Diseño de experimentos: ANOVA

Elisa Mª Molanes López

Un ejemplo introductorio

- Un ingeniero de desarrollo de productos desea maximizar la resistencia a la tensión de una nueva fibra sintética que se utilizará para fabricar camisas.

- Por experiencia, parece que la resistencia (o fortaleza) se ve influida por el % de algodón presente en la fibra.

- También se sospecha que valores elevados de % de algodón repercuten negativamente en otras cualidades de calidad que se desean (por ej. que la fibra pueda recibir un tratamiento de planchado permanente).

- Ante esta situación, el ingeniero decide tomar cinco muestras para diferentes niveles de % de algodón y medir la fortaleza de las fibras así producidas.

Un ejemplo introductorio

Estos datos suman **49** y su media es **9.8**

Media global de las 25 observaciones

Lo que obtiene se puede resumir en la siguiente tabla:

% de algodón	Observaciones (fortaleza de las 25 fibras fabricadas)					Total	Promedio
15%	7	7	15	11	9	49	9.8
20%	12	17	12	18	18	77	15.4
25%	14	18	18	19	19	88	17.6
30%	19	25	22	19	23	108	21.6
35%	7	10	11	15	11	54	10.8
Suma total de los 25 valores de fortaleza obtenidos						376	15.04

Un ejemplo introductorio

- A la hora de fabricar las 25 fibras anteriores se debe seguir una secuencia aleatorizada.
- Esta aleatorización en la secuencia de fabricación es necesaria para evitar que los datos observados (la fortaleza de los tejidos), sean contaminados por el efecto de otras variables que no conocemos y por tanto no podemos controlar.
- Supongamos que se fabrican las 25 fibras sin un mecanismo aleatorizado, es decir, siguiento el orden original (primero se fabrican las 5 fibras con un 15 % de algodón, luego las 5 fibras con un 20% de algodón, y así sucesivamente).
- En esta situación, si la máquina que mide la fortaleza de la fibra presentase un efecto de calentamiento de modo que a mayor tiempo de funcionamiento diese menores lecturas de resistencia, entonces los datos se contaminarían. Por ese efecto de calentamiento, la fortaleza de las fibras fabricadas con un 35% de algodón resultarían negativamente muy contaminadas. No pasaría lo mismo con las fabricadas con un 15% de algodón.
- Si aleatorizamos la fabricación de las 25 fibras, se espera que este efecto esté presente por igual en todos los % de algodón, de modo que las comparaciones entre los distintos niveles siguen siendo válidos.

Un ejemplo introductorio

El análisis de la varianza nos ayudará a responder las siguientes cuestiones:

- ¿Influye el % de algodón en la fortaleza de la fibra fabricada?

- Si es así, ¿qué niveles de % de algodón son similares y cuáles no?

Analysis Of Variance (ANOVA)

En general, tendremos:

Factor	Observaciones (variable dependiente de interés, y)					Total	Promedio
Nivel o grupo 1	y_{11}	y_{12}	…	…	y_{1n1}	$y_{1\bullet}$	$\bar{y}_{1\bullet}$
Nivel 2	y_{21}	y_{22}	…	…	y_{2n2}	$y_{2\bullet}$	$\bar{y}_{2\bullet}$
Nivel I-ésimo	y_{I1}	y_{I2}	…	…	y_{InI}	$y_{I\bullet}$	$\bar{y}_{I\bullet}$
						$y_{\bullet\bullet}$	$\bar{y}_{\bullet\bullet}$

Notación

y_{ij} se refiere a la observación *j*-ésima de la variable *y* (fortaleza) en el grupo *i*-ésimo del factor (% de algodón).

$$y_{i\bullet} = \sum_{j=1}^{n_i} y_{ij}$$

El punto significa que sumamos sobre el índice que sustituye.

Es la suma de las n_i observaciones del grupo *i*

$$\bar{y}_{i\bullet} = \frac{y_{i\bullet}}{n_i}$$

Es la media de la n_i observaciones del grupo *i*

$$y_{\bullet\bullet} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}$$

$$\bar{y}_{\bullet\bullet} = \frac{y_{\bullet\bullet}}{n_1 + n_2 + \ldots + n_I}$$

$$n_1 + n_2 + \ldots + n_I = n$$

El modelo teórico

Las observaciones se describen según el siguiente modelo lineal:

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

Es la media global de *y*

Lo que se desvía la media de *y* en el grupo *i*-ésimo con respecto a la media global de *y*

Es el error aleatorio. Lo que se desvía la observación y_{ij} de su media de grupo. Es la pertubación debida al error experimental

$\mu_i = \mu + \tau_i$ Media de *y* en el grupo *i*-ésimo

Hipótesis del modelo

Los errores del modelo son variables aleatorias con distribución normal, de media *cero* y varianza σ^2

Esta varianza se supone constante para todos los niveles (o grupos) del factor

Es importante comprobar que estas hipótesis se verifican para poder sacar conclusiones fiables a partir de un análisis de la varianza.

Más adelante veremos con un ejemplo, cómo comprobar que los datos cumplen las hipótesis del modelo.

Estimación del modelo

- En el modelo teórico existen ciertos parámetros desconocidos que estimaremos utilizando los datos observados.
- Existen I+1 parámetros desconocidos, las I medias de grupo y la varianza del error experimental.
- Para estimar estos parámetros utilizaremos el método de máxima verosimilitud.
- Para ello, primero necesitamos definir la función de verosimilitud *L* y maximizarla.
- Maximizar *L* será equivalente a maximizar el logaritmo neperiano de *L*, ln*(L)*.
- Para maximizar ln*(L)*, derivamos con respecto a los I+1 parámetros desconocidos, igualamos a cero las I+1 derivadas que obtenemos y resolvemos el sistema de I+1 ecuaciones que resulta (en este sistema las incógnitas son los parámetros desconocidos del modelo).

