Notebook sobre supuestos de Normalidad

Introducción

Durante el diseño de un experimento es importante contar con el tamaño apropiado de la muestra, éste es uno de los aspectos más importantes en esta etapa. Lo anterior es debido a que la elección del tamaño de la muestra y la probabilidad β del **Error tipo** II guardan una estrecha relación.

Teniendo en cuenta la información indicada en **Diseño y Análisis de Experimentos - Montgomery pp 55** Suponga que se está probando la hipótesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Y que las medias no son iguales, por lo que $\delta = \mu_1 - \mu_2$. Puesto que $H_0 : \mu_1 = \mu_2$ no es verdadera, la preocupación principal es cometer la equivocación de no rechazar Ha. La probabilidad del **error tipo II** depende de la verdadera diferencia en las medias δ. A una gráfica de β contra δ para un tamaño particular de la muestra se le llama **la curva de operación característica**, o curva **OC**, de la prueba. El error β también es una función del tamaño de la muestra. En general, para un valor dado de δ, el error β se reduce cuando el tamaño de la muestra se incrementa. Es decir, es más fácil detectar una diferencia especificada en las medias para tamaños grandes de la muestra que para los tamaños pequeños.

Como conclusión de un análisis sobre **OCs** se determina:

1. Entre más grande sea la diferencia en las medias, $\mu_1 - \mu_2$, menor será la probabilidad del error tipo II para un tamaño de la muestra y un valor de α dados. Es decir, para un tamaño de la muestra y un valor de α especificados, la prueba detectará con mayor facilidad las diferencias grandes que las pequeñas.
2. Cuando el tamaño de la muestra se hace más grande, la probabilidad del error tipo II se hace más pequeña para una diferencia en las medias y un valor de α dados. Es decir, para detectar una diferencia δ espcificada, se puede aumentar la potencia de la prueba incrementando el tamaño de la muestra

Una vez se cuenta con un tamaño de muestra adecuado, es necesario verificar la distribución de los datos, particularmente que la métrica de calidad se distribuya como una $N(\mu, \sigma^2)$, lo anterior es debido a que La validez de los resultados obtenidos en cualquier análisis de varianza queda supeditado a que los supuestos del modelo se cumplan. Los supuestos en su orden son: *1. Normalidad, 2. Homocedasticidad (Varianza constante), 3. Independencia. * No obstante, para evaluar dichos supuestos es necesario determinar previamente qué tan adecuado es el modelo.

Adecuación del modelo

El análisis de descomposición de la varianza (ANOVA) para un único factor (ANOVA Monofactorial) se plantea a partir de la relación algebraica entre la media de cada nivel y cada observación, lo anterior, mediante la siguiente relación:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Donde y_{ij} corresponde a cada uno de las observaciones u experimentos (corridas) realizado (donde i es el nivel de cada factor, y j corresponde con la réplica realizada); μ es la tendencia general de los datos o media global; τ_i el ajuste o efecto de pertenecer a cada uno de los niveles en el factor bajo estudio; y ε_{ij} indica el error o la variabilidad que no puede ser expresada mediante los componentes de la ecuación, estos errores se caracterizan por seguir uns distribución Normal con media cero y varianza desconocida pero constante $N(0, \sigma^2)$. Es importante resaltar que durante la ejecución de experimentos se suele trabajar con muestras pequeñas y, en consecuencia, es posible que se presenten fluctuaciones significativas de los valores; sin embargo, si la desviación de la distribución Normal es moderada no implica necesariamente una violación seria de los supuestos. Ahora bien, revisar las desviaciones del supuesto de normalidad se puede hacer mediante el análisis de la distribución de los residuales, éstos son definidos mediante la siguiente ecuación:

