalidad

- En general, las desviaciones moderadas de la normalidad no son motivo de preocupación en el análisis de varianza de efectos fijos
- Es más preocupante una distribución de los errores con colas considerablemente más gruesas o más delgadas que la distribución normal. Una distribución sesgada (a izquierda o derecha) no representa tanto problema.
- La prueba F sólo se afecta ligeramente por estas desviaciones, por lo que se dicen que el análisis de varianza es **robusto** con respecto al supuesto de normalidad.

Verificación de la normalidad

- Las desviaciones de la normalidad hacen que tanto el verdadero nivel de significación como la verdadera potencia (α y β) difieran ligeramente de los valores anunciados
- La potencia generalmente es más baja.
- El modelo de efectos aleatorios sí se afecta en forma más severa por la no normalidad.

Puntos atípicos en los gráficos de probabilidad normal

- Una anomalía muy común en estas gráficas es un residual que es mucho más grande que cualquier otro (llamado **Punto atípico**)
- La presencia de uno o más puntos atípicos suele introducir serias distorsiones en el análisis de varianza.
- Se requiere entonces una investigación atenta en estos casos.
- Posibles causas:
  - Error en los cálculos
  - Error al codificar o copiar los datos
- Si estas no son las causas, deben estudiarse las circunstancias experimentales que rodean esta corrida particular.

Puntos atípicos en los gráficos de probabilidad normal

- Si las circunstancias de los puntos atípicos muestran condiciones especialmente deseables (alta resistencia, bajo costo, etc), el punto atípico puede ser más informativo que el resto de los datos.
- Las observaciones atípicas no deben rechazarse o descartarse a menos que se tengan razones no estadísticas de peso para hacerlo.
- En el peor de los casos puede llegarse a tener dos análisis, uno don el punto atípico y otro sin él.

Procedimientos para detectar puntos atípicos

- Existen varios procedimientos estadísticos formales para detectar puntos atípicos.
- Una forma es examinar los residuales estandarizados, que se calculan como:

$$d_{ij} = \frac{e_{ij}}{\sqrt{MS_E}}$$

- Si los errores tienen media cero, los residuales estandarizados deberán ser aproximadamente normales con media cero y varianza unitaria.
- En consecuencia se cumplirá que: el 68% de los residuales estandarizados deberán estar incluidos dentro de los límites +/- 1, cerca del 95% de ellos deberán estar incluidos dentro del +/- 2 y, virtualmente, todos ellos deberán estar incluidos dentro de +/- 3. Un residual mayor que 3 o 4 desviaciones estándar a partir de cero es un punto atípico potencial.

Procedimientos para detectar puntos atípicos

- Para el ejemplo de la resistencia a la tensión, la gráfica obtenida de probabilidad normal no produce indicio alguno de puntos atípicos.
- El residual estandarizado mayor es:

$$d_{13} = \frac{e_{13}}{\sqrt{MS_E}} = \frac{5.2}{\sqrt{8.06}} = \frac{5.2}{2.84} = 1.83$$

- Por lo que no es motivo de preocupación.

Secuencia temporal de los residuales

- La grafica de los residuales en el orden temporal de la recolección de datos es útil para detectar correlaciones entre los residuales.
- Una tendencia a tener corridas de residuales positivos o negativos indica una correlación positiva. Esto implicaría que el supuesto de independencia de los errores ha sido violado.
- Esto es un problema serio y de difícil solución. Es importante evitar este problema cuando se recolecten los datos.
- Esto se logra con la correcta aleatorización del experimento.

Secuencia temporal de los residuales

- Otro error detectable es que en ocasiones las habilidades del experimentador (o los sujetos) pueden cambiar conforme el experimento avanza, o el proceso bajo estudio puede vagar sin rumbo o volverse más errático.
- Esto produce un cambio en la varianza del error con el tiempo. Esta condición se observa en la gráfica como una dispersión mayor en uno de los extremos que en el otro.
- Una varianza no constante es un problema serio para poder aplicar Anova.