Estimación por máxima verosimilitud

En base a las hipótesis del modelo se verifica que:

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

Estos parámetros del modelo se suponen fijos, y por tanto, no aleatorios

$u_{ij} \sim N(0, \sigma^2)$ ⇒ $y_{ij} \sim N(\mu_i, \sigma^2)$

La función de verosimilitud es: $\mathrm{L}(\mu_1, \ldots, \mu_I, \sigma^2) = \prod_{i=1}^{I} \prod_{j=1}^{n_i} f(y_{ij})$

donde:

$$f(y_{ij}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_{ij} - \mu_i)^2}{2\sigma^2} \right\}$$

es la función de densidad de una normal con media μ_i y varianza σ^2

Estimación por máxima verosimilitud

Derivamos el logaritmo de *L* con respecto a los parámetros desconocidos e igualamos a cero dichas derivadas.

$$\ln(L(\mu_1,\ldots,\mu_I,\sigma^2)) = -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2$$

$$\frac{\partial \ln(L(\mu_1,\ldots,\mu_I,\sigma^2))}{\partial \mu_i} = 0 \quad \Rightarrow \quad \tfrac{1}{\sigma^2}\sum_{j=1}^{n_i}(y_{ij}-\mu_i) = \tfrac{n_i}{\sigma^2}(\bar{y}_{i\bullet}-\mu_i) = 0$$

$$\Rightarrow \quad \hat{\mu}_i = \bar{y}_{i\bullet}$$

Cada media de grupo se estima mediante la media muestral de las observaciones *y* obtenidas en ese grupo

$$\frac{\partial \ln(L(\hat{\mu}_1,\ldots,\hat{\mu}_I,\sigma^2))}{\partial \sigma^2} = 0 \quad \Rightarrow \quad \tfrac{-n}{2\sigma^2} + \tfrac{1}{2(\sigma^2)^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\hat{\mu}_i) = 0$$

$$\Rightarrow \quad -n + \tfrac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\hat{\mu}_i)^2 = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\hat{\mu}_i)^2}{n}$$

Estimación de la varianza

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2}{n}$$

Este estimador de la varianza presenta un problema. Se trata de un estimador sesgado.

Un buen estimador de la varianza debería ser insesgado, i.e. debería verificar que su media fuese igual a la varianza (el parámetro que estima). Sin embargo sucede que:

$$E(\hat{\sigma}^2) \neq \sigma^2$$

Buscaremos otro estimador de la varianza que sea insesgado.

Pero antes de ello, definiremos los residuos y veremos cómo expresar $\hat{\sigma}^2$ en función de los residuos.

Estimación de la varianza

De acuerdo con el modelo: $u_{ij} = y_{ij} - \mu_i$

(Se sustituye por su estimación)

Así que podemos estimar los errores mediante: $\hat{u}_{ij} = y_{ij} - \hat{\mu}_i$

A estas estimaciones de los errores o perturbaciones del modelo, se les llama **residuos** y los denotaremos por e_{ij}

$$e_{ij} = \hat{u}_{ij} = y_{ij} - \bar{y}_{i\bullet}$$

Estos residuos miden la variabilidad de *y* no explicada por el modelo.

Además, sucede que: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} e_{ij}^2 = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(e_{ij} - \bar{e})^2$

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} e_i = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})$$

$$= \frac{1}{n}\sum_{i=1}^{I}(y_{i\bullet} - n_i\bar{y}_{i\bullet}) = \frac{1}{n}\sum_{i=1}^{I}(y_{i\bullet} - y_{i\bullet}) = 0$$

$\hat{\sigma}^2$ es la varianza de los residuos

Estimación de la varianza

Los residuos no son todos independientes entre sí.

Nótese que los residuos satisfacen las I ecuaciones (véase pág. 12) que nos permitieron obtener estimadores para la media de cada grupo, i.e: Para cada *i=1,…,I*, se verifica que:

$$\sum_{j=1}^{n_i}(y_{ij}-\hat{\mu}_i)=0$$

Son los residuos

Esto implica que si conocemos el valor de n-I residuos, podemos encontrar los restantes I residuos resolviendo las I ecuaciones anteriores.

Así que, sólo n-I residuos son independientes entre sí.

Para estimar la varianza del error, consideraremos una modificación de $\hat{\sigma}^2$ por grados de libertad, es decir, dividiremos entre el número de residuos independientes en lugar de entre el total de residuos.

Esto dará lugar a la varianza residual: $\hat{s}_R^2=\frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}e_{ij}^2$

Estimación de la varianza

Como ya sucedió en otras ocasiones, utilizaremos entonces la varianza residual para estimar la varianza del error, que es una corrección de $\hat{\sigma}^2$ por grados de libertad.

$$\hat{s}^2_R = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} e^2_{ij}$$

Dividimos entre (*n-I*) en lugar de *n*

$$= \frac{1}{n-I}\sum_{i=1}^{I}(n_i - 1)\hat{S}^2_i$$

Se trata de una media ponderada de las cuasivarianzas de cada grupo

$$\hat{S}^2_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2$$

Cuasivarianza de *y* en el grupo *i*-ésimo

Estimación de la varianza

Se puede comprobar que $\hat{s}_R^2 = \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} e_{ij}^2$ sí es un estimador insesgado para σ^2

Propiedades de los estimadores

$$\hat{\mu}_i \sim N(\mu_i, \sigma^2/n_i)$$

Si conociésemos sigma, un Intervalo de Confianza con nivel de confianza *1-alpha*, para la media del grupo *i*, vendría dado por:

$$\hat{\mu}_i \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n_i}}$$

Es el valor de una normal estándar que deja a su derecha una probabilidad de magnitud: $\alpha/2$

Pero σ es desconocido, así que se sustituye por la raíz cuadrada de la cuasivarianza de *y* en el grupo *i* y lo que se obtiene es el siguiente Intervalo de Confianza:

Es el valor de una *t* de Student con n_i-1 g.l. que deja a su derecha una probabilidad de magnitud: $\alpha/2$

$$\hat{\mu}_i \pm t_{\alpha/2, n_i - 1}\frac{\hat{S}_i}{\sqrt{n_i}}$$

Propiedades de los estimadores

$$\hat{s}_R^2 = \frac{1}{n-I}\sum_{i=1}^{I}(n_i-1)\hat{S}_i^2$$

Se verifica que:

$$\frac{(n_i-1)\hat{S}_i^2}{\sigma^2} = \frac{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\bullet})^2}{\sigma^2} \sim \chi^2_{n_i-1}$$

La suma de variables aleatorias chi cuadrado sigue una distribución chi cuadrado con g.l igual a la suma de los g.l de cada componente en la suma

$$\frac{(n-I)\hat{s}_R^2}{\sigma^2} = \frac{\sum_{i=1}^{I}(n_i-1)\hat{S}_i^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{I}(n_i-1)}$$

Son los grados de libertad (g.l.)

$$\frac{(n-I)\hat{s}_R^2}{\sigma^2} = \frac{\sum_{i=1}^{I}(n_i-1)\hat{S}_i^2}{\sigma^2} \sim \chi^2_{n-I}$$

$$\sum_{i=1}^{I}(n_i-1) = n-I$$

Objetivo: Comparar los grupos

Una vez estimadas las medias de grupo y la varianza del error, a partir de los datos, podremos realizar comparaciones entre grupos.