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

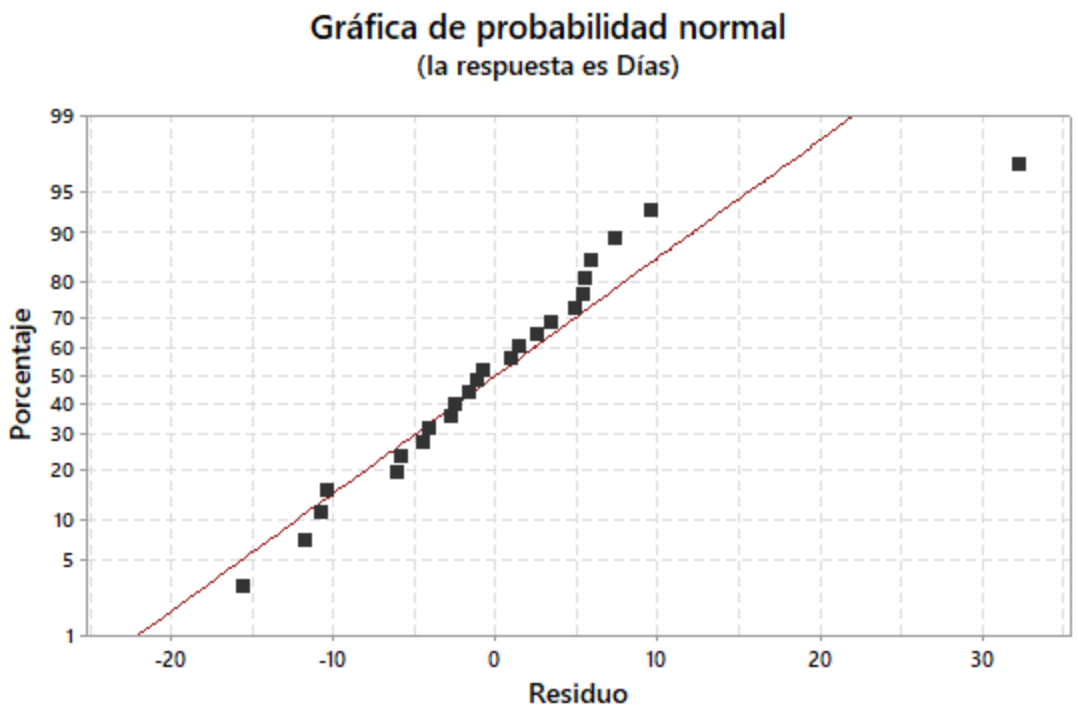
Es decir, el valor residual e_{ij} corresponde con la diferencia entre cada observación y_{ij} y su valor estimado esperado $\hat{y}_{ij}$, donde este último se puede hallar de la siguiente manera:

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i$$
$$\hat{y}_{ij} = \overline{y}_{ij} + (\overline{y}_{i.} - \overline{y}_{ij})$$
$$\hat{y}_{ij} = \overline{y}_{i.}$$

Teniendo en cuenta lo anterior (es decir, el valor esperado de cada pronóstico), se determina que la estimación de cualquier observación en el *iésimo* tratamiento corresponde con el promedio del tratamiento ($\overline{y}_{i.}$). Si bien para comprobar cada supuesto existen pruebas analíticas y gráficas, a continuación se aplicarán pruebas gráficas que si bien se prestan más a la interpretación, en casos de experimentos con residuales alejados de una distribución Normal brindan la suficiente información para concluir los supuestos.