Secuencia temporal de los residuales

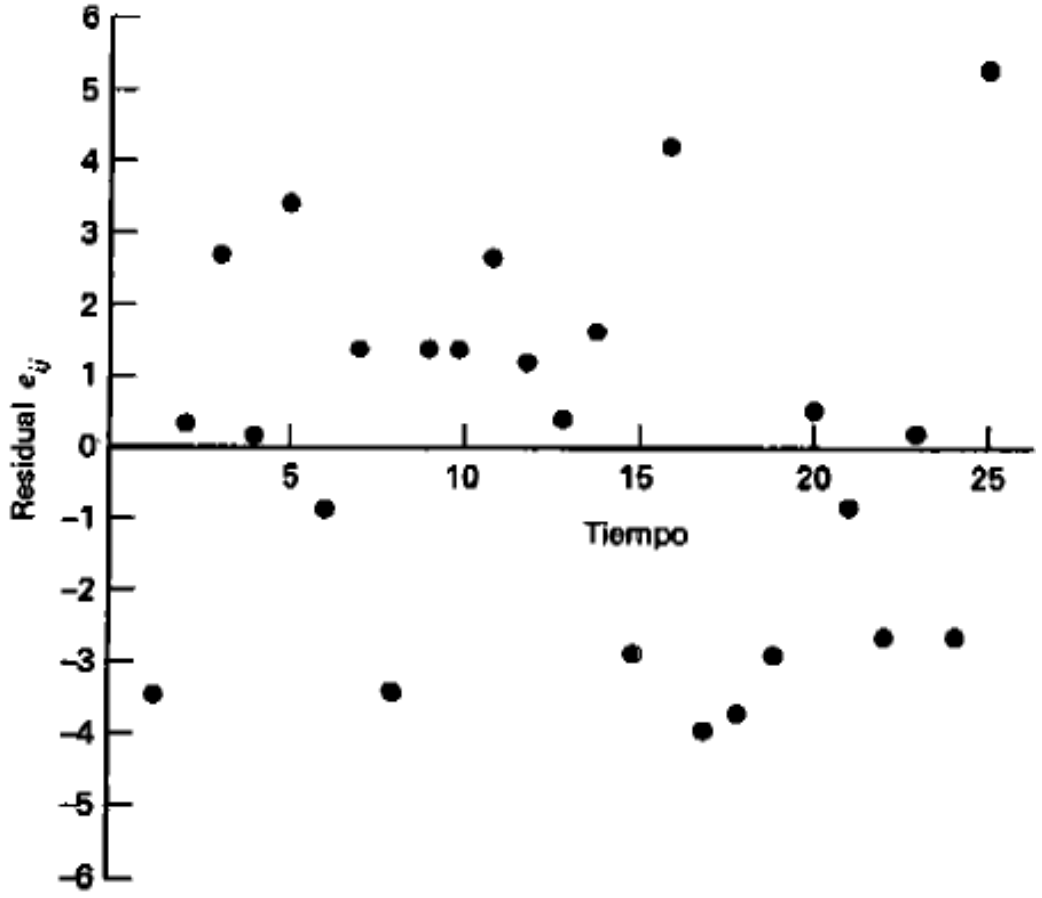

Figura 3-5 Gráfica de los residuales contra el tiempo.

- En el gráfico se aprecia que no hay correlación entre los errores
- Tampoco se presenta una varianza inconstante.

Gráfica de los residuales contra los valores ajustados

- Si el modelo es correcto y se satisfacen los supuestos, los residuales deberán aparecer sin estructura
- Los residuales no deberán estar relacionados con ninguna otra variable, incluyendo la respuesta predicha.
- Una verificación simple es graficar los residuales contra los valores ajustados (estimados) $\hat{y}_{ij}$
- Esta gráfica no deberá mostrar ningún patrón obvio.