Los grupos se compararán a través de sus medias de grupo, pero también teniendo en cuenta su variabilidad.

ANOVA

Nos interesará, contrastar en primer lugar si existen diferencias estadísticamente significativas entre las medias de grupo.

Método de Fischer

Si este contraste nos indica que sí existen diferencias, entonces en segundo lugar nos interesará saber qué par de medias (es decir, qué par de grupos) se diferencian entre sí

Comparación de medias cuando hay dos niveles

Si sólo hay dos grupos podemos utilizar los intervalos de confianza y contrastes de hipótesis para comparar las medias de dos poblaciones normales.

Un estimador puntual de $\mu_1 - \mu_2 \rightarrow \bar{y}_{1\bullet} - \bar{y}_{2\bullet} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$

Consideremos la hipótesis nula de igualdad de medias: H_0

Interesa contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$

frente a la hipótesis alternativa $H_1 : \mu_1 \neq \mu_2$

Estandarizando y bajo H_0

$$d = \frac{\bar{y}_{1\bullet} - \bar{y}_{2\bullet}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Comparación de medias cuando hay dos niveles (contraste de hipótesis)

$$d = \frac{\bar{y}_{1\bullet} - \bar{y}_{2\bullet}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$ bajo H_0

σ^2 es desconocida → $\hat{S}_T^2 = \frac{(n_1-1)\hat{S}_1^2 + (n_2-1)\hat{S}_2^2}{n_1+n_2-2}$

(Se estima utilizando una media ponderada de las cuasivarianzas de *y* en el grupo 1 y 2)

Se verifica que: $\frac{(n_1+n_2-2)\hat{S}_T^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$

$$t = \frac{\bar{y}_{1\bullet} - \bar{y}_{2\bullet}}{\hat{S}_T\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Si $\left| \frac{\bar{y}_{1\bullet} - \bar{y}_{2\bullet}}{\hat{S}_T\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > t_{\alpha/2, n_1+n_2-2}$

se rechaza la hipótesis nula H_0

Comparación de medias cuando hay dos niveles (IC para la diferencia de medias)

$$\bar{y}_{1\bullet} - \bar{y}_{2\bullet} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

$$d = \frac{(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Además, $\frac{(n_1+n_2-2)\hat{S}_T^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$

$$t = \frac{(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) - (\mu_1 - \mu_2)}{\hat{S}_T\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Intervalo de confianza para $\mu_1 - \mu_2$ con nivel de confianza $1 - \alpha$

$$(\bar{y}_{1\bullet} - \bar{y}_{1\bullet}) \pm t_{\alpha/2, n_1+n_2-2}\hat{S}_T\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Comparación de medias cuando hay más de dos niveles

Cuando existen más de dos grupos, la comparación de medias se hará a través del análisis de la varianza

ANOVA →

Primero contrastaremos la hipótesis nula de igualdad de las I medias frente a la alternativa de que al menos una de las medias difiere de las demás. Esto lo haremos a través de la tabla ANOVA (en la que veremos cómo se descompone la variabilidad total de los datos *y*).

Método de Fischer →

Si este contraste nos indica que debemos rechazar la hipótesis nula, entonces trataremos de ver qué par de medias difieren entre sí, a través de un contraste conjunto en el que simultáneamente se contrastará la igualdad de todos los pares posibles de medias.

Existen varios métodos para llevar a cabo este contraste simultáneo. Aquí veremos el método de Fischer o LSD (least square deviation).

Descomposición de la variabilidad de la variable dependiente *y*

$$VT = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2$$

(La distancia entre la observación y_{ij} de la media global se descompone en la suma de lo que la observación y_{ij} dista de su media de grupo *i* + lo que dista la media de grupo *i* de la media global.)

$$(y_{ij} - \bar{y}_{\bullet\bullet}) = (y_{ij} - \bar{y}_{i\bullet}) + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$$

↓ (elevamos al cuadrado)

$$(y_{ij} - \bar{y}_{\bullet\bullet})^2 = (y_{ij} - \bar{y}_{i\bullet})^2 + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 + 2(y_{ij} - \bar{y}_{i\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$$

↓ (sumamos en *i* y en *j*)

$$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2 + \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$

↑ (el término cruzado se anula)

$$2 \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) = 2 \sum_{i=1}^{I} \left((\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet}) \right)$$

$$y_{i\bullet} - n_i \bar{y}_{i\bullet} = y_{i\bullet} - y_{i\bullet} = 0$$

Descomposición de la variabilidad de la variable dependiente *y*

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\bullet\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2$$

VT = variabilidad total

VNE= variabilidad no explicada o residual, también conocida como *variabilidad intra grupos*

VE = variabilidad explicada por el modelo, también conocida como *variabilidad entre grupos*

Nótese que:

$$\hat{s}_R^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} e_{ij}^2 = \frac{VNE}{n-I}$$

Anteriormente vimos que: $E(\hat{s}_R^2) = \sigma^2 \Rightarrow E(\frac{VNE}{n-I}) = \sigma^2$

También se puede demostrar que: $E(\frac{VE}{I-1}) = \sigma^2 + \frac{\sum_{i=1}^{I} n_i \tau_i^2}{I-1}$

ANOVA. Contraste de hipótesis

Estamos interesados en contrastar la hipótesis nula de igualdad de medias: $H_0 : \mu_1 = \ldots = \mu_I = \mu$

frente a la alternativa: $H_1 : \mu_j \neq \mu_k$, para algún $j, k \in \{1, \ldots, I\}$

$$\mu_i = \mu + \tau_i$$

Esto es equivalente a contrastar: $H_0 : \tau_1 = \ldots = \tau_I = 0$

frente a la alternativa: $H_1 : \tau_j \neq 0$, para algún $j \in \{1, \ldots, I\}$

Sabemos que:

$$E\left(\frac{VNE}{n-I}\right) = \sigma^2$$

$$E\left(\frac{VE}{I-1}\right) = \sigma^2 + \frac{\sum_{i=1}^{I} n_i \tau_i^2}{I-1}$$

Es un término >=0

Bajo H_0 tenemos dos estimadores insesgados de la varianza.

