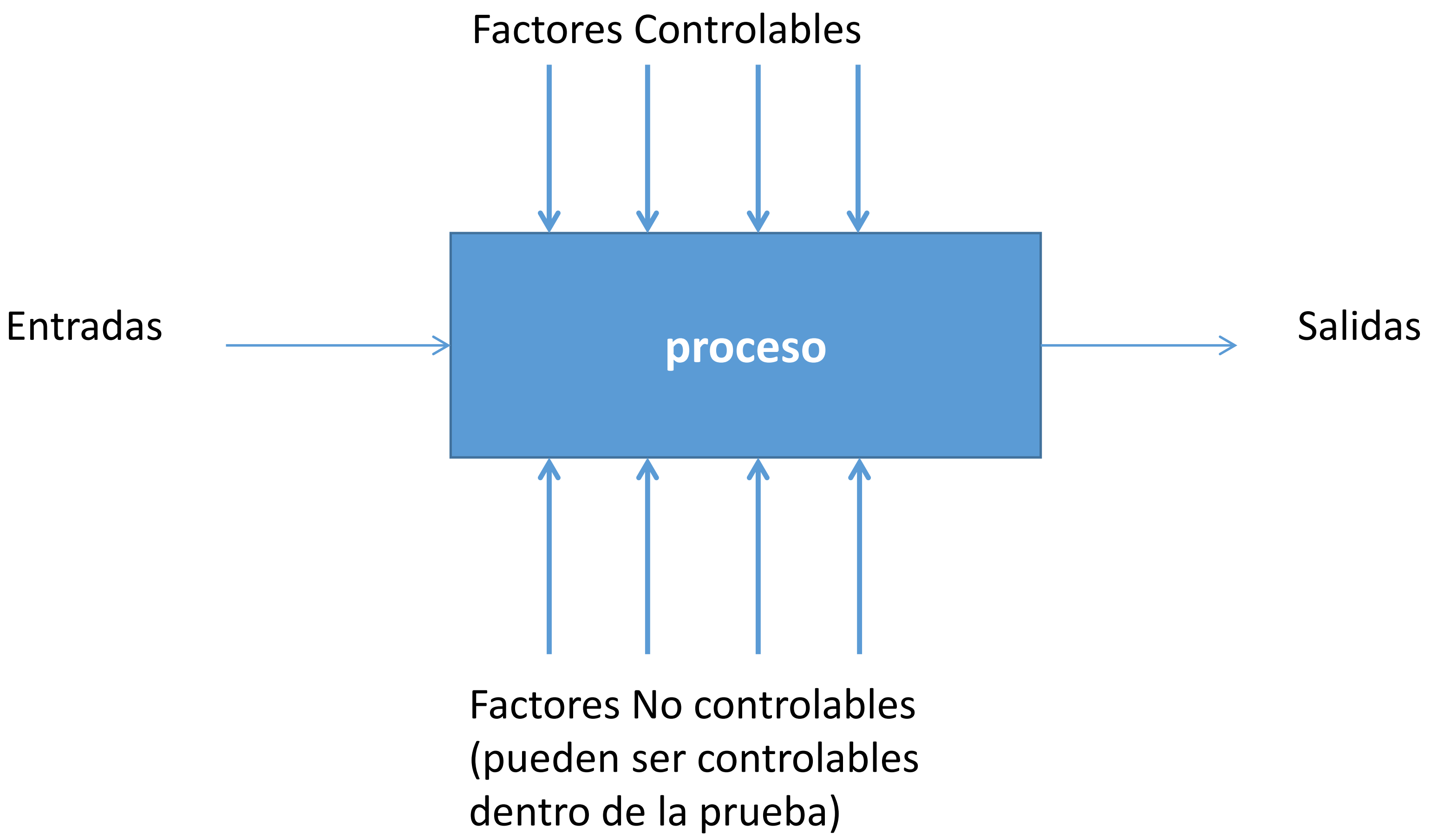
Experimentos con un solo factor: El análisis de varianza

Jhon Jairo Padilla Aguilar, PhD.

Experimentación en sistemas aleatorios:



Motivación

- Hasta ahora hemos comparado dos condiciones o tratamientos
- Otra forma de describir este tipo de experimento es como un experimento con un solo factor, con dos niveles del factor
- El factor es por ejemplo la formula utilizada para un compuesto
- Los niveles son los dos métodos diferentes para construir la fórmula
- Muchos experimentos de este tipo involucran más de dos niveles del factor
- Ahora estudiaremos los métodos para el diseño y el análisis de los experimentos con un solo factor con *a* niveles del mismo (ó *a* tratamientos)

Ejemplo: El objetivo del experimento

- Un ingeniero de desarrollo tiene interés en investigar la resistencia a la tensión de una fibra sintética nueva que se usará para hacer camisas para caballero.
- Por experiencia previa, se conoce que la resistencia a la tensión se afecta por el peso porcentual del algodón en la mezcla de materiales de la fibra.
- Se sospecha que al aumentar el contenido de algodón se incrementará la resistencia. (**Hipótesis**)
- Se sabe que el contenido de algodón deberá variar entre 10 y 40 por ciento para que el producto final tenga otras características de calidad que se desean (como la capacidad de ser sometido a un tratamiento de planchado permanente). (**Condiciones/restricciones de operación**)

Diseño del experimento

- El ingeniero decide probar ejemplares en cinco niveles de peso porcentual de algodón: 15, 20, 25, 30 y 35 por ciento.
- Se decide que se probarán 5 ejemplares de cada nivel de contenido de algodón.
- Este es un experimento con un solo factor (porcentaje de algodón) con a=5 niveles (5 porcentajes diferentes de algodón)
- También se decide probar con 5 ejemplares de cada nivel de contenido de algodón.