Gráfica de los residuales contra los valores ajustados: Ejemplo

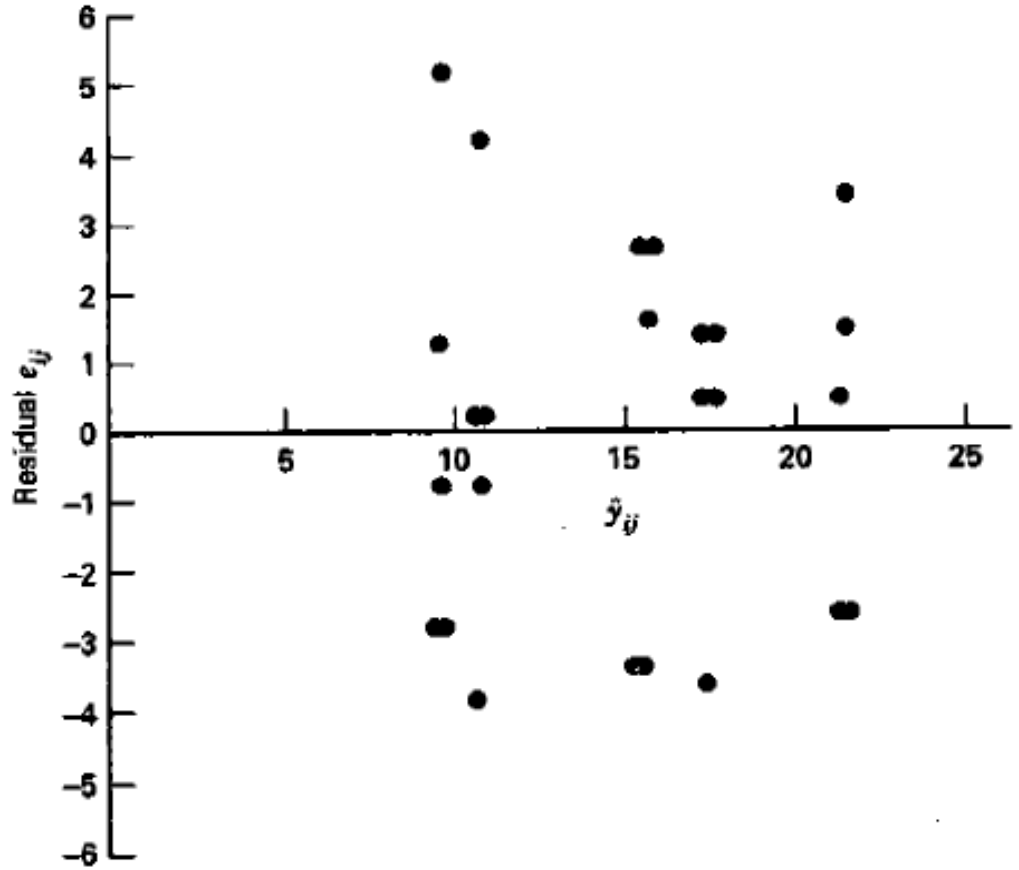

Figura 3-6 Gráfica de los residuales contra los valores ajustados.

Problemas en las varianzas

- Un defecto que se puede detectar con este gráfico es el de la varianza no constante.
- En ocasiones la varianza de las observaciones se incrementa cuando la magnitud de la observación se incrementa.
- Una causa de esto puede ser que el error o ruido de fondo fuera un porcentaje constante de la magnitud de la observación (suele ocurrir con muchos instrumentos de medición: el error es un porcentaje de la escala de medición)
- La gráfica de los residuales contra y_{ij} se vería como un embudo o un megáfono viendo hacia la izquierda

Problemas en las varianzas

- Una varianza no constante también surge en los casos en que los datos siguen una distribución no normal, sesgada (en las distribuciones sesgadas la varianza tiende a ser una función de la media)
- Cuando se viola el supuesto de homogeneidad de las varianzas, la prueba F:
  - Sólo resulta ligeramente afectada en el modelo balanceado
  - En diseños no balanceados o en casos en que una de las varianzas es considerablemente más grande que las demás, el problema es más grave.

Problemas en las varianzas

- Si los niveles de factor que tienen varianzas mayores corresponden también con los tamaños de las muestras más pequeños, el índice de error tipo I real es mayor que lo previsto (los intervalos de confianza tiene niveles de confianza reales más bajos de los esperados)
- Si los niveles del factor con varianzas mayores tienen también los tamaños de las muestras mayores, los niveles de significación son mucho menores que lo anticipado (los niveles de confianza son más altos)
- **Por estas razones es preferible escoger tamaños de las muestras iguales siempre que sea posible.**

Solución para usar ANOVA en los casos vistos

- Para poder usar ANOVA con datos que presentan los problemas de varianza mencionados, se requiere utilizar una **Transformación** con el fin de **estabilizar la varianza**.
- ANOVA se correrá con los datos transformados.
- Por tanto, las conclusiones de ANOVA sólo se aplican a las poblaciones transformadas!!!!