Si H_0 es falsa, se espera que

$$\frac{VE/(I-1)}{VNE/(n-I)} > 1$$

Además, cuanto más grande sea este cociente, más evidencia habrá de que H_1 es cierta y no H_0.

ANOVA. Contraste de hipótesis

$\frac{VE/(I-1)}{VNE/(n-I)} > 1$

¿Cuánto de grande debe ser este cociente para rechazar H_0? Si es ligeramente mayor que 1, no rechazaremos H_0.

Para responder a esta pregunta necesitamos conocer la distribución de este cociente bajo H_0.

Ya que valores grandes nos dan evidencia de que H_0 es falsa, la región de rechazo habrá que buscarla en la cola derecha de la distribución de ese cociente (que es la cola de la distribución correspondiente a valores más grandes).

Ya vimos que: $\hat{s}_R^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} e_{ij}^2 = \frac{VNE}{n-I}$

$\frac{(n-I)\hat{s}_R^2}{\sigma^2} = \frac{\sum_{i=1}^{I}(n_i-1)\hat{S}_i^2}{\sigma^2} \sim \chi^2_{n-I}$ → $\frac{VNE}{\sigma^2} \sim \chi^2_{n-I}$

ANOVA. Contraste de hipótesis

Bajo H_0 se verifica que: $\frac{VE}{\sigma^2} \sim \chi^2_{I-1}$

Una distribución F de Snedecor sabemos que se obtiene a partir de distribuciones chi cuadrado del siguiente modo:

$$F_{n,m} = \frac{\chi^2_n / n}{\chi^2_m / m}$$

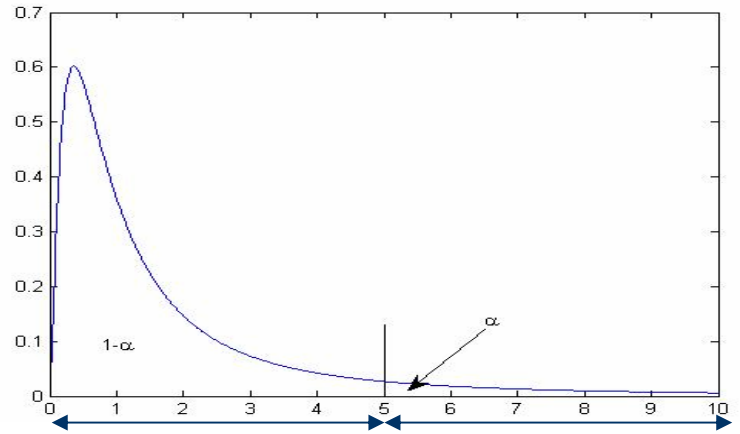
$$\frac{\frac{VE}{\sigma^2(I-1)}}{\frac{VNE}{\sigma^2(n-I)}} = \frac{VE/(I-1)}{VNE/(n-I)} \sim F_{I-1,n-I}$$

$\hat{s}^2_e$ (numerador) $\qquad$ $\hat{s}^2_R$ (denominador)

Hay que ver dónde cae este valor

$$F_{I-1,n-I} = \frac{\hat{s}^2_e}{\hat{s}^2_R}$$

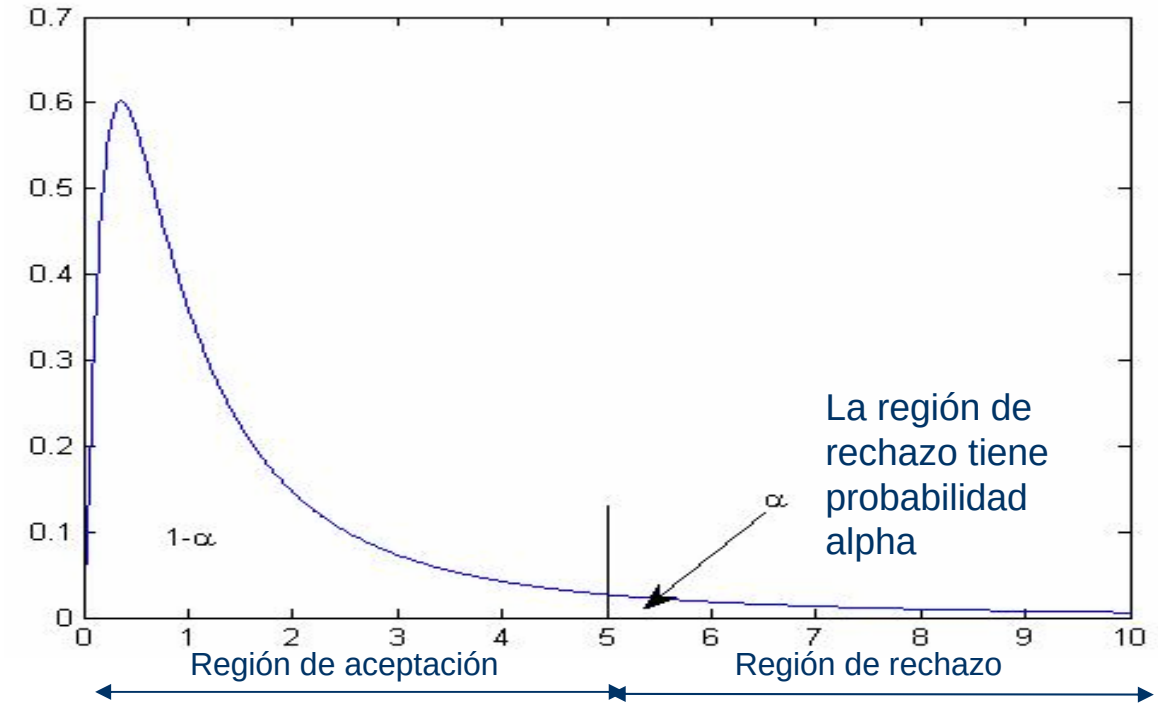
si en la región de rechazo o en la de aceptación.



ANOVA. Contraste de hipótesis

$$\frac{\hat{s}_e^2}{\hat{s}_R^2}$$

En base a este valor y su *p-valor* asociado, podremos rechazar o no, la hipótesis nula de igualdad de medias.



El *p-valor* asociado al test F

$$test\ F = \frac{\hat{s}_e^2}{\hat{s}_R^2}$$

es la probabilidad que queda a la derecha de ese valor.