Justificación de la aleatoriedad en la secuencia

- La secuencia de prueba aleatorizada es necesaria para evitar que los efectos de variables perturbadoras desconocidas (fuera de control durante el experimento) contaminen los resultados.
- En el ejemplo podría suceder que si las corridas se hacen en orden ascendente de porcentaje de algodón (primero 5 corridas con 15%, luego 5 corridas con 20%, etc), y si la máquina de pruebas sufriese un calentamiento en la medida que pasan las pruebas, de tal manera que el calentamiento afecta las mediciones, esto destruirá la validez del experimento.

Determinación de la secuencia de las corridas

- Las 25 corridas deberán realizarse de manera aleatoria.
- Suponga que las corridas se numeran de la siguiente manera:

Peso porcentual del algodón	Número de corrida experimental				
15	1	2	3	4	5
20	6	7	8	9	10
25	11	12	13	14	15
30	16	17	18	19	20
35	21	22	23	24	25

Determinación de la secuencia de las corridas

- Para ejecutar las corridas se selecciona un número aleatorio entre 1 y 25.
- Suponga que el número aleatorio que salió escogido primero fue el 8 (20% de algodón).
- El procedimiento se repite hasta que las 25 observaciones tengan asignada una posición en la secuencia de la prueba.

Secuencia aleatoria de las observaciones

- Suponga que la secuencia de la prueba fue la siguiente:

Secuencia de prueba	Número de corrida	Peso porcentual del algodón
1	8	20
2	18	30
3	10	20
4	23	35
5	17	30
6	5	15
7	14	25
8	6	20
9	15	25
10	20	30
11	9	20
12	4	15
13	12	25
14	7	20
15	1	15
16	24	35
17	21	35
18	11	25
19	2	15
20	13	25
21	22	35
22	16	30
23	25	35
24	19	30
25	3	15

Resultados obtenidos

- Suponiendo que se siguió el orden aleatorio descrito anteriormente, se muestran las observaciones que se realizaron para la resistencia a la tensión en la siguiente tabla

Tabla 3-1 Datos (en lb/pulgada^2) del experimento de la resistencia a la tensión

Peso porcentual del algodón	Observaciones					Total	Promedio
	1	2	3	4	5		
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

Revisión gráfica de los resultados

- Siempre es bueno hacer una revisión gráfica de los resultados para observar ciertas tendencias.
- Diagrama de cajas

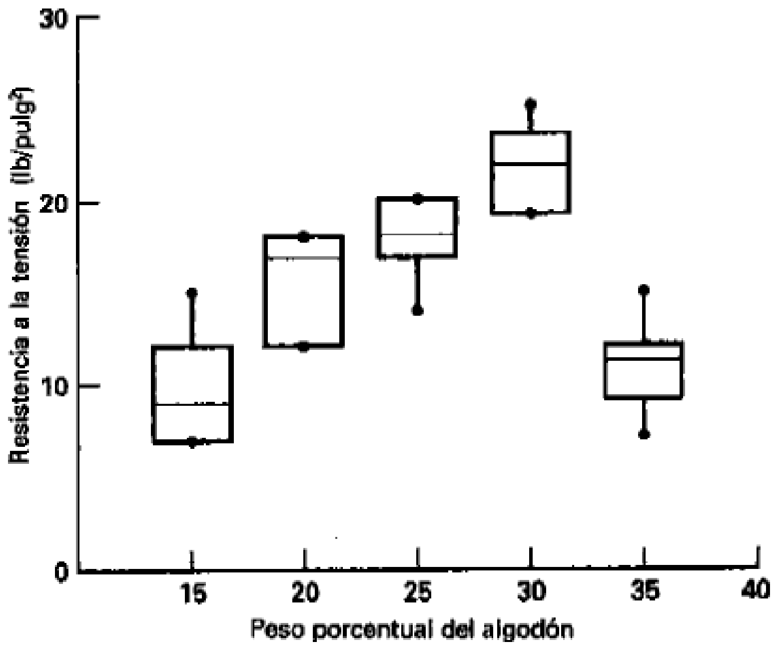
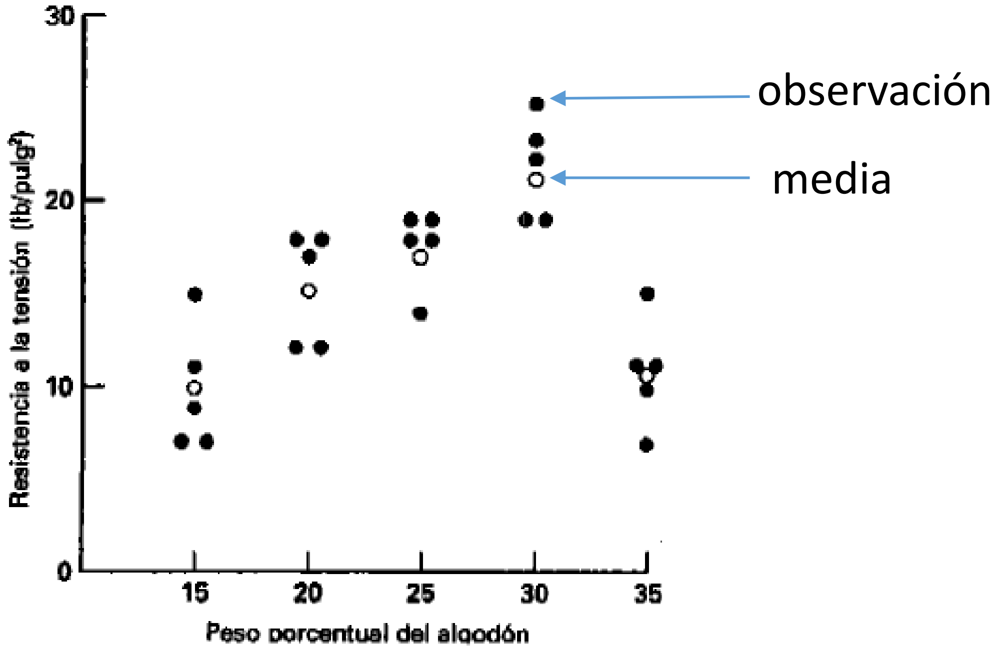