Transformaciones para estabilizar la varianza

- Se han dedicado considerables esfuerzos de investigación para la selección de una transformación adecuada
- Si se conoce la distribución teórica de las observaciones, se pueden usar algunos casos estudiados para una transformación adecuada:

Distribución de las observaciones	Transformación a la observación y_{ij}
Poisson	Raíz cuadrada
Lognormal	logarítmo
Binomial (datos como fracciones)	Arcsen de la raíz cuadrada

- Las transformaciones que se hacen a las observaciones afectan también la forma de la distribución del error. En la mayoría de los casos hace que la distribución del error se acerque a la distribución normal.

Transformaciones empíricas

- Sea E(y)=μ la media de y, y suponga que la desviación estándar de y es proporcional a una potencia media de *y* tal que

$$\sigma_y \propto \mu^{\alpha}$$

- Quiere encontrarse una transformación de *y* que produzca una varianza constante
- Suponga que la transformación es una potencia de los datos originales, por ejemplo: $y^* = y^{\lambda}$
- Puede demostrarse entonces que $\sigma_{y^*} \propto \mu^{\lambda+\alpha-1}$
- Si se hace $\lambda = 1 - \alpha$, la varianza de los datos transformados es constante.

Transformaciones para estabilizar la varianza

Tabla 3-9 Transformaciones para estabilizar la varianza

Relación entre σ_y y μ	α	$\lambda = 1 - \alpha$	Transformación	Comentario
$\sigma_y \propto$ constante	0	1	Sin transformación	
$\sigma_y \propto \mu^{1/2}$	1/2	1/2	Raíz cuadrada	Datos (números) de Poisson
$\sigma_y \propto \mu$	1	0	Log	
$\sigma_y \propto \mu^{3/2}$	3/2	–1/2	Raíz cuadrada recíproca	
$\sigma_y \propto \mu^2$	2	–1	Recíproco	

Ejemplo

EJEMPLO 3-5

Un ingeniero civil está interesado en determinar si cuatro métodos diferentes para estimar la frecuencia de las inundaciones producen estimaciones equivalentes de la descarga pico cuando se aplican a la misma cuenca. Cada procedimiento se usa seis veces en la cuenca, y los datos de las descargas resultantes (en pies cúbicos por segundo) se muestran en la parte superior de la tabla 3-7. El análisis de varianza de los datos, el cual se resume en la tabla 3-8, implica que hay una diferencia en las estimaciones de la descarga pico promedio obtenidas en los cuatro procedimientos. La gráfica de los residuales contra los valores ajustados, la cual se muestra en la figura 3-7, es preocupante porque la forma de embudo con la boca hacia afuera indica que no se satisface el supuesto de una varianza constante.

Observaciones del ejemplo

Tabla 3-7 Datos de la descarga pico

Método de estimación	Observaciones						$\bar{y}_{i.}$	$\tilde{y}_i$	S_i
1	0.34	0.12	1.23	0.70	1.75	0.12	0.71	0.520	0.66
2	0.91	2.94	2.14	2.36	2.86	4.55	2.63	2.610	1.09
3	6.31	8.37	9.75	6.09	9.82	7.24	7.93	7.805	1.66
4	17.15	11.82	10.95	17.20	14.35	16.82	14.72	15.59	2.77

Método de estimación	Desviaciones d_{ij} para la prueba de Levene modificada					
1	0.18	0.40	0.71	0.18	1.23	0.40
2	1.70	0.33	0.47	0.25	0.25	1.94
3	1.495	0.565	1.945	1.715	2.015	0.565
4	1.56	3.77	4.64	1.61	1.24	1.23

Análisis de varianza para el ejemplo

Tabla 3-8 Análisis de varianza de los datos de la descarga pico

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor P
Métodos	708.3471	3	236.1157	76.07	<0.001
Error	62.0811	20	3.1041		
Total	770.4282	23			

Gráfica de los residuales contra los valores estimados

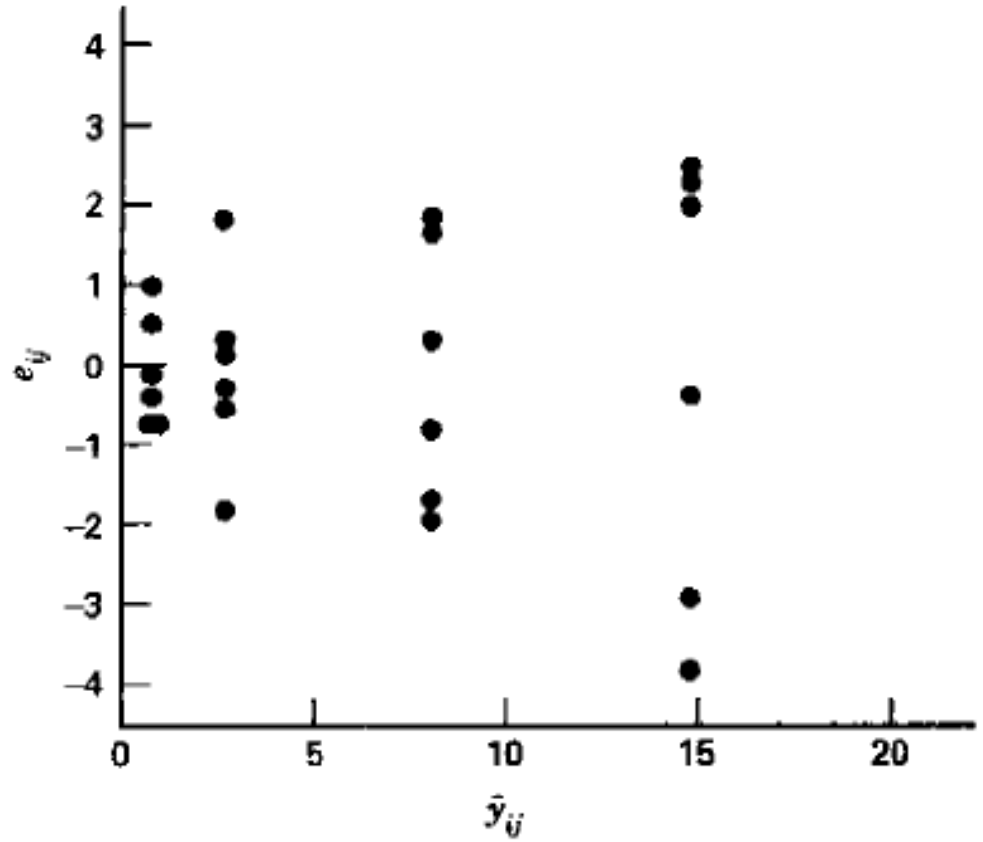

Figura 3-7 Gráfica de los residuales contra $\hat{y}_{ij}$ para el ejemplo 3-5.