Si es menor que alpha, el test F cae en la región de rechazo, así que rechazamos H_0.

En caso contrario, aceptamos H_0. No hay evidencia suficiente para recharzarla.

Tabla ANOVA: descomposición de la variabilidad

Fuentes de variación	Suma de Cuadrados (SC)	Grados de Libertad (g.l.)	Varianza (cuadrado medio) (SC/g.l.)	Test F $F_{I-1,n-I}$
Variabilidad explicada = variabilidad entre grupos	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 = \sum_{i=1}^{I} n_i(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2$	I-1	$\hat{s}_e^2$	$\frac{\hat{s}_e^2}{\hat{s}_R^2}$
Variabilidad no explicada = Variabilidad intra grupos	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\bullet})^2$	n-I	$\hat{s}_R^2$	Si el *p-valor* asociado al test F es pequeño, se rechaza H_0. Un *p-valor* pequeño significa que el test F ha caído muy a la derecha, en la cola derecha de la distribución, y por tanto el F test ha salido muy grande.
Variabilidad total	$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\bullet\bullet})^2$	n-1	$\hat{S}_y^2$	

Método de Fischer o LSD (Least significative distance)

Hemos visto anteriormente, que para hacer un contraste de la igualdad de dos medias, podíamos utilizar:

$$t = \frac{(\bar{y}_{1\bullet} - \bar{y}_{2\bullet}) - (\mu_1 - \mu_2)}{\hat{S}_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$\hat{S}_T^2 = \frac{(n_1-1)\hat{S}_1^2 + (n_2-1)\hat{S}_2^2}{n_1+n_2-2}$$

En el caso de que existan más de dos grupos, como estamos trabajando bajo la hipótesis de que en todos los grupos la variabilidad es la misma, es decir estamos suponiendo que tienen la misma varianza σ^2, podremos entonces, utilizar la información contenida en los datos de todos los grupos para estimar esa varianza, en vez de usar simplemente los datos de los dos grupos, cuyas medias queremos comparar.

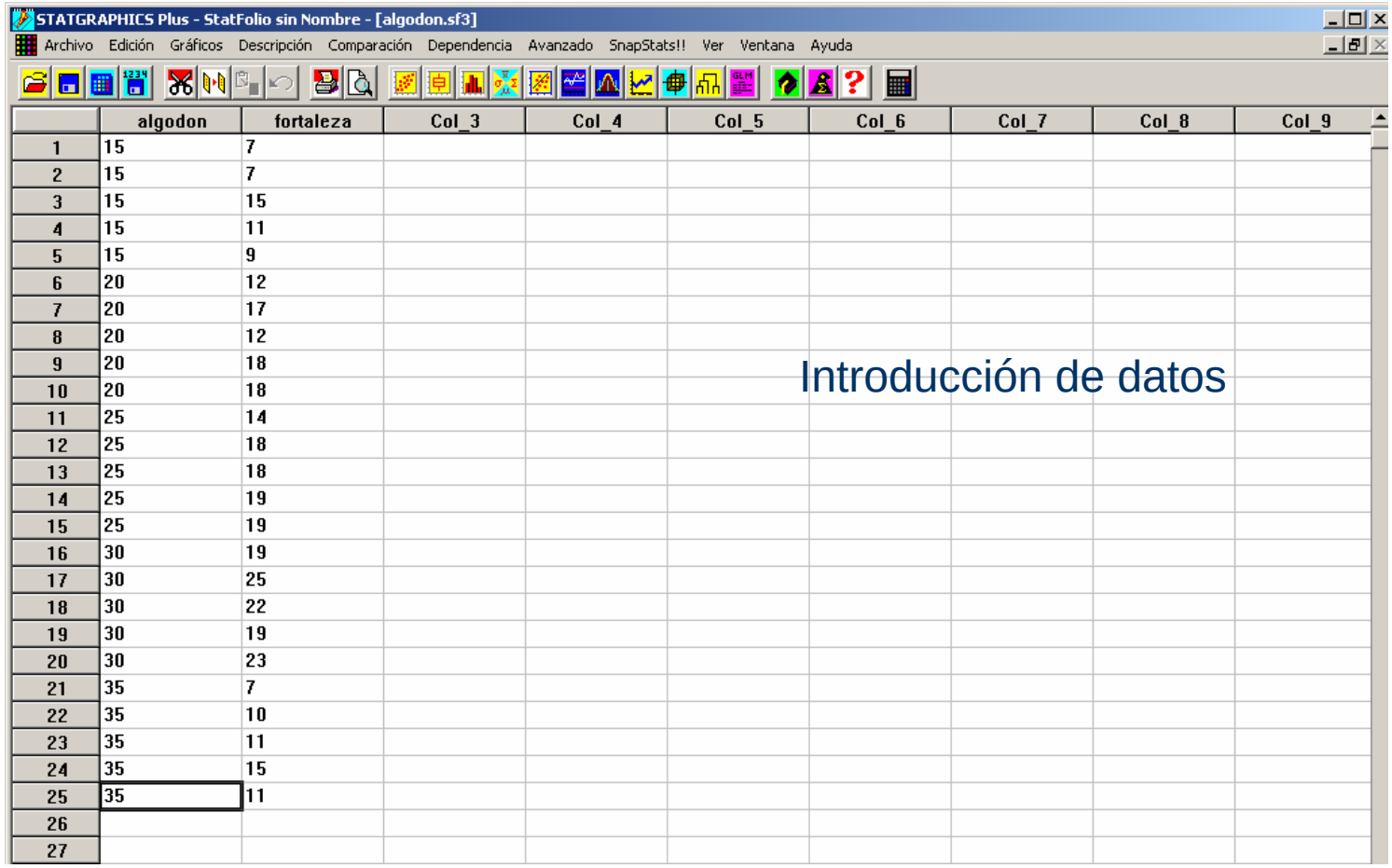
Así que, en vez de utilizar $\hat{S}_T^2$, utilizaremos la varianza residual $\hat{s}_R^2$, en la *t* de Student (con n_1+n_2-2 g.l.) que nos permite realizar el contraste.

En esto consiste el método de Fischer o LSD. La ventaja es que se realizan las comparaciones dos a dos de modo simultáneo y se consiguen detectar diferencias más pequeñas.