Diagrama de dispersión (puntos)



Conclusiones con base en los gráficos

- Ambas gráficas indican que la resistencia a la tensión se incrementa cuando el contenido de algodón se incrementa, hasta alcanzar el 30% de algodón.
- Después del 30% de algodón, hay un marcado descenso en la resistencia a la tensión.
- No hay evidencia sólida que sugiera que la variabilidad de la resistencia a la tensión alrededor del promedio dependa del peso porcentual del algodón.
- Se tienen firmes sospechas:
  - El contenido del algodón afecta la resistencia a la tensión
  - Alrededor del 30% de algodón produce la resistencia máxima

Conclusiones más objetivas

- Suponga que se quiere probar que las medias de los 5 niveles del factor son iguales.
- Una forma sería hacer comparaciones de a dos medias (con la prueba t), tomando todas las posibles combinaciones con las medias de los 5 niveles del factor.
- Este proceso es engorroso.
- Hay 10 posibles pares a comparar.
- Si la probabilidad de aceptar correctamente la hipótesis nula en cada prueba individual es de $1-\alpha=0.95$, la probabilidad de aceptar correctamente la hipótesis nula en las 10 pruebas es de $(0.95)^{10}$, si las pruebas son independientes. Por tanto, ocurre un incremento sustancial en el error de tipo I.
- El procedimiento correcto para probar la igualdad de varias medias es el **análisis de varianza**.

El análisis de varianza

ANOVA: Analysis of Variance

- Suponga que se tienen *a* tratamientos o niveles de un solo factor
- Se desea comparar los diferentes niveles del factor
- La variable de salida a comparar con cada uno de los niveles es una variable aleatoria

Tabla 3-2 Datos típicos de un experimento de un solo factor

Tratamiento (nivel)	Observaciones				Totales	Promedios
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$\underline{y_{a.}}$	$\underline{\bar{y}_{a.}}$
					$y_{..}$	$\bar{y}_{..}$

Modelos para los datos

- Las observaciones se pueden describir mediante un modelo matemático sencillo: **Modelo de las medias**

$$y_{ij} = \mu_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

- Donde:
  - y_{ij} es la observación ij-ésima,
  - μ_i es la media del nivel del factor o tratamiento i-ésimo, y
  - ε_{ij} es un componente de error aleatorio (incorpora demás fuentes de variabilidad del experimento: mediciones, factores no controlados, diferencias en los materiales de prueba, variabilidad con el tiempo, medio ambiente, etc)
- Es conveniente suponer que los errores tienen media cero para mantener la media

Modelos para los datos

- Otro modelo es considerar que la media tiene dos componentes:

$$\mu_i = \mu + \tau_i, \qquad i = 1, 2, \ldots, a$$

- A μ_i se le llama media del nivel, la cual se compone de una media global (μ) y un τ_i que es el *efecto del tratamiento i-ésimo*.
- Por tanto, el modelo matemático queda en definitiva como:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases}$$

- A este modelo se le llama el **modelo de los efectos**

Modelos de los datos

- Tanto el **modelo de las medias** como el **modelo de los efectos** son modelos lineales.
- La variable de respuesta y_{ij} es una función lineal de los parámetros del modelo
- Ambas formas son útiles, pero el **modelo de los efectos** se utiliza más en la literatura del diseño de experimentos.
- A cualquiera de estos modelos se les llama **modelo de análisis de varianza simple** o de un solo factor (únicamente se investiga un solo factor)

Modelo de análisis de varianza simple

- Requisitos:
  - El experimento debe llevarse a cabo en orden aleatorio (ambiente en que se apliquen los tratamientos debe ser lo más uniforme posible)
  - El diseño experimental es el de un experimento completamente aleatorizado
- Objetivos:
  - Probar las hipótesis apropiadas acerca de las medias de los tratamientos y estimarlas
- Supuestos:
  - Los errores del modelo son variables aleatorias que siguen una distribución normal e independiente con media cero y varianza σ^2.
  - La varianza es constante para todos los niveles del factor
  - Por tanto, las observaciones tienen una distribución normal, es decir, $y_{ij} \sim N(\mu+\tau_i, \sigma^2)$
  - Las observaciones son mutuamente independientes

Factor fijo o aleatorio?

- Modelo de Efectos fijos:

Los *a* tratamientos o niveles pudieron ser elegidos expresamente por el experimentador:

  - Quieren probarse hipótesis acerca de las medias de los tratamientos.
  - Las conclusiones se aplicarán únicamente a los niveles del factor considerados en el análisis.
  - Las conclusiones no pueden extenderse a tratamientos similares que no fueron considerados explícitamente.
  - También podría tratar de estimarse los parámetros del modelo (μ, τ_i, varianza)

Factor fijo o aleatorio?

- Modelo de Efectos aleatorios:
  - Los *a* tratamientos podrían ser una muestra aleatoria de una población más grande de tratamientos
  - Sería deseable poder extender las conclusiones (se basan en la muestra de los tratamientos) a la totalidad de los tratamientos de la población.
  - Las τ_i son variables aleatorias.
  - Lo que se obtiene sobre las τ_i particulares que se investigaron es relativamente inútil
  - Se suelen probar hipótesis sobre la variabilidad de las τ_i y se intenta estimar su variabilidad. A este modelo también se le llama **modelo de los componentes de la varianza**.
  - **Este modelo NO lo estudiaremos en este curso!!!!**

Modelo de Efectos fijos

Supuestos iniciales

- Se tiene un solo factor
- Se tienen efectos fijos
- Se realizará el análisis de varianza para esta situación

Definiciones

- $y_{i.}$: total de observaciones bajo el tratamiento i-ésimo.
- $\bar{y}_{i.}$: promedio de las observaciones bajo el tratamiento i-ésimo.
- $y_{..}$ Representa el gran total de todas las observaciones
- $\bar{y}_{..}$ Gran promedio de todas las observaciones

$$y_{i.} = \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{i.} = y_{i.} / n \qquad i = 1,2,\ldots,a$$

$$y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{..} = y_{..} / N \qquad N = an$$

N: Número total
de observaciones

Objetivo del experimento

- El interés se centra en probar la igualdad de las *a* medias de los tratamientos
- Es decir, se quiere probar que: $E(y_{ij}) = \mu + \tau_i = \mu_i \quad , \quad i = 1,2,\ldots,a$
- Por tanto, las Hipótesis nula y alternativa son:

$$H_0 = \mu_1 = \mu_2 = \ldots. = \mu_a$$

$$H_1 : \mu_i \neq \mu_j$$

Para al menos un par (*i,j*)

Objetivo del experimento

- Otra manera de ver el objetivo del experimento es:
  - Teniendo en cuenta que la media global es la media de las medias por factor:

$$\frac{\sum_{i=1}^{a} \mu_i}{a} = \mu$$

  - Y si se supone que los efectos de los factores se anulan

$$\sum_{i=1}^{a} \tau_i = 0$$

  - Tendremos que las hipótesis se pueden replantear como:

$$H_0 = \tau_1 = \tau_2 = ... = \tau_a = 0$$

$$H_1 : \tau_i \neq 0$$ Para al menos una i

Concluyendo…

- Por tanto, para probar la igualdad de las medias se puede probar que los efectos de los tratamientos (τ_i) son cero.
- Es decir, se puede utilizar el análisis de varianza para probar esto.

Análisis de Varianza

- Su nombre se deriva de la partición de la variabilidad total en sus partes componentes.
- La variabilidad global de los datos se mide mediante la suma de cuadrados total corregida:

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$$

- Esto es razonable porque si la SS_T tuviese que dividirse por el número apropiado de grados de libertad (*an-1=N-1*), se obtendría la varianza muestral de las *y.*

Identidad de Análisis de Varianza

- Se puede demostrar que:

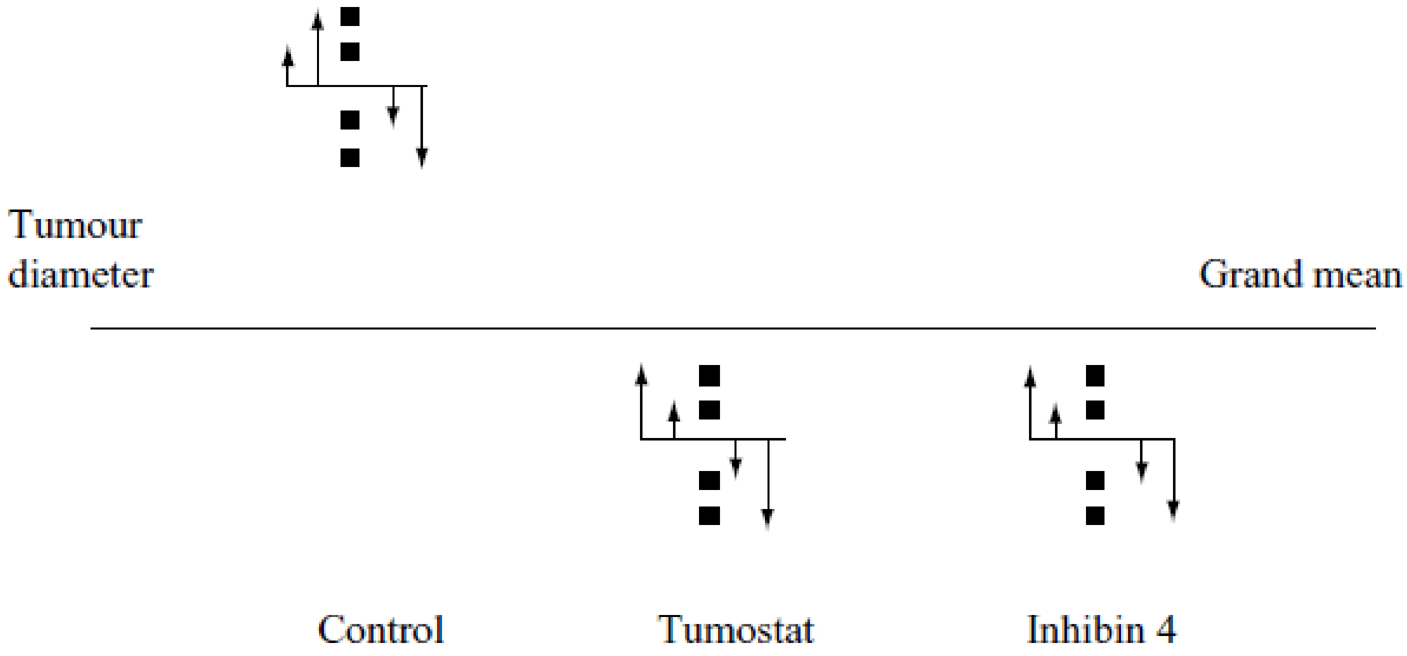
$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

Pueden deberse al error aleatorio

Medida de las diferencias entre las medias de los tratamientos

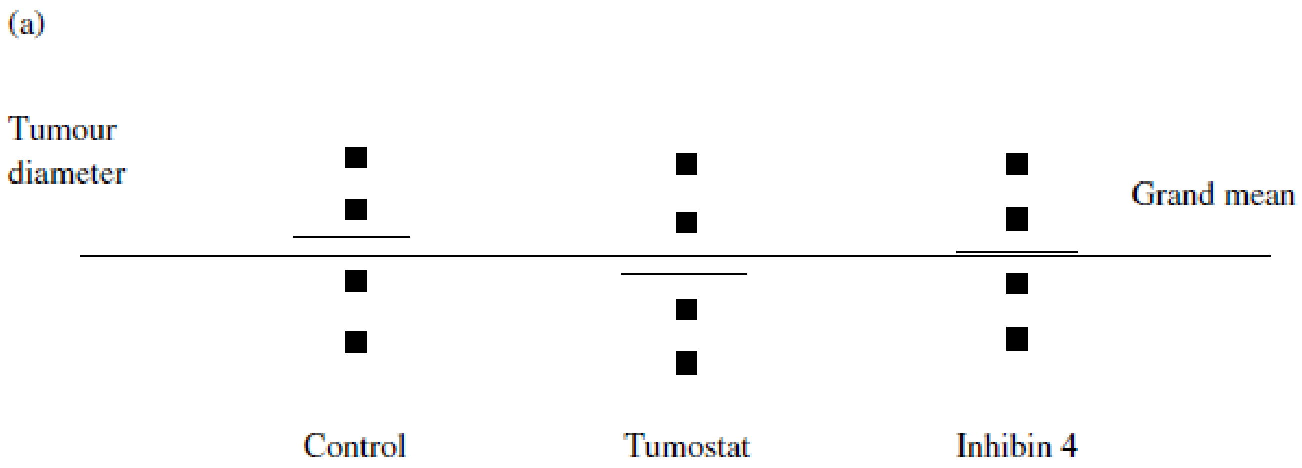
Variación debida al error

- Las observaciones se mueven con respecto a la media de su nivel de factor (variación dentro del grupo)



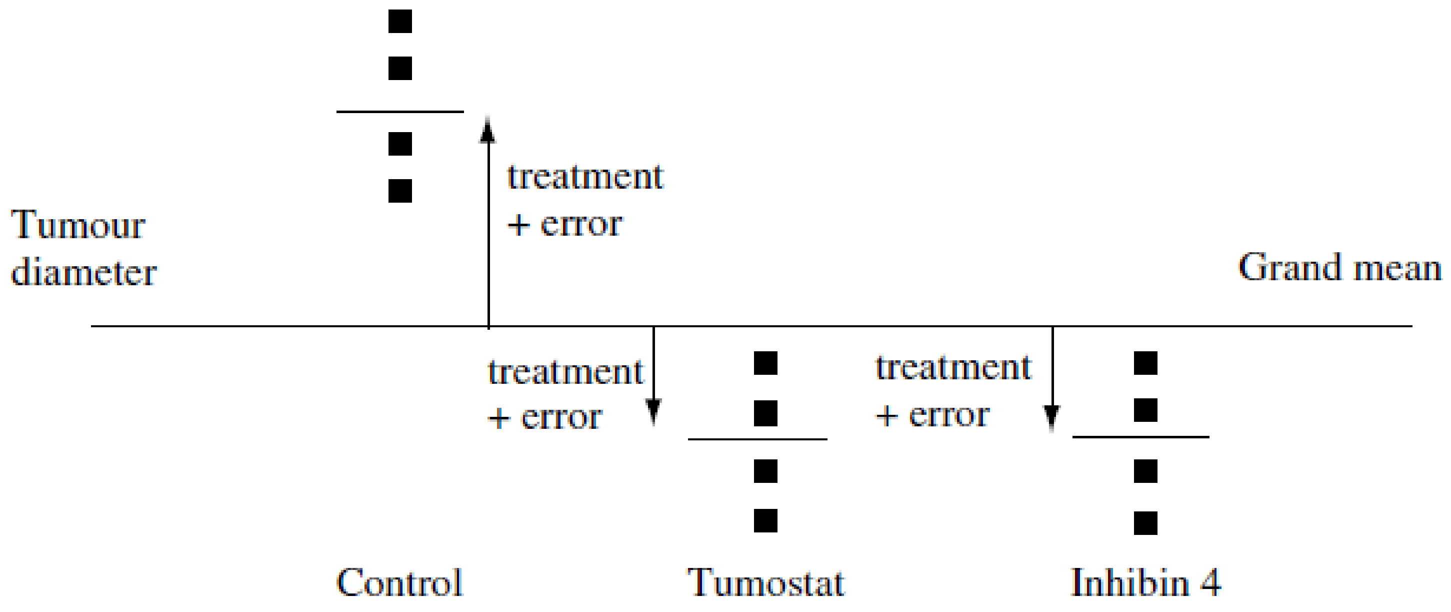
Variación sin efecto del tratamiento

- Sólo se observan las variaciones debidas al error.
- Las medias debidas a los niveles del tratamiento son cero.



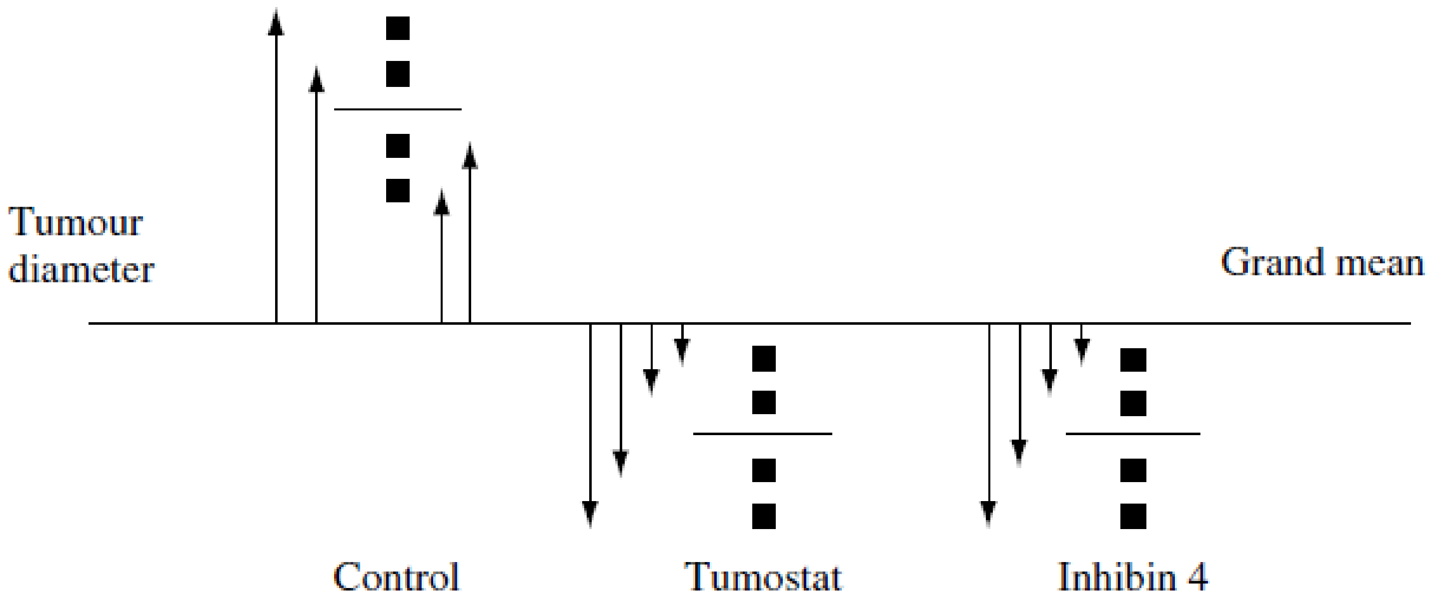
Variación de las medias

Se observa una variación en los datos debida al error sumada a la variación de las medias debida a los niveles del tratamiento (variación entre grupos)



Lo que mide ANOVA

- Lo que se mide con ANOVA es la variación de cada observación con respecto a la media Global.



$$SS_T = SS_{\text{Tratamientos}} + SS_E$$

- $SS_{\text{Tratamientos}}$: sumas de los cuadrados debida a los tratamientos (entre los tratamientos)
- SS_E: Suma de cuadrados debida al error (dentro de los tratamientos)
- Hay *an=N* observaciones en total
- Por tanto, SS_T tiene N-1 grados de libertad
- Como hay *a* tratamientos o niveles de factor, $SS_{\text{Tratamientos}}$ tiene *a*-1 grados de libertad
- Por tanto, para el error se tienen *a(n-1)=an-a=N-a* grados de libertad

- Observando el término SS_E:

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{a} \left[\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2\right]$$

Si se divide por n-1 el término entre corchetes, se obtiene la varianza del tratamiento i-ésimo

$$S_i^2 = \frac{\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2}{n-1} \qquad i = 1, 2, \ldots, a$$

- Y también, si se hace

$$\frac{(n-1)S_1^2+(n-1)S_2^2+\ldots+(n-1)S_a^2}{(n-1)+(n-1)+\ldots+(n-1)}=\frac{\sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]}{\sum_{i=1}^{a}(n-1)}$$

$$=\frac{SS_E}{(N-a)}$$

- Al dividir el término debido al error por N-*a*, se obtiene una estimación combinada de la varianza común dentro de cada uno de los a tratamientos.

- Si no hubiese diferencias entre las medias de los *a* tratamientos (el error se anula), podría usarse la variación de los promedios de los tratamientos y el gran promedio para estimar la varianza:

$$\frac{SS_{\text{Tratamientos}}}{a-1} = \frac{n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2}{a-1}$$

- Esta expresión es una estimación de la varianza si las medias de los tratamientos son iguales

El criterio definitivo…

- Se observa que la identidad de análisis de varianza nos proporciona dos estimaciones de la varianza, una basada en la variabilidad inherente dentro de los tratamientos y una basada en la variabilidad entre los tratamientos.

$$\sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$$

- **Si no hay diferencias en las medias de los tratamientos, estas dos estimaciones deberán ser muy similares, y si no lo son, se sospecha que la diferencia observada puede ser causada por diferencias en las medias de los tratamientos.**

Viéndolo de otra manera…

- Si definimos los cuadrados medios como:

$$MS_{\text{Tratamientos}} = \frac{SS_{\text{Tratamientos}}}{a-1}$$

$$MS_E = \frac{SS_E}{N-a}$$

- Y calculando el valor esperado de los cuadrados medios, se puede obtener que:

$$E(MS_E) = \sigma^2 \qquad E(MS_{\text{Tratamientos}}) = \sigma^2 + \frac{n\sum_{i=1}^{a} \tau_i^2}{a-1}$$

- Por tanto, MS_E estima la varianza
- Si no hay diferencias en las medias de los tratamientos ($\tau_i=0$), $MS_{Tratamientos}=MS_E=varianza$
- Si las medias de los tratamientos difieren, el valor esperado del cuadrado medio de los tratamientos es mayor que la varianza.
- Por tanto, **se puede realizar una prueba de hipótesis de que no hay diferencias en las medias de los tratamientos comparando $MS_{Tratamientos}$ y MS_E.**

La prueba de hipótesis: el objetivo

- Ahora se estudiará cómo puede hacerse una prueba formal de hipótesis de que no hay diferencias en las medias de los tratamientos.
- Es decir, $H_0 = \mu_1 = \mu_2 = ... = \mu_a$
- Ó lo que es equivalente, $H_0 = \tau_1 = \tau_2 = ... = \tau_a = 0$

La prueba de hipótesis: supuestos iniciales

- Se supone que los errores ε_{ij} siguen una distribución normal e independiente con media cero y varianza σ^2
- Por tanto, las observaciones y_{ij} tienen una distribución normal e independiente con media $\mu+\tau_i$ y varianza σ^2.
- En consecuencia, SS_T es una suma de cuadrados de variables aleatorias con una distribución normal
- Entonces, puede demostrarse que SS_T/σ^2 tiene una distribución ji-cuadrada con N-1 grados de libertad.
- También puede demostrarse que SS_E/σ^2 es una variable ji-cuadrada con N-a grados de libertad
- Y que $SS_{Tratamientos}/\sigma^2$ es una variable ji-cuadrada con *a*-1 grados de libertad si la hipótesis nula $H_0{:}\,\tau_i=0$ es verdadera

Independencia de las Sumas de cuadrados

- Las tres sumas de cuadrados no son necesariamente independientes (SS_T es la suma de $SS_{Tratamientos}$ y SS_E)
- Se requiere el teorema de Cochran para establecer la independencia de SS_E y $SS_{Tratamientos}$
- Teorema de Cochran:
  - Sea Z_i una variable aleatoria con distribución normal estándar. Esto se cumple para $i=1, 2, \dots, v$
  - Y si $$\sum_{i=1}^{v} Z_i^2 = Q_1 + Q_2 + \dots + Q_s$$
  - Donde $s \leq v$, y Q_i tiene v_i grados de libertad ($i=1, 2, 3, \dots, s$)
  - Entonces $Q_1, Q_2, \dots, Q_s$ son variables aleatorias ji-cuadrada independientes con $v_1, v_2, \dots, v_s$ grados de libertad respectivamente, si y solo si $v=v_1+v_2+\dots v_s$

Observaciones sobre el teorema de Cochran

- Puesto que los grados de libertad de $SS_{\text{Tratamientos}}$ y SS_E suman N-1 (el número total de grados de libertad), el teorema de Chochran implica que $SS_{\text{Tratamientos}}/\sigma^2$ y SS_E/σ^2 son variables aleatorias ji-cuadrada con distribuciones independientes.
- Por tanto, si la hipótesis nula de que no hay diferencias en las medias de los tratamientos es verdadera, el cociente

$$F_0 = \frac{SS_{\text{Tratamientos}}/(a-1)}{SS_E/(N-a)} = \frac{MS_{\text{Tratamientos}}}{MS_E}$$

- Se distribuye como F con *a-1* y *N-a* grados de libertad.
- **Esta expresión es el estadístico de la prueba para la hipótesis de que no hay diferencias en las medias de los tratamientos.**

- Por los cuadrados medios esperados se observa que, en general MS_E es un estimador insesgado de σ^2
- Bajo la hipótesis nula, $MS_{Tratamientos}$ es un estimador insesgado de σ^2.
- Si la hipótesis nula es falsa, el valor esperado de $MS_{Tratamientos}$ es mayor que σ^2.
- Por tanto, bajo la hipótesis alternativa, el valor esperado del numerador del estadístico de prueba F_o, es mayor que el valor esperado del denominador.
- En consecuencia, H_o deberá rechazarse para valores del estadístico de prueba que son muy grandes. (Esto implica una región crítica de una sola cola superior)

- Por lo tanto, H_o deberá rechazarse y concluirse que hay diferencias en las medias de los tratamientos si $F_0 > F_{\alpha, a-1, N-a}$
- Donde F_o se calcula como

$$F_0 = \frac{SS_{\text{Tratamientos}} / (a-1)}{SS_E / (N-a)} = \frac{MS_{\text{Tratamientos}}}{MS_E}$$

- Se puede usar el enfoque de los valores P para tomar la decisión.

- Para calcular los SS se pueden utilizar también las siguientes fórmulas equivalentes:

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

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

- Y calcular SS_E como: $SS_E = SS_T - SS_{\text{Tratamientos}}$

Resumen del análisis de varianza con un solo factor y efectos fijos

Tabla 3-3 Tabla de análisis de varianza para el modelo con un solo factor y efectos fijos

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0
Entre los tratamientos	$SS_{\text{Tratamientos}} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$	$a-1$	$MS_{\text{Tratamientos}}$	$F_0 = \dfrac{MS_{\text{Tratamientos}}}{MS_E}$
Error (dentro de los tratamientos)	$SS_E = SS_T - SS_{\text{Tratamientos}}$	$N-a$	MS_E	
Total	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	$N-1$		

Ejemplo: El objetivo del experimento

- Un ingeniero de desarrollo tiene interés en investigar la resistencia a la tensión de una fibra sintética nueva que se usará para hacer camisas para caballero.
- Por experiencia previa, se conoce que la resistencia a la tensión se afecta por el peso porcentual del algodón en la mezcla de materiales de la fibra.
- Se sospecha que al aumentar el contenido de algodón se incrementará la resistencia. (**Hipótesis**)
- Se sabe que el contenido de algodón deberá variar entre 10 y 40 por ciento para que el producto final tenga otras características de calidad que se desean (como la capacidad de ser sometido a un tratamiento de planchado permanente). (**Condiciones/restricciones de operación**)

Diseño del experimento

- El ingeniero decide probar ejemplares en cinco niveles de peso porcentual de algodón: 15, 20, 25, 30 y 35 por ciento.
- Se decide que se probarán 5 ejemplares de cada nivel de contenido de algodón.
- Este es un experimento con un solo factor (porcentaje de algodón) con a=5 niveles (5 porcentajes diferentes de algodón)
- También se decide probar con 5 ejemplares de cada nivel de contenido de algodón.

Resultados obtenidos

- Suponiendo que se siguió el orden aleatorio descrito anteriormente, se muestran las observaciones que se realizaron para la resistencia a la tensión en la siguiente tabla

Peso porcentual del algodón	Resistencia a la tensión observada (lb/pulg^2)					Totales y_i	Promedios $\bar{y}_i$
	1	2	3	4	5		
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
						$y_{..} = 376$	$\bar{y}_{..} = 15.04$

- Hipótesis nula

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

- Hipótesis alternativa

  H_1: algunas medias son diferentes

- Se calculan las sumas de cuadrados requeridas

$$SS_T = \sum_{i=1}^{5}\sum_{j=1}^{5} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$= (7)^2 + (7)^2 + (15)^2 + \cdots + (15)^2 + (11)^2 - \frac{(376)^2}{25} = 636.96$$

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

$$= \frac{1}{5}[(49)^2 + \cdots + (54)^2] - \frac{(376)^2}{25} = 475.76$$

$$SS_E = SS_T - SS_{\text{Tratamientos}}$$

$$= 636.96 - 475.76 = 161.20$$

Tabla 3-4 Análisis de varianza de los datos de la resistencia a la tensión

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor P
Peso porcentual del algodón	475.76	4	118.94	$F_0 = 14.76$	<0.01
Error	161.20	20	8.06		
Total	636.96	24			

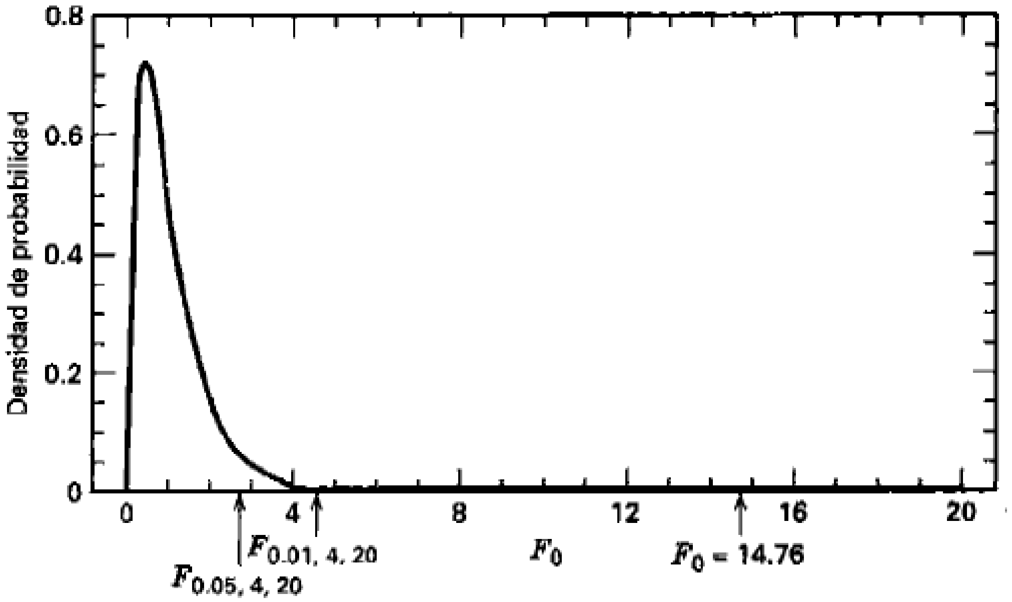
- Obsérvese que el cuadrado medio de los tratamientos (118.94) es varias veces mayor que el cuadrado medio dentro de los tratamientos (ó cuadrado medio del error = 8,06)
- **Por tanto, no es posible que las medias de los tratamientos sean iguales.**

Prueba más formal

- Calculando el coeficiente F: $F_o=118.94/8.06=14.76$
- Obtención del F de referencia con un $\alpha=0.05$ de la tabla de la distribución F: $F_{0.05,4,20}=2.87$
- Por tanto, $14.76>2.87$, se rechaza H_o y se concluye que las medias de los tratamientos difieren.
- En otras palabras, el peso porcentual del algodón en la fibra afecta de manera significativa la resistencia a la tensión media.

Valores P

- Podría calcularse un valor P para el estadístico de la prueba.
- Obsérvese que con un valor P de 0.01, $F_{0.01,4,20}=4.43$ y $F_o>4.43$
- Incluso el valor de P exacto es $P=9.11\text{x } 10^{-6}$



Propiedades de las observaciones

- Las observaciones pueden ser trasladadas mediante la suma o resta de una constante a todas, manteniendo el mismo análisis de varianza.
- Ejemplo:
  - Al restar 15 a las observaciones del ejemplo se obtienen los siguientes resultados

Tabla 3-5 Datos codificados de la resistencia a la tensión del ejemplo 3-2

Peso porcentual del algodón	Observaciones 1	2	3	4	5	Totales y_i
15	–8	–8	0	–4	–6	–26
20	–3	2	–3	3	3	2
25	–1	3	3	4	4	13
30	4	10	7	4	8	33
35	–8	–5	–4	0	–4	–21

Propiedades de las observaciones

- Y se obtienen los mismos valores para el análisis de la varianza:

$$SS_T = (-8)^2 + (-8)^2 + \cdots + (-4)^2 - \frac{(1)^2}{25} = 636.96$$

$$SS_{\text{Tratamientos}} = \frac{(-26)^2 + (2)^2 + \cdots + (-21)^2}{5} - \frac{(1)^2}{25} = 475.76$$

$$SS_E = 161.20$$

Propiedades de las observaciones

- Suponga que cada observación se multiplica por 2.
- Se obtienen los siguientes parámetros: $SS_T = 2547.84, SS_{\text{Tratamientos}} = 1903.04$
- Observe que son los parámetros de las observaciones originales multiplicados por 4 (es decir 2^2). Por ejemplo:

$$1903.04/4 = 475.76$$

- Además, el factor F se mantiene igual: $F = (1903.04/4)/(644.80/20) = 14.76$
- Por tanto, los análisis de varianza son equivalentes!!!!