Gráfica de varianza contra log de las observaciones

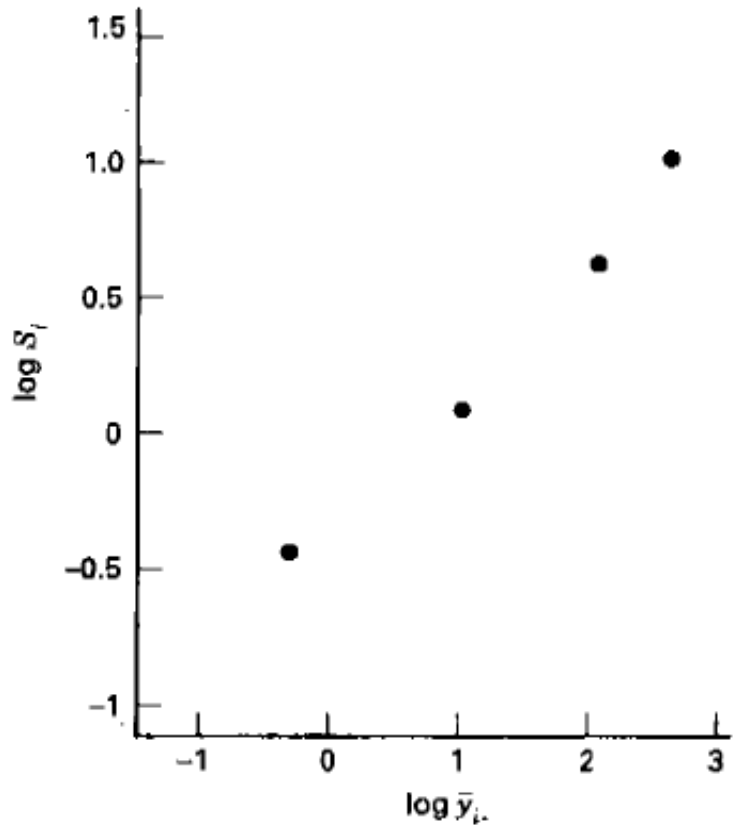

Figura 3-8 Gráfica de $\log S_i$ contra $\log \bar{y}_{i\cdot}$ para los datos de la descarga pico del ejemplo 3-5.

Análisis de varianza para los datos transformados

Tabla 3-10 Análisis de varianza de los datos transformados de la descarga pico, $y^* = \sqrt{y}$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor P
Métodos	32.6842	3	10.8947	76.99	<0.001
Error	2.6884	19	0.1415		
Total	35.3726	22			

Gráfico de residuales contra los valores estimados con los valores transformados

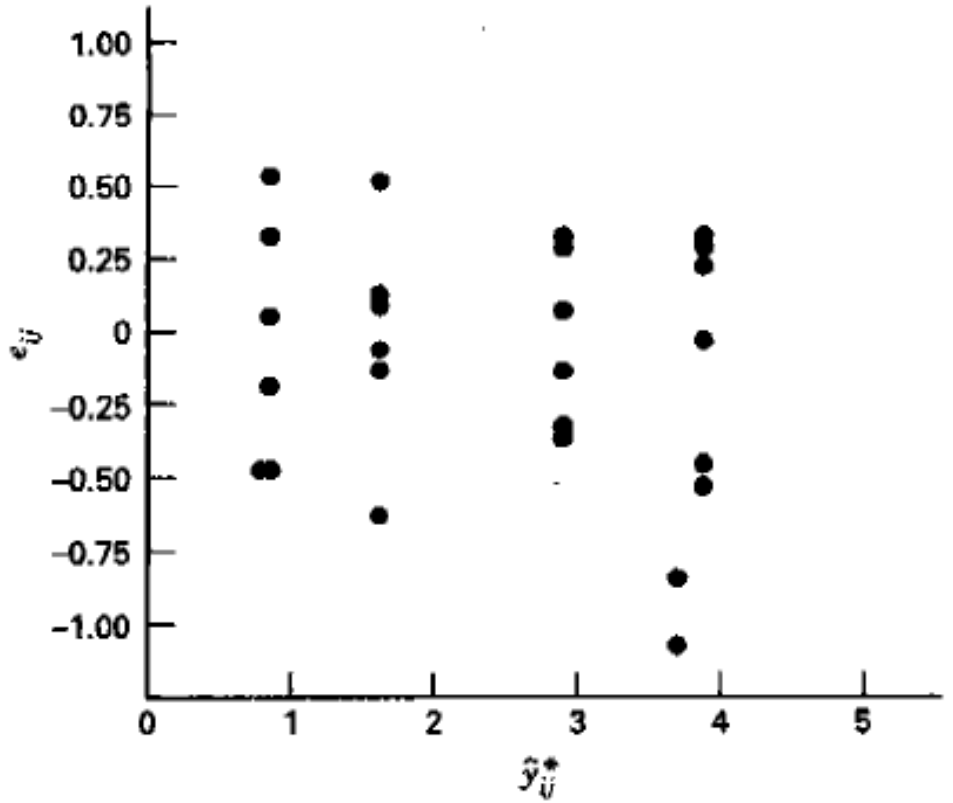

Figura 3-9 Gráfica de los residuales de los datos transformados contra $\hat{y}_{ij}^{*}$ para los datos de la descarga pico del ejemplo 3-5.

Concluyendo…

- En la práctica, muchos experimentadores seleccionan la forma de la transformación probando varias alternativas y observando el efecto de cada transformación en la gráfica de los residuales contra la respuesta predicha. Entonces se selecciona la transformación que produjo la gráfica residual más satisfactoria.