Volviendo al ejemplo introductorio

% de algodón	Observaciones (fortaleza de las 25 fibras fabricadas)					Total	Promedio
15%	7	7	15	11	9	49	9.8
20%	12	17	12	18	18	77	15.4
25%	14	18	18	19	19	88	17.6
30%	19	25	22	19	23	108	21.6
35%	7	10	11	15	11	54	10.8
						376	15.04

Analicémoslo con el Statgraphics

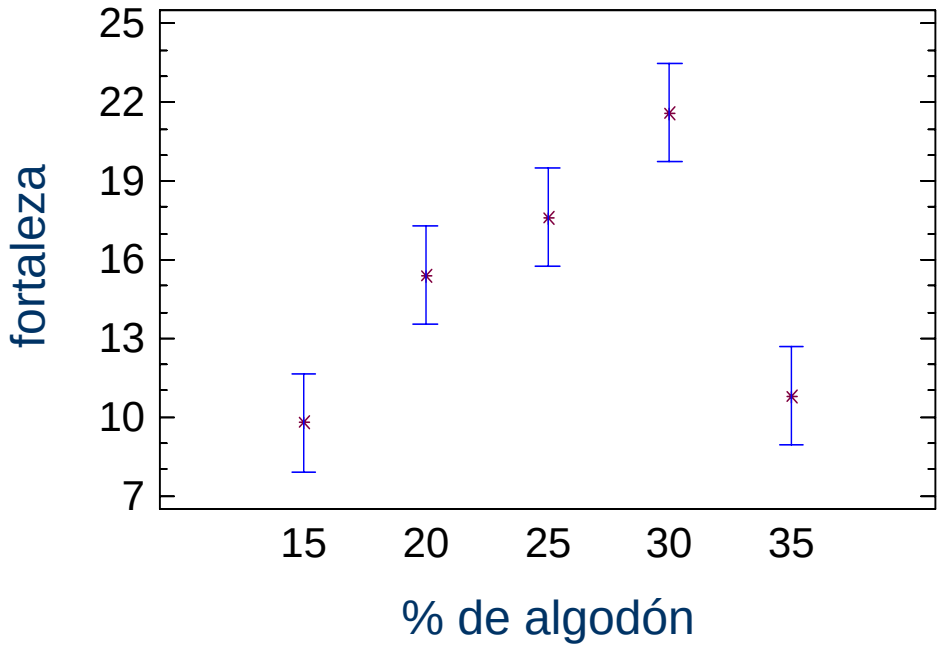


	algodon	fortaleza	Col_3	Col_4	Col_5	Col_6	Col_7	Col_8	Col_9
1	15	7							
2	15	7							
3	15	15							
4	15	11							
5	15	9							
6	20	12							
7	20	17							
8	20	12							
9	20	18							
10	20	18							
11	25	14							
12	25	18							
13	25	18							
14	25	19							
15	25	19							
16	30	19							
17	30	25							
18	30	22							
19	30	19							
20	30	23							
21	35	7							
22	35	10							
23	35	11							
24	35	15							
25	35	11							
26									
27									

Introducción de datos

Medias por cada grupo

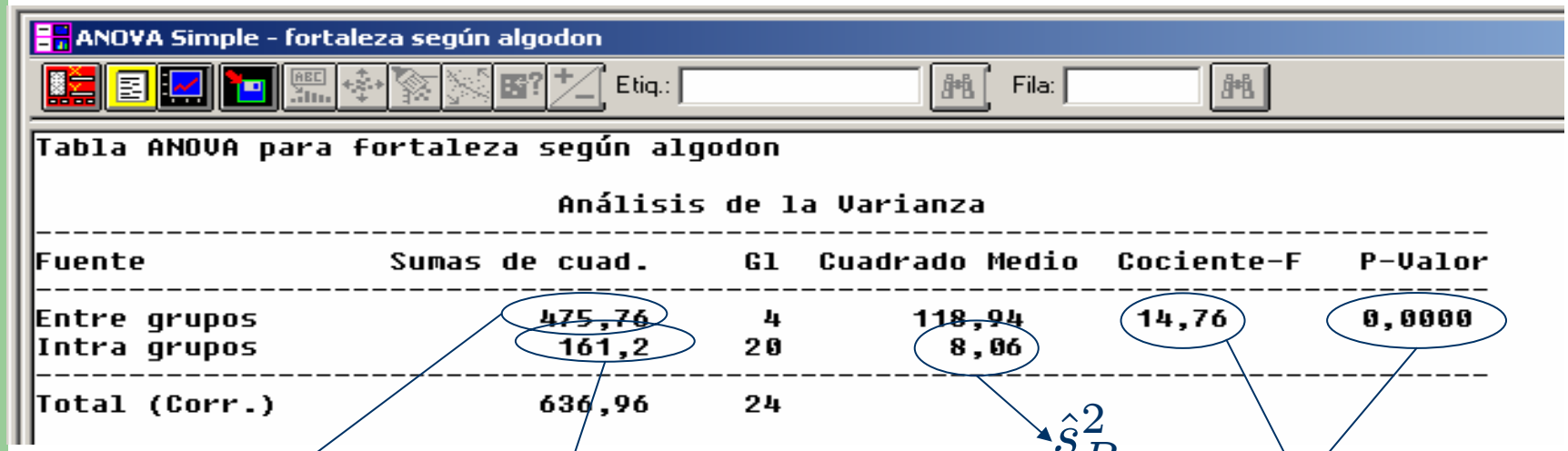
¿Son todas las medias iguales?



A la vista de este gráfico de medias, se ve que las medias difieren unas de otras.

Usando un 30% de algodón parece que se fabrican las mejores fibras, es decir, las de mayor fortaleza

Tabla ANOVA



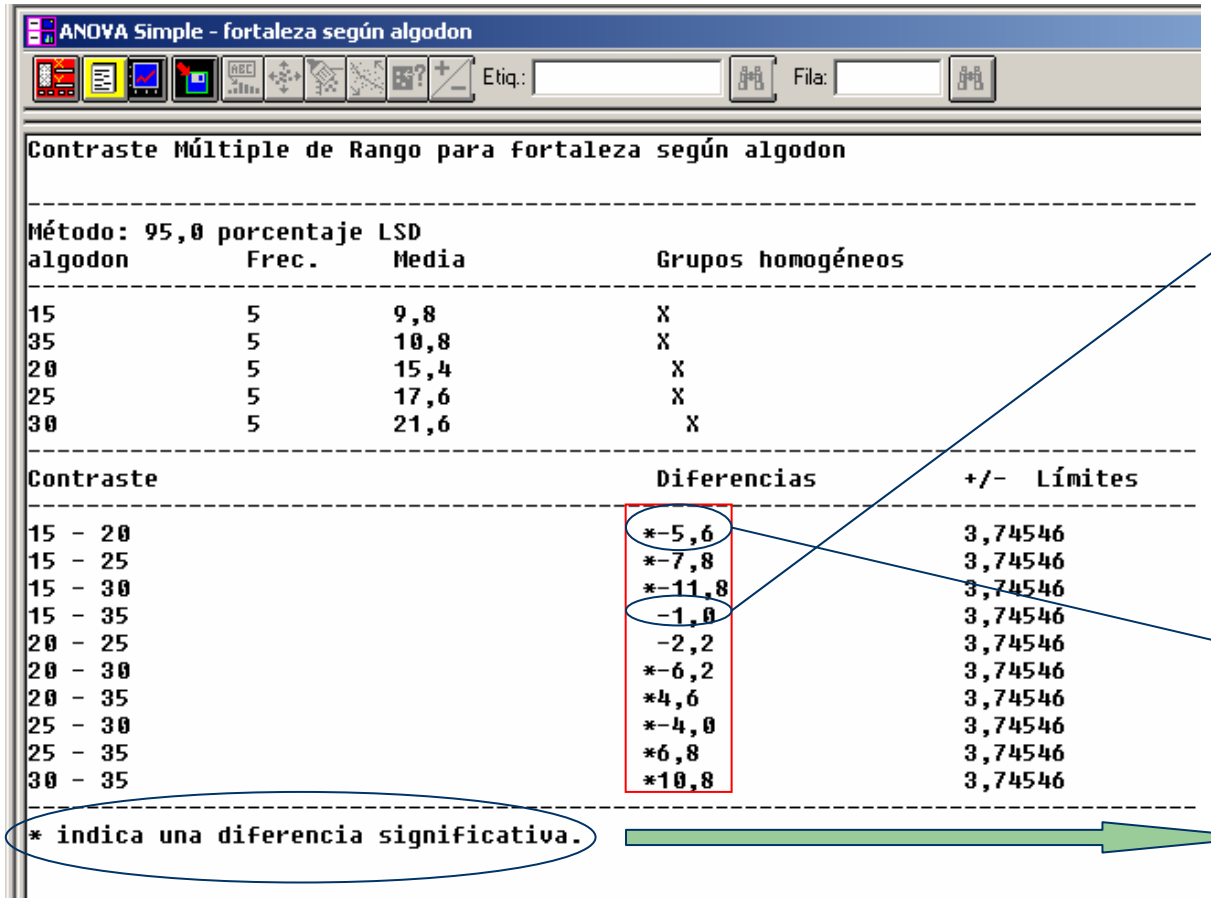
Variabilidad explicada por el modelo, también conocida como variabilidad entre grupos.

Variabilidad no explicada por el modelo, también conocida como variabilidad intra grupos

Estadístico o test F

Se detectan diferencias significativas entre las medias.

Comparación simultánea de cada par de medias (método de Fischer o LSD)



Contraste Múltiple de Rango para fortaleza según algodon

Método: 95,0 porcentaje LSD

algodon	Frec.	Media	Grupos homogéneos
15	5	9,8	X
35	5	10,8	X
20	5	15,4	X
25	5	17,6	X
30	5	21,6	X

Contraste	Diferencias	+/- Límites
15 - 20	*-5,6	3,74546
15 - 25	*-7,8	3,74546
15 - 30	*-11,8	3,74546
15 - 35	-1,0	3,74546
20 - 25	-2,2	3,74546
20 - 30	*-6,2	3,74546
20 - 35	*4,6	3,74546
25 - 30	*-4,0	3,74546
25 - 35	*6,8	3,74546
30 - 35	*10,8	3,74546

* indica una diferencia significativa.

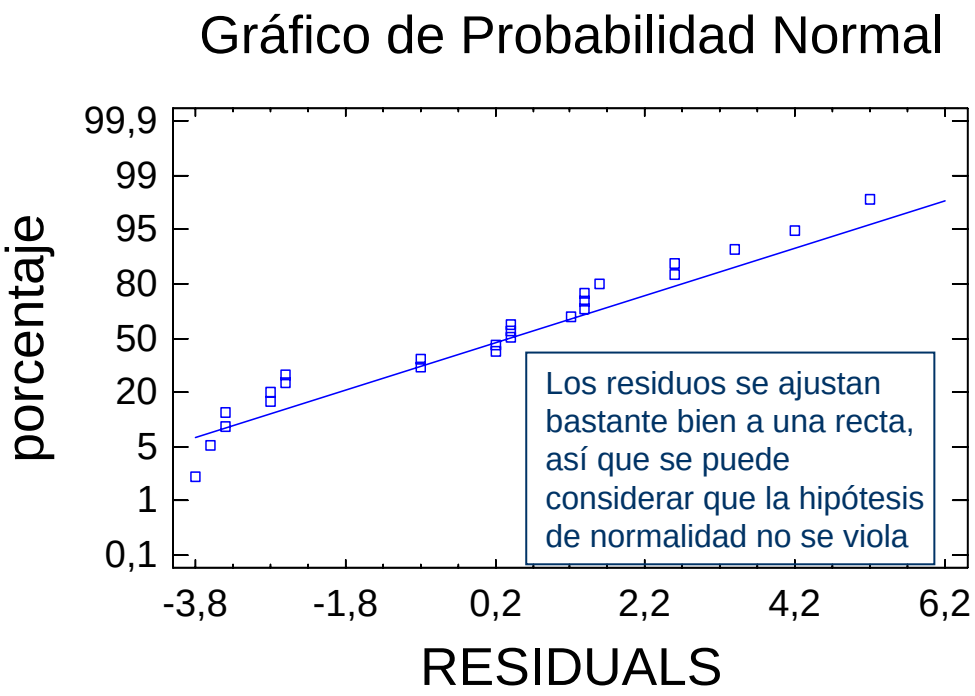
Los niveles de 15% y 35% de algodón no son significativamente distintos.

En cambio, sí se han detectado diferencias entre los niveles de 15% y 35% de algodón

Diagnosis: Normalidad

Para comprobar la suposición de normalidad podemos utilizar la gráfica de probabilidad normal de los residuos.

Si la distribución de los errores es normal, esta gráfica parecerá una línea recta.



Pasos a seguir

1. Después de haber realizado el análisis ANOVA de un factor, en el botón de “guardar resultados”, le pediremos que nos guarde los residuos (RESIDUALS). Aparecerá en la hoja de datos una nueva columna con los residuos.
2. Vamos a hacer un análisis unidimensional de los residuos: Menú Descripción>Datos Numéricos>Análisis unidimensional y metemos los residuos (RESIDUALS) en Datos.
3. En las opciones gráficas del análisis unidimensional pedimos que nos represente el gráfico de probabilidad normal.

Diagnosis: Normalidad

La gráfica de probabilidad normal es una representación gráfica de la distribución acumulada de los residuos sobre papel de probabilidad normal.

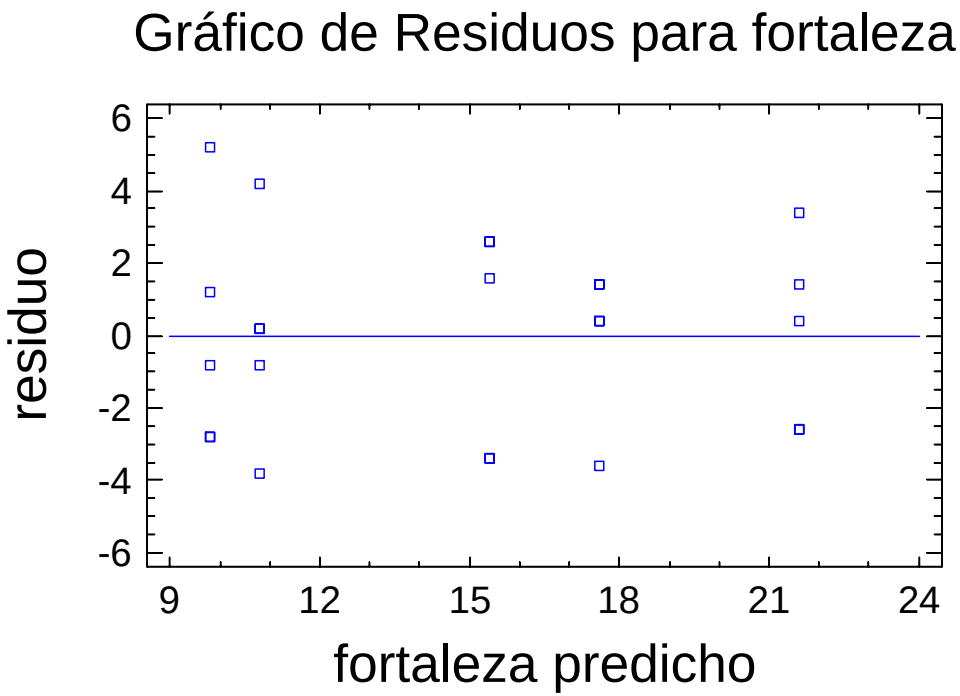
Cuando hablamos de papel de probabilidad normal nos referimos a aquel en el que la escala de ordenadas (el eje Y) es tal que si representamos la distribución acumulada de una normal lo que obtenemos es una recta.

Para obtener la gráfica de probabilidad normal, se ordenan los n residuos de menor a mayor y se representa el k-ésimo residuo ordenado contra su punto de probabilidad acumulada: (k-0.5)/n, en papel de probabilidad normal.

Este proceso lo hace directamente el Statgraphics siguiente los pasos descritos en la página anterior.

Diagnosis: linealidad y homocedasticidad

El gráfico de residuos versus predichos puede ayudarnos a detectar desviaciones de las hipótesis de linealidad y homocedasticidad.



Después de haber realizado el análisis ANOVA de un factor, en el botón de "opciones gráficas", seleccionaremos la opción "Residuo frente a predicho" para que nos represente dicho gráfico.

En este gráfico no se observa ningún patrón ni forma de embudo, por lo que se puede considerar que los datos satisfacen las hipótesis de linealidad y homocedasticidad.

Diagnosis: linealidad y homocedasticidad

Es también interesante graficar los residuos frente a los valores ajustados (o predicciones). En esta gráfica no se deben revelar patrones obvios que, en el caso de aparecer, indicarían que la suposición de linealidad no se satisface.

Esta gráfica también sirve para detectar una violación de la hipótesis de homocedasticidad (o igualdad de varianzas). En ciertas ocasiones ocurre que la variabilidad de los datos aumenta a medida que aumenta la magnitud del dato. Esto suele suceder en instrumentos de medición, el error del instrumento de medición es proporcional a la escala de lectura. En situaciones como esta, la gráfica de residuos frente a predichos se ensanchará como un embudo hacia la derecha.

Cuando se viola la hipótesis de homogeneidad, el test F se ve sólo ligeramente afectado cuando se han tomado el mismo número de observaciones por grupo (es decir cuando estamos ante un diseño balanceado: $n_1 = n_2 = \ldots = n_I$).

Sin embargo, en diseños no balanceados, el problema es más importante, y especialmente si una de las varianzas es mucho mayor que el resto.

El problema de heterocedasticidad (distintas varianzas) se puede corregir transformando adecuadamente los datos mediante un logaritmo o una potencia. La transformación adecuada dependerá de cada conjunto de datos particular.

Diagnosis: Independencia

Para comprobar la suposición de independencia de los datos, es interesante graficar los residuos frente al orden temporal en el que éstos han sido recopilados.

Si en este gráfico se detecta una tendencia a tener rachas con residuos positivos y negativos, esto nos indicará que los datos no son independientes.

Si se han tomados los datos mediante un procedimiento de aleatorización (como ya se comentó al comienzo de esta presentación, véase pág. 4), entonces es de esperar que hayamos asegurado la independencia de las observaciones y que no se observen esas rachas.

Supongamos que a medida que avanza el proceso la habilidad del experimentador o experimentadores cambia a medida que el experimento se desarrolla (se hace más errático, debido al cansancio, o por el contrario, se hace más experto,por la experiencia adquirida). En situaciones como esta puede suceder que la varianza de los datos cambie con el tiempo. Este tipo de problema se puede detectar en el gráfico de residuos frente al tiempo, porque se verá cómo la dispersión de los residuos se hace mayor o menor a medida que el tiempo transcurre.

Es muy importante evitar este tipo de problemsa en el momento de la recogida de datos (en el momento de la experimentación). El análisis de la varianza es válido si, entre otros supuestos, se cumple el de varianza constante e independencia.