Supuesto de Normalidad

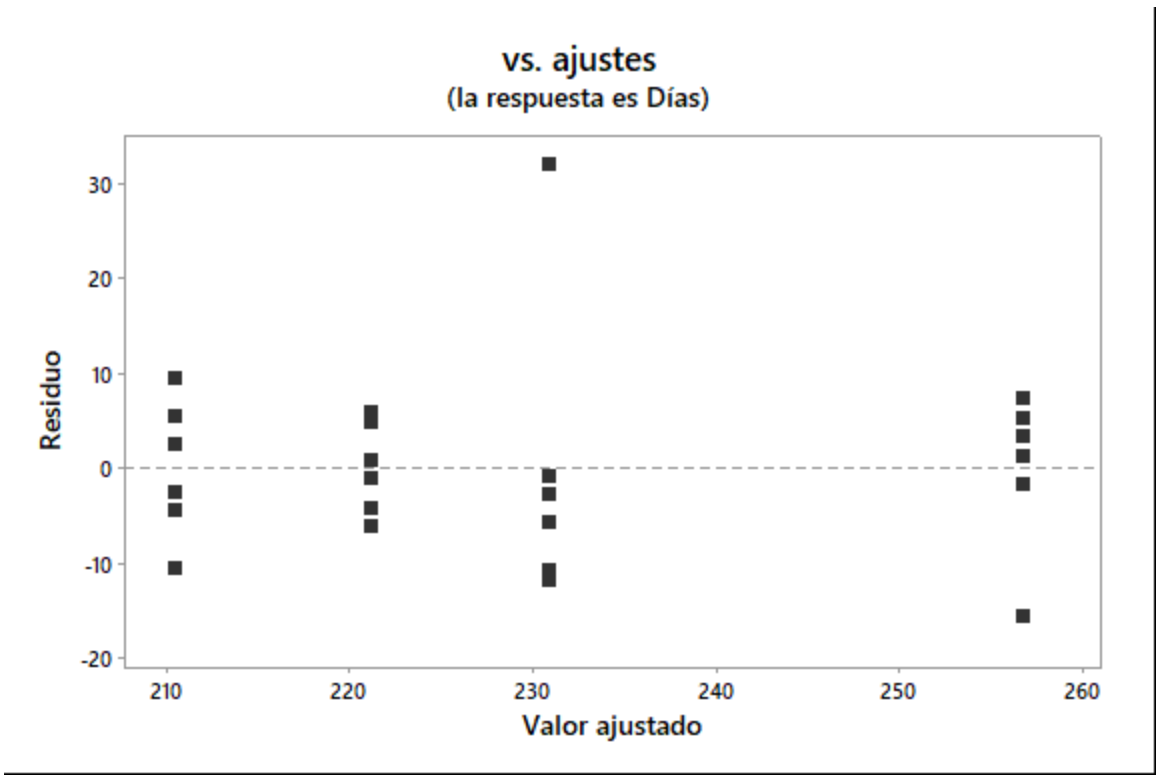
En el primer supuesto se busca una aproximación gráfica a la siguiente afirmación: los e_{ij} siguen una distribución normal con media cero. Para ello se reliza una gráfica de dispersión donde existen dos variables de interés. Para el caso de trabajar en una escala lineal (donde hay equidistancia en ambos ejes) en el eje de las abscisas (x) se registran de manera ascendente los valores de los residuales r_i, y en el eje de las ordenadas (y) la función Normal estándar acumulada evaluada en Z_i, es decir, cada punto en la gráfica corresponde con la ubicación (r_i, Z_i). A modo de ejemplo, a continuación, se muestra una gráfica que contiene los elementos anteriormente descritos.



Como recomendación, si gusta hay un vídeo donde se explica brevemente los pasos para realizar la gráfica Q-plot y se encuentra en el siguiente enlace: Gráfica de Normalidad (https://www.powtoon.com/c/cCrNK5TS7Dz/2/m)

Supuesto de Homocedasticidad

En el segundo supuesto se busca una aproximación gráfica a la siguiente afirmación: los residuos de cada tratamiento tienen la misma varianza σ^2. Para ello una manera de realizar la aproximación es graficando los residuales en función del grupo o nivel al cual pertenecen los experimentos (i, e_{ij}). A continuación, se muestra un ejemplo de la gráfica; es importante analizar si a medida que cambia el nivel bajo estudio los datos se encuentran igualmente dispersos (homocedasticidad) u diferentemente dispersos (heterocedasticidad), en caso de darse esta última condición, se rechaza el supuesto de varianza constante.



Supuesto de Independencia

En el tercer supuesto se busca una aproximación gráfica a la siguiente afirmación: los e_{ij} son independientes entre sí. El supuesto de independencia en los residuos puede verificarse de manera gráfica al contrastar el orden en que se colectó un dato (i, j) (o se hizo el experimento) en el eje de las abscisas, contra el residuo correspondiente $e_{i,j}$ en el eje de las ordenadas. A continuación, se muestra un ejemplo de la gráfica descrita, es necesario verificar que no se encuentren patrones aparentes: