DISEÑOS EXPERIMENTALES:
Teoría y práctica para experimentos agropecuarios

Julio Gabriel Ortega
Alfredo Valverde Lucio
Blanca Indacochea Ganchozo
Carlos Castro Piguave
Máximo Vera Tumbaco
José Alcívar Cobeña
Raquel Vera Velásquez

compAS
Grupo de capacitación e investigación pedagógica

UNIVERSIDAD ESTATAL DEL SUR DE MANABI
EXCELENCIA ACADÉMICA PARA EL DESARROLLO
Feb. 7 - 2001
UNESUM

DISEÑOS EXPERIMENTALES:
Teoría y práctica para experimentos agropecuarios

compAS
Grupo de capacitación e investigación pedagógica

UNIVERSIDAD ESTATAL DEL SUR DE MANABI
EXCELENCIA ACADEMICA PARA EL DESARROLLO
Feb. 7 - 2001
UNESUM

Julio Gabriel Ortega
Alfredo Valverde Lucio
Blanca Indacochea Ganchozo
Carlos Castro Piguave
Máximo Vera Tumbaco
José Alcívar Cobeña
Raquel Vera Velásquez

compAs
Grupo de capacitación e investigación pedagógica

UNIVERSIDAD ESTATAL DEL SUR DE MANABI
EXCELENCIA ACADEMICA PARA EL DESARROLLO
Feb. 7 - 2001
UNESUM

DISEÑOS EXPERIMENTALES:
Teoría y práctica para
experimentos agropecuarios

©Julio Gabriel Ortega
Alfredo Valverde Lucio
Blanca Indacochea Ganchozo
Carlos Castro Piguave
Máximo Vera Tumbaco
José Alcívar Cobeña
Raquel Vera Velásquez

Editores
Julio Gabriel Ortega
Alfredo Valverde Lucio

Segunda edición 2021,
Publicado por acuerdo con los autores.

Guayaquil-Ecuador

Editado en Guayaquil - Ecuador

ISBN : 978-9942-33-381-0

compÁs
Grupo de capacitación e investigación pedagógica

UNIVERSIDAD ESTATAL DEL SUR DE MANABI
UNESUM

Cita.
Gabriel, J., Valverde, A., Indacochea, B., Castro, C., Vera, M., Alcívar, J., Vera, R., (2021). DISEÑOS EXPERIMENTALES: Teoría y práctica para experimentos agropecuarios. Segunda edición, Editorial Grupo Compás. Universidad Estatal del Sur de Manabí. Guayaquil, Ecuador.

Jipijapa, Ecuador 2020

PARTE I

GENERALIDADES DEL METODO CIENTÍFICO Y LOS DISEÑOS EXPERIMENTALES ESTANDAR

UNIDAD 1

MÉTODO CIENTÍFICO Y CONCEPTOS FUNDAMENTALES DEL DISEÑO DE EXPERIMENTOS

Julio Gabriel Ortega

Blanca Indacochea Ganchozo

"El propósito de la ciencia estadística es suministrar una base objetiva para el análisis de problemas en los que los datos se apartan de las leyes de la causalidad exacta. Se ha ideado un sistema lógico general de razonamiento inductivo. Aplicable a datos de esta naturaleza, y en la actualidad se utiliza ampliamente en la investigación científica. Es importante comprender sus principios, tanto para los investigadores científicos como para aquellos cuyos intereses residen en la aplicación de avances tecnológicos resultantes de dichas investigaciones. Esto es especialmente cierto para las ciencias agrícolas y biológicas."

D.J. Finney

Razonamiento deductivo

El razonamiento que parte de un principio general hacia una conclusión específica es un proceso deductivo.

Razonamiento inductivo

El razonameinto inductivo llega a un pricipio general a partir de una conclusión particular .

Los experimentos son conducidos para sumunistrar hechos específicos a partir de los cuales se establecen las conclusiones generales o principios contemplando el razonamiento inductivo.

Importancia de la variabilidad

La variabilidad es una característica del material biológico y plantea el problema de decidir si las diferencias entre unidades experiemntales se deben a la variabilidad no ponderada o a los efectos reales del tratamiento. La ciencia estadística ayuda a superar esta dificultad, requiriendo el acopio de datos para suministrar estimaciones imparciales de los efectos de tratamiento y para evaluar las diferencias de tratamiento mediante pruebas de significación basados en mediciones de la variabilidad no ponderada.

Principios del diseño experimental

Repetición

Selección aleatoria

Control local

El método científico

Contempla un flujo desde los hechos observados hacia la hipótesis para la experimentación, la cual suministra más hechos que anularán, ampliarán o alterarán la hipótesis.

Un diseño bién concebido y diseñado deberá ser lo mas simple posible, tener grandes posibilidades de alcanzar su objetivoy evitar los errores tendenciosos y sistemáticos. Sus conclusiones deberán poseer un ampli rango de validez, y los datos recabados a partir del mismo deben estar sujetos al análisis a través de procedimientos estadísticos válidos.

Procedimiento para experimentación

Contempla los siguientes pasos:

1. Planificación

Definir el problema

El desarrollo agroindustrial, el descubrimiento de nuevas tecnologías así como las exigencias alimenticias de una población mundial en constante crecimiento, obligan a los profesionales agropecuario a realizar investigaciones que contribuyan a solucionar sus problemas, destacando de manera particular los inherentes a la baja productividad. Sin embargo el investigador debe definir la magnitud del problema puesto que un inadecuado análisis provocaría como consecuencia resultados insipientes y de limitada transcendencia.

Establecer los objetivos

Una vez identificado el problema se debe definir los objetivos, tanto el general como los específicos. Definir bien los objetivos es importante pues su adecuada identificación aportaría a la solución de los problemas. Es oportuno indicar que los objetivos específicos deben formularse en función a las varibles dependientes e independientes que se establecen como base de la investigación.

Determinar las variables a medir

Es fundamental establecer las variables de respuestas así como la estrategia de mediación, la razón; garantizar la confiabilidad de los resultados, por lo cual se recomienda utilizar los instrumentos de mediación apropiados y que certifiquen la fidelidad de los resultados. La toma inadecuada de datos afecta los resultados y por tanto sesga la decisión aumentando drásticamente el error experimental.

Establecer y seleccionar los tratamientos, niveles o factores de estudios

Los factores a investigarse deben ser definidos en función a los resultados que se esperan alcanzar más que por el criterio del investigador, además se debe considerar tomar los factores de mayor efecto. otro aspecto a considerar es sin lugar a dudas el económico, puesto que a mayor tratamientos o factores, mayor será el número de unidades experimentales.

Es oportuno establecer la diferencia entre tratamientos, niveles y factores:

- Factor es la variable independiente o de investigación a estudiar, un factor puede ser una época, una variedad de semilla, hormonas, etc.
- Los niveles son las dosis, tiempos, pesos u otros parámetros que impliquen la tomas de datos, expresan al factor.
- Tratamientos es la interacción de los niveles y factores (ejemplo épocas y variedades de semillas), destacando que en los diseños de un solo factor cada nivel es un tratamiento.

El material experimental

En lo que respecta a material experimental, podríamos citar a nivel agropecuario como a la parcela de tierra, la especie animal de interés zootécnico, la caja de Petri si el experimento es a nivel de laboratorio, el semillero, etc.

El material experimental debe ser cuidadosamente seleccionado a fin de garantizar la calidad de los resultados.

El número de repeticiones

El número de repeticiones que se realicen en una investigación es importante, pues estadistícamente a mayor repeticiones menor será el error experimental.

Podríamos definir a las repeticiones como un reprís de los tratamientos, sin embargo es importante tener claro el número de repeticiones, recomendando entre 4 a 5 repeticiones para los diseños simples. Otro aspecto importante es la aleatorización de las repeticiones, pues tal acción permite acercar a la realidad la investigación

Ejemplo:

Diseño experimental

DISEÑO BIFACTORIAL (Tabla 1.1)

FACTOR A (Bioestimulantes)

A1. Bioestimulante 1

A2. Bioestimulante 2

A3. Bioestimulante 3

FACTOR B (Intervalos de aplicación)

B1. Aplicación cada 8 días

B2. Aplicación cada 12 dias

Tratamientos en estudio

Tabla 1.1. Tramientos aplicados en un diseño bofactorial**.**

N°	Nomenclatura	Factor A	Factor B
1	A1 X B1	A1. Bioestimulante 1	B1. Aplicación cada 8 días
2	A1 X B2	A1. Bioestimulante 1	B2. Aplicación cada 12 dias
3	A2 X B1	A2. Bioestimulante 2	B1. Aplicación cada 8 días
4	A2 X B2	A2. Bioestimulante 2	B2. Aplicación cada 12 dias
5	A3 X B1	A3. Bioestimulante 3	B1. Aplicación cada 8 días
6	A3 X B2	A3. Bioestimulante 3	B2. Aplicación cada 12 dias

2. Ejecución

Controlar los efectos entre unidades adyacentes o vecinas

Este tema nos conlleva a considerar los efectos de borde, que particularmente se dan en las investigaciones agrícolas, donde las plantas que se encuentran en los bordes suelen tener ventajas o desventajas con respecto a las plantas que se encuentran en el centro, por tanto se recomienda tomar datos de las plantas centrales, a fin de évitar el sesgar los resultados y por ende la toma de decisiones.

Muestreo

Es oportuno indicar que previa al tipo de muestreo se debe definir el tamaño de la muestra, sin embargo se debe considerar que según la investigación y su diseño experimental, este ejercicio solo sería aplicable cuando las unidades experimentales exijan un número considerable de plantas o animales.

Los tipos de muestro son: Probabilísticos y no probabilísticos.

Probabilísticos: Aleatorio simples, aleatorio estratificado, aleatorio conglomerado, Aleatorio sistemático.

No Probabilisticos: Por cuotas o por conveniencias.

Los muestreos más recomendados a nivel de investigaciones científicas son los aleatorios, la utilización de uno de los tipos de muestreo garantiza que el investigador no sesgue los resultados.

Recabar datos y analizar

La toma de datos es uno de los aspectos de mayor importancia dentro del desarrollo de una investigación, por lo que es importante que se considere dentro de la planificación el optar por equipos y herramientas que garanticen la confiabilidad de los datos tomados.

Tomados los datos de manera correcta, permitirá al investigador realizar un análisis confiable del experimento. Cabe destacar que el análisis se lo puede realizar de manera manual, en un documento Excel y hasta en programas o software computacionales creados para tales fines.

Los resultados que se alcancen gracias a la realización del análisis de los datos, permitirá tomar decisiones sobre el o los productos agropecuarios investigados.

Tamaño de la parcela (Figuras 1.1 y 1.2)

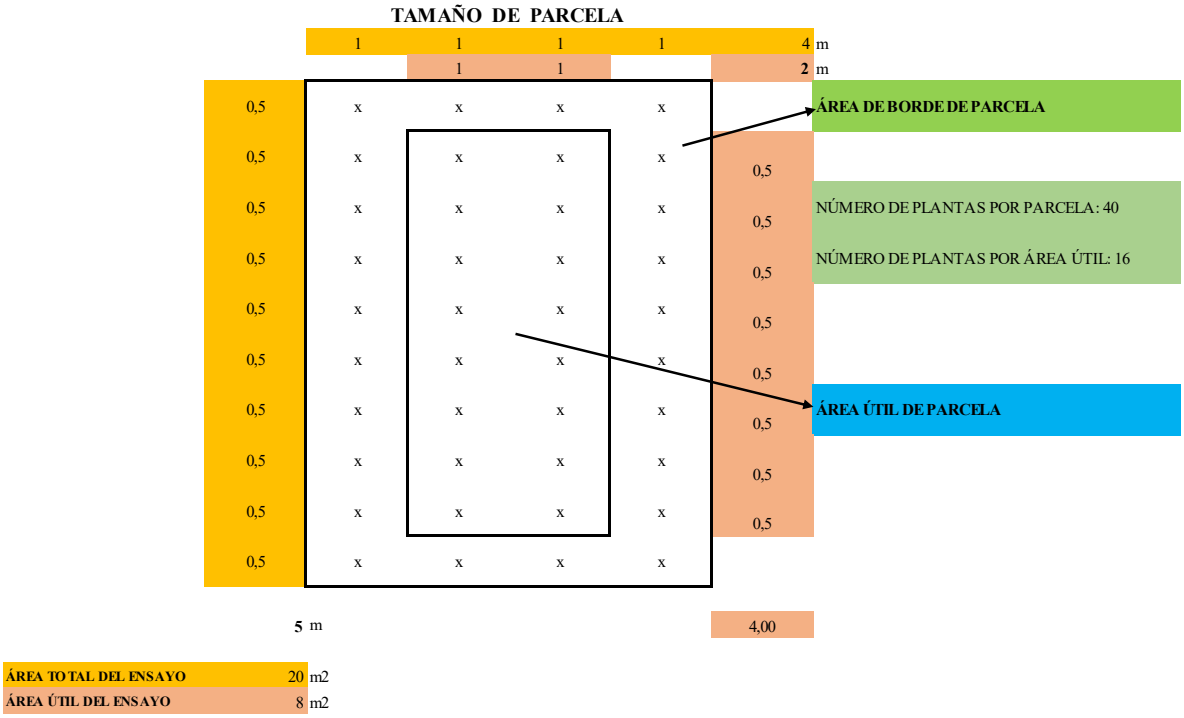


Figura 1.1. Croquis de campo del ensayo.

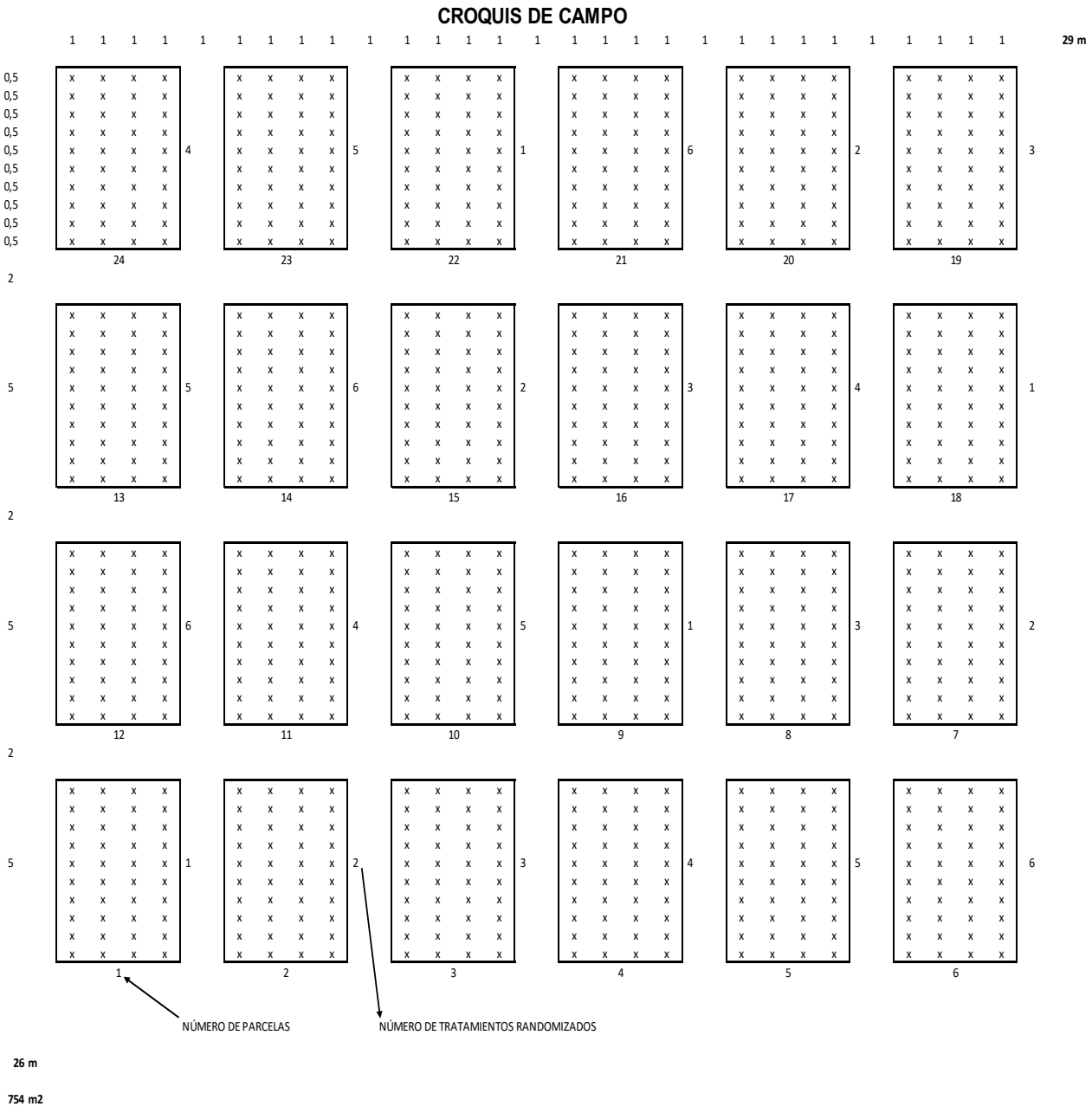


Figura 1.2. Distribucón de ls bloque y tratamientos en campo.

3. Resultados

4. Conclusiones

Las conclusiones de una investigación representan el resumen del trabajo realizado, específicamente de los resultados alcanzados y su aporte a la ciencia. Las conclusiones se deben hacer a partir de los objetivos planteados.

Interpretar e informar los resultados

Implica la aplicación del ejercicio experimental, y es mediante este procedimiento que se comprueba o no la hipótesis, la significación o no de los tratamientos, niveles e interacción entre

factores, el error experimental entre lo más destacado, datos que son importantes para la toma de decisión.

Difundir

Es el acto culminante de la investigación y su labor consiste en redactar técnicamente los resultados y difundirlos en revistas con cobertura nacional e internacional a manera de artículos científicos, de esta manera una investigación contribuye a la ciencia mundial.

Definiciones importantes

Definición de diseño experimental

Se define como el método, procedimiento o conjunto de reglas para asignar los tratamientos aleatoriamente a las unidades experimentales.

En campo es el arreglo geométrico de un experimento.

Clases de experimentos

1. Experimentos exploratorios, se someten a prueba un gran número de tratamientos de los cuales se conoce poco. No se hace énfasis en la precisión.
2. Experimentos analíticos, en la que se hace énfasis en averiguar el cómo y el por qué de las cosas, interesa averiguar qué pasa. El número de repeticiones y el control son mayores.
3. Experimentos de validación, sirven para comprobar si las inferencias que se hicieron, resultantes del experimento analítico, son aplicables en un ambiente determinado.

Error experimental

- Inherente a la naturaleza del material experimental (introduce heterogeneidad).
- Falta de uniformidad en el manejo de la UE.
- Utilizar el diseño más simple. De los más simple a lo más complejo, buscando precisión y menor costo posible.

Uso de información adicional.

- **Análisis de varianza**: Variación se reparte entre diferentes fuentes o causas.
- **Análisis de covarianza:** Aumenta la precisión (o información) del experimento.

Manejo del Experimento

- Definir con precisión las variables y asegurarse de lograr la mayor exactitud y precisión posible.

Responderse las preguntas tales como:

- Qué tamaño de experimento debe hacerse
- Cuántos tratamientos
- Que tamaño de parcela,
- Cuántas repeticiones.

Estas preguntas serán contestadas una vez que hayamos definido los siguientes aspectos:

- El objetivo del trabajo.
- Disponibilidad de recursos.
- Precisión que quiere lograrse.
- Costo de insumos y mano de obra.
- Variabilidad del material experimental y sustrato.

Número de tratamientos

Dependerá del tipo de experimento y de la magnitud y variabilidad de la región de exploración que quiera estudiar. La selección de los tratamientos tiene importancia decisiva en la precisión de un experimento.

Unidad experimental

La unidad experimental es una unidad material a la cual se aplica un solo tratamiento dentro de una repetición o bloque, puede ser parcela de terreno, animal, una hoja de vegetales, un árbol, una maceta, un lote de semilla etc.

Tamaño de la Unidad Experimental (UE)

Depende la precisión y costo. Mayor precisión con mayor número de repeticiones y menor tamaño de UE. Pero el costo incrementa.

El tamaño de la parcela debe ser el adecuado para conseguir los objetivos del ensayo, el tamaño de las unidades experimentales debe satisfacer los requerimientos de la investigación. Lo deseable debe estar en equilibrio con lo posible. Es importante indicar que el número de repeticiones debe ser el mismo, ya sea con parcelas pequeñas o grandes. Hay una tendencia a creer en que grandes unidades experimentales hacen mejores ensayos y que por lo tanto se necesitaran menos repeticiones. Esto no es así parcelas más grandes aumentan el costo de los ensayos y también la probabilidad de un mayor error experimental debido a la heterogeneidad dentro de las parcelas. En general, el tamaño de parcela estará determinado por la cantidad de terreno disponible para el ensayo y por la cantidad de mano de obra o insumos disponibles para desarrollar el ensayo.

Forma

No existe mucha información. En experimentos agronómicos se prefieren parcelas largas y angostas en dirección al gradiente de fertilidad o humedad del suelo.

Se debe mencionar que es importante tener el mayor detalle posible de las características del experimento, como se indica en la Tabla 1.2.

Tabla 1.2. Características de la parcela experimental.

No.	Delineamiento experimental	Características
1	Unidades o parcelas experimentales	24
2	Número de repeticiones	4
3	Número de tratamientos	6
4	Hileras por parcela	4
5	Hileras útiles	2
6	Hileras borde por parcela	2
7	Número de plantas por unidad experimental	40
8	Número de plantas por parcela útil	18
9	Número de plantas evaluadas en parcela útil	10
10	Distancia entre hileras	1 m
11	Distancia entre plantas	0,50 m
12	Distancia entre repeticiones	2 m
13	Longitud de parcela	5 m
14	Ancho de parcela	4 m
15	Área total de la parcela	20 m^2 (5mx4m)
16	Área útil de la parcela	8 m^2 (4mx2m)
17	Área útil del ensayo	192 m^2 (8m2x24)
18	Área total del ensayo	754m^2 (29mx26m)

Guía para elaborar y ejecutar proyectos de investigación

I.- Titulo
II.- Antecedentes
III.- El Problema
IV.- Justificación
V.- Objetivos
    5.1.- Objetivo general
    5.2.- Objetivo específicos
VI.- Hipòtesis
VII.- Marco teórico
VIII.- Materiales y métodos
    A.- Materiales
    B.- Métodos
        1.- Ubicación
        2.- Factores en estudio
        3.- Tratamientos
        4.- Diseño experimental
        5.- Características de las unidades experimentales
        6.- Análisis estadístico
        7.- Variables en estudio
        8.- Manejo del experimento
IX.- Cronograma de actividades
X.- Presupuesto
XI. Bibliografía

UNIDAD 2

DISEÑOS EXPERIMENTALES ESTANDAR MAS UTILIZADOS

Julio Gabriel Ortega

Alfredo Valverde Lucio

Máximo Vera Tumbaco

Definiciones

Las modalidades o niveles que afectan un factor en estudio se llama tratamiento. Por ejemplo cultivares (o variedades), dosis de nitrógeno, raciones de alimentos, etc. Se llama bloque a un conjunto de unidades experimentales más o menos homogéneas con el propósito de eliminar fuentes de variación sistemática (Figura 2.1).

Pendiente (+) (-)

R1t1	R1t2	R1t3	R1t4	R1t5	R1t6	R1t7	R1t8	R1t9	Bloque 1
R2t9	R2t3	R2t7	R2t2	R2t8	R2t6	R2t5	R2t1	R2t4	Bloque 2
R3t3	R3t1	R4t5	R8t8	R7t7	R3t9	R3t4	R5t6	R3t2	Bloque 3

Figura 2.1. Bloqueo de una fuente de variabilidad (pendiente), utilizando un diseño de bloques completamente aleatorizados (DBCA).

Aleatorización

Asignación de los tratamientos sin que intervenga la voluntad del investigador

Características de los diseños estándar

1. Cada tratamiento se repite el mismo número de veces
2. Todos los bloques del experimento son del mismo tamaño
3. Un tratamiento puede o no ocurrir en un bloque particular, pero aun cuando ocurre sólo se observa una vez
4. Los tratamientos deben aplicarse en forma aleatoria

Clasificación de los diseños estándar

1. **Diseños completamente aleatorios (DCA)**
   - Ausencia de bloques.
   - Cada unidad experimental tiene la misma probabilidad de recibir cualquier tratamiento, el cual se repite en dos o más ocasiones.
   - Son apropiados para los casos en el que el material experimental es completamente homogéneo.
2. **Diseños de bloques completamente aleatorizados (DBCA)**
   - Las unidades experimentales se agrupan en dos o más bloques completos.
   - En cada unidad experimental se alojan una vez los tratamientos.

- Son apropiados en casos donde se observa una cierta tendencia de variación en el material experimental.

3. **Diseños aleatorios de bloques incompletos (DBI)**
   - No todos los tratamientos se encuentran representados en cada bloque.
   - Son apropiados en casos donde ensayan muchos tratamientos.
   - Se agrupan en bloque más pequeños.

UNIDAD 3

NORMALIDAD, HOMOGENEIDAD DE VARIANZAS Y TRANSFORMACIÓN DE DATOS

Julio Gabriel Ortega

Alfredo Valverde Lucio

Carlos Castro Piguave

Caractacterísticas

Un investigador que se conforma con aprender las ‘’recetas’’ para llevar a cabo un análisis de varianza, sin buscar el dominio y la comprension de los principales inherentes al mismo, puede encontrarse con serios problemas. Sea que lo comprenda o no, el investigador hará ciertas suposiciones, dicho análisis puede dar lugar a que el investigador lleguen a conclusiones que no tiene justificación. Así mismo el investigador puede descubrir conclusiones importantes que se alcanzarian si los datos fueran analizados adecuadamente.

Supuestos del análisis de varianza

Los supuestos sobre lo que se basa un análisis de varianza son, en resumen:

1. Los términos de error son aleatoria, independiente y normalmente distribuidos.

2. Las varianzas de las diferentes muestras son homogéneas.

3. Las varianzas y las medias de las distintas muestras no estan correlacionadas.

4. Los efectos principales son aditivos.

Normalidad

Existen pruebas de normalidad, que resulta bastante útil aplicarlas, a menos que el número de muestras con las que estamos trabajando sea definitivamente grande. La independencia implica que no hay relacion entre el tamaño de los términos de error y la agrupación experimental a la cúal pertenecen. Puesto quie las parcelas adyacentes de un campo tienden a estar más estrechemante relacionadas entre si que las parcelas aleatoriamente distribuidas, resulta importane evitar que todas las parcelas que reciben el mismo tratamiento ocupen posiciones adyacentes en el campo. Esta es una de las principles razones para insistir en no dividir en subparcelas una parcela que recibe cierto tratamiento y referirnos a las mismas como repeticiones. La mayor seguridad contra violaciones evidentes del primer supuesto del análisis de varianza consiste en llevar acabo la distribución aleatoria adecuada al diseño experimental en particular, que estamos utilizando.

Homogeneidad de Varianzas

Si las varianzas dentro de tratamientos diferentes fuesen de hecho distintas, no tendriamos justificación para combinarlas. Supongamos, por ejemplo, que las repeticiones de dos de los tratamientos fueron en realidad muestras de poblaciones con grandes varianzas, mientras que aquellas de los dos otros tratamientos se obtuvieron de poblaciones con varianzas mucho menores (Tabla 3.1).

Resulta obvio que la diferencia requerida para la significación seria mayor para los dos tratamientos con más alta variabilidad que patra los dos de menor variabilidad. Promediar la varianzas mayores y menores podrian arrojar resultados engañosos. La diferencia entre dos tratamientos con varianzas grandes puede ser considerada significativa, cuando en realidad esta pudo haber ocurrido facilmente, por casualidad. Por otro lado la diferencia entre tratamientos cada uno repetido cinco veces, ilustraron esta situación (Tabla 3.1).

Tabla 3.1. Experimento completamente aleatorio con cuatro tratamientos y cinco repeticiones (no bloques).

Tratamiento	**Repetición**					**Total**	**Media**	**S^2**
	1	2	3	4	5			
A	3	1	5	4	2	15	3	2.5
B	6	8	7	4	5	30	6	2.5
C	12	6	9	3	15	465	9	22.5
D	20	14	11	17	8	70	14	22.5

Realizando el análisis de varianza en forma habitual, obtenemos (Tabla 3.2):

Tabla 3.2. Análisis de varianza para el experimento mencionado.

FV	**gl**	**SC**	**CM**	**F**
Tratamientos	3	330	110	8.8 ** a
Error	16	200	12.5	

*: Significativo al p<0.05 de probabilidad, **: Altamente significativo al p<0.01 de probabilidad.

Nóteseque en la tabla 3.2, que el cuadrado medio del error es el promedio de las cuatro promedios de las varianzas individuales dentro de los tratamientos. El valor F calculado es altamente significativo.

Calculemos ahora una la DMS:

$$DSM_{0.5} = t\sqrt{2ESM / r} = 2.12\sqrt{5} = 4.74$$

Puesto que la diferencia de medias entre los tratamientos A y B es de solamente de 3, podriamops concluir que esta no es significativa. La diferencia entre las medias de C y D es igual a 5; por tanto; esta podrìa denominarse significativa en el nivel de 5%; sin embargo, notamos que las varianzas de C y D son 9 veces mayores que las A y B (Tabla 3.3). **Tabla 3.3**. Análisis de varianza para A y B.

FV	**gl**	**SC**	**CM**	**F**
Tratamientos	1	22.5	22.5	9 **
Error	8	20.0	2.5	

*: Significativo al p<0.05 de probabilidad, **: Altamente significativo al p<0.01 de probabilidad.

El supuesto de que las varianzas son homogéneas es puesto en seria duda. Por tanto resultaría más razonable analizar A y B por separado de C y D (Tabla 3.4).

Tabla 3.4. Análisis de varianza para C y D.

FV	**gl**	**SC**	**CM**	**F**
Tratamientos	1	62.5	62.5	2.78 ns
Error	8	180.	22.5	

*: Significativo al p<0.05 de probabilidad, **: Altamente significativo al p<0.01 de probabilidad.

Esto nos conduce precisamente a las conclusiones opuestas respecto de las diferencias entre A y B y entre C y D. Posteriormente veremos cómo probar los datos para homocedasticidad u homogéneidad de varianzas, y lo que podemos hacer cuando que las varianzas de los datos no son homogeneas. Hay diversos procesos que podemos seguir. **Primero** podemos separar los datos en grupos, de modo que dentro de cada grupo las varianzas sean homogéneas. Luego cada grupo puede analizarse por separado, como lo hicimos en el ejemplo anterior. **Segundo** podemos utilizar métodos más avanzados de estadistica, el cual contempla procedimientos bastante complicado para ponderar la medias de acuerdo a sus varianzas. **Tercero** podemos tranformar lo datos en forma tal que estos sean homogéneos. Este último método es el que analizamos mas ampliamente en el presente artículo.

Independencia de medias y varianzas

En algunos datos existe una relación definida entre las medias de las muestras y sus varianzas. Este es un caso especial y la causa más común de heterogeneidad de varianza. Una correlación positiva entre medias y varianzas suele encontrarse cuando existe un amplio rango de medias de la muestra.

Supongamos, por ejemplo, que estuvimos probando los efectos de diversos insecticidas sobre el pulgón, y que medimos su eficacia mediante el conteo del número de pulgones por hoja, después de la aplicación. Si las medias de dos tratamientos bastante ineficaces fueran 305 y 315, naturalmente vacilaríamos en otorgar demasiada importancia a esta diferencia. Por otro lado, si las medias de otros dos tratamientos fuesen 5 y 15, tenderíamos a pensar que esta diferencia fue apreciable, impresionados por el hecho de que una de ellas fue tres veces mayor que la otra. Bajo el supuesto de que las varianzas son homogéneas y no relacionadas con las medias, tendríamos que otorgar tanta importancia a la diferencia entre 305 y 315 como a la diferencia entre 5 y 15, puesto que las diferencias reales son las mismas en ambos casos. Probablemente enfrentaríamos un sentimiento de intranquilidad de que algo estuvo mal. El examen de las diversas muestras revelaría casi seguramente que en general las muestras con medias grandes presentarían también grandes varianzas y que aquellas con medias pequeñas presentarían varianzas reducidas. Entonces el supuesto de que las medias y las varianzas no están correlacionadas resultaría falso, y un análisis de varianza ordinario de los datos en bruto no tendría validez (sería erróneo hacerla).

Tomemos un ejemplo más extremo. Un investigador desea probar el efecto de una nueva vitamina sobre el peso de algunos animales. Es su deseo de incluir un amplio rango de animales en sus pruebas, de modo que elige ratones, gallinas y ovejas. El sentido común diría que una diferencia de media libra en los pesos medios de dos lotes de ovejas se consideraría despreciable y se atribuiría fácilmente a la casualidad. Una diferencia de media libra en los pesos medios de dos lotes de gallinas se consideraría muy grande, pero no más allá de las posibilidades. Una diferencia de media libra en los pesos medios de dos lotes de ratones, se consideraría algo absolutamente fantástico. Este es un ejemplo reconocidamente extremo y casi absurdo, pero sirve para enfatizar el punto de que el supuesto de independencia de varianzas y medias no debe aceptarse ciegamente. Debemos examinar los datos y, si es necesario, probar la validez del supuesto, antes de proceder con el análisis de varianza.

Otros tipos de datos, que frecuentemente muestran una relación entre varianzas y medias, son aquellos basados en conteos y los que consisten en proporciones de porcentajes. Ahora, supóngase que encontramos que existe una relación entre varianzas y medias; ¿significa esto que estamos forzados a abandonar el análisis de varianza como método para analizar los datos?

Afortunadamente, esto no sucede con frecuencia. A menudo podemos transformar los datos en forma tal que el supuesto de independencia entre varianzas y medias sean válidas. Luego podemos proceder con el análisis de varianza de los datos trasformados.

Aditividad

Para cada diseño experimental existe un modelo matemático denominado **modelo lineal aditivo**. Para un diseño completamente aleatorio, este modelo es: $X_i = \mu + t_i + e$, que expresa que el valor de cualquier unidad experimental está compuesto por la media general más el efecto de tratamiento, más un término de error. El modelo correspondiente para un diseño de bloques completamente aleatorios es: $X_{ij} = \mu + t_¡ + b_j + e_{ij}$; que expresa que cualquier unidad experimental está compuesta por la media general más un efecto de tratamiento, el efecto de bloque, más un término de error. El aspecto importante, que debe notarse en estos modelos, es que los **términos se suman**; de ahí el término aditividad.

El modelo para un diseño en bloques completamente aleatorios, por ejemplo, implica que un efecto de tratamiento es el mismo para todos los bloques y que el efecto de bloque es el mismo para todos los tratamientos. En otras palabras, si se encuentra que un tratamiento incrementa la producción en cierta cantidad promedio por encima de la media general, suponemos que éste tiene el mismo efecto en los bloques de alta producción que en los bloques de baja producción.

Podemos concebir diversas situaciones en las que este supuesto no sería correcto; por ejemplo, en un experimento para probar el efecto de nitrógeno (N) sobre la producción, algunos bloques pueden producir menos que otros, debido a un bajo nivel natural de N en el suelo. Cabe esperar que las parcelas de dichos bloques se beneficien más con la adición de N que las parcelas de bloques, en que la reserva natural de N ya adecuada. Por otro lado, supóngase que la baja producción se debió a una reserva inadecuada de humedad. Entonces, cabe esperar que el suministro de N tenga resultados poco halagüeños en estos bloques de baja producción, pero que origine un notable incremento de la producción en los bloques donde hubo suficiente agua. Otra situación puede ser aquella en la cual el efecto de un tratamiento es incrementar la producción en cierto porcentaje o proporción. Esto recibe el nombre de efecto multiplicativo de tratamiento.

En cualquiera de los casos anteriores, el supuesto de aditividad sería incorrecto; este hecho debe reconocerse en el análisis de los datos. En el caso de los efectos multiplicativos de tratamiento, se requiere nuevamente transformaciones que cambiarán los datos para ajustarlos al modelo aditivo.

Pruebas para las violaciones de los supuestos

Ahora estamos en condiciones de brindar algunos ejemplos específicos de datos que no cumplen uno o más de los supuestos del análisis de varianza. Mostraremos cómo probar dichos supuestos y las formas en que pueden transformarse los datos, de modo que resulten adecuados. A continuación, presentamos algunos datos hipotéticos que pueden obtenerse en un experimento como los estudiados en párrafos anteriores, sobre los efectos de una nueva vitamina en ratones, gallinas y ovejas (Tabla 3.5):

Tabla 3.5. Pesos en libras, de animales tratados con vitaminas y de control, en un experimento de bloques completamente aleatorios.

Especies - tratamiento	Bloque					Media
	I	II	III	IV	Total	
Ratones-control	0.18	0.30	0.28	0.44	1.2	0.3
Ratones-vitamina	0.32	0.40	0.42	0.46	1.6	0.4
Subtotales	0.50	0.70	0.70	0.90	2.8	0.35
Gallinas-control	2.0	3.0	1.8	2.8	9.6	2.40
Gallinas –vitaminas	2.5	3.3	2.5	3.3	11.6	2.90
Subtotales	4.5	6.3	4.3	6.1	21.2	2.65
Ovejas-control	108.0	140.0	135.0	165.0	548.0	137.0
Ovejas-vitamina	127.0	153.0	178.0	176.0	604.0	151.0
Totales principales	240.0	300.0	288.0	348.0	1176.0	49.0

Examinando los datos dan como resultado el siguiente análisis de varianza (Tabla 3.6):

Tabla 3.6. Análisis de varianza de animales tratados con vitaminas y de control.

FV	gl	SC	CM	F
Bloques	3	984.0	328.0	2.63
Especies	2	108321.16	54160.58	434.51**
Vitaminas	1	142.11	142.11	1.14
Especies x vitaminas	2	250.41	125.20	1.00
Error	15	1869.72	124.65	

*: Significativo al $p<0.05$ de probabilidad, **: Altamente significativo al $p<0.01$ de probabilidad.

La diferencia altamente significativa al $p<0.01$ de probabilidad entre especies no debe sorprendernos en lo más mínimo. **Puede parecer muy extraño** que no hayamos encontrado una diferencia significativa debida a las vitaminas, especialmente porque cada animal que recibió la vitamina mostró en cada repetición un mayor peso que el animal de control correspondiente. **Parece también extraño** que no encontremos evidencias de interacción entre los efectos de la vitamina y las especies, puesto que la respuesta aparente a las vitaminas es sumamente diferente en las distintas especies.

Si aceptamos este análisis en su valor nominal, tendríamos que concluir que el **experimento fue virtualmente un fracaso total**. Al parecer, todo lo que aprendimos con este experimento fue que los ratones, las gallinas y las ovejas difieren de peso. Incluso si separáramos aquí los efectos de especies en dos comparaciones, una comparando a las ovejas con las gallinas y los ratones, y la otra comparando a las gallinas con los ratones, encontramos que no podríamos siquiera mostrar una diferencia significativa entre gallinas y ratones. Fijémonos en los datos con los supuestos del análisis de varianza en mente, y veamos qué puede hacerse si algunos de los mismos resultan falsos. Primero, podemos fijarnos en los términos de error para comprobar si los mismos están aleatoriamente independiente y normalmente distribuidos. Para hacerlo, removeremos la media general, los efectos de tratamiento y los efectos de bloque de cada celda. Esto da la siguiente tabla de términos de error (Tabla 3.7).

Tabla 3.7. Componentes del error en un experimento con vitaminas.

Especies—tratamiento	Bloques				Total
	I	I	111	IV	
Ratones- control	8.88	-1.00	0.98	-8.86	0
Ratones-vitamina	8.92	-1.00	1.02	-8.94	0
Gallinas-control	8.60	- 0.40	0.40	-8.60	0
Gallinas-vitamina	8.60	- 0.60	0.60	-8.60	0
Ovejas-control	-0.00	2.00	-1.00	19.00	0
Ovejas-vitamina	-15.00	1.00	-2.00	16.00	0
Totales	0.00	0.00	0.00	0.00	

Estos términos de error no parecen estar aleatoriamente distribuidos. En apariencia, éstos no son independientes, puesto que en cada bloque los términos de error para los dos miembros de cada especie están estrechamente relacionados. Por último, parece que su distribución se desviara considerablemente de la normal, puesto que existen dos clases de modelo, uno entre 8.5 y 9.0, y el otro entre -8.5 y -9.0. El primer supuesto del análisis de varianza no logró sostener muy bien bajo una inspección minuciosa.

A continuación, examinaremos el supuesto de homogeneidad de varianzas. Para esto, necesitamos aprender una prueba conocida como la **prueba de Bartlett.**

Primero, necesitamos calcular la varianza entre las cuatro repeticiones de cada combinación de tratamiento. Para los controles del ratón será:

$[-18^2 + .30^2 + .28^2 + 4.4^2\text{-}(1.2\ ^2\ 4)$ número de repeticiones - 1)] = 0.0115.

Después de calcular cada una de las varianzas, los valores obtenidos se agrupan en una tabla como la que se presenta a continuación (Tabla 3.8):

Tabla 3.8. Varianza y sus logaritmos para grupos en un experimento con vitaminas.

Tratamiento	gl	S_i^2	codificado S_i^2	logaritmos de S_i^2 codificado
Ratón — control	3	0.0115	11.5	1.06
Ratón—vitamina	3	0.0035	3.5	0.54
Gallinas—control	3	0.3407	346.7	2.54
Gallinas-vitamina	3	0.2133	213.3	2.33
Oveja-control	3	546.0	546000.0	5.74
Oveja-vitamina	3	425.3	425 300.0	5.63
Totales	18	971	875.0	17.84
Media				161979.0
Logaritmo de la media				5.209

El propósito de codificar las varianzas es evitar los logaritmos negativos. Podemos multiplicar las varianzas por una constante cualquiera, sin alterar la prueba. Resulta deseable hacer todos los valores codificados iguales o mayores que 1, de modo que realizamos nuestra codificación multiplicando cada 2 por 1000. Resulta más sencillo utilizar logaritmos comunes; dos dígitos en la mantisa suelen ser suficientes. La media de las varianzas codificadas se encuentra al dividir sus totales entre el número de muestras, y el log de esta media se incluye en la tabla. Ahora estamos listos para calcular lo que se denomina **ji cuadrada** (X^2) no ajustada, a partir de la fórmula:

y2 = 2.3026 [(S gl x log de la media) — (gl por muestra x 2 logs)]
= 2.3026 [(18 x 5.209) - (3 x 17.84)1
= 92.66.

El factor 2.3026 es el factor para convertir logaritmos comunes a logaritmos naturales.

El valor de ji cuadrada que hemos calculado requiere un ajuste, y para hacerlo necesitamos:

$$C = 1 + \frac{1}{3(\textit{númeromuestras} - 1)}\left[\frac{\textit{Númeromuestras}}{\textit{glpormuestra}} - \frac{1}{\sum gl}\right] = 1 + \frac{1}{3(6-1)}\left[\frac{6}{3} - \frac{1}{18}\right] = 1.13$$

Entonces X^2 ajustada = X^2 no ajustada /C = 92.66/1.13 = 82.00.

Consultamos ahora la tabla de X^2 cuadrada con 5 grados de libertad (uno menos que el número de muestras), y encontramos que 82 excede ampliamente el valor tabular en el nivel de significación al $p<0.01$ de probabilidad. La evidencia de que las varianzas son heterogéneas resultan, por tanto, convincentes.

Hemos presentado aquí una forma simplificada de la prueba de Bartlett, basada en tamaños iguales de muestra, puesto que este es el caso más comúnmente encontrado. La prueba puede realizarse con muestras de tamaño desigual, pero los cálculos son más laboriosos. Pueden hallarse mayores detalles sobre esta prueba en los textos de Steel y Torrie (1990) o de Snedecor y Cochran (1989).

El próximo supuesto que examinaremos es el relativo a la independencia entre medias y varianzas. Una rápida observación a los datos es suficiente para convencernos de que dicho supuesto resulta ciertamente incorrecto, puesto que las medias más elevadas tienen varianzas muy grandes y las medias reducidas presentan varianzas muy pequeñas.

Una pregunta importante que debe contestarse con el fin de decidir la transformación que a utilizarse, es la de cuáles están más cercanamente proporcionales a las medias: si las varianzas o las desviaciones estándar. Para esto hemos construido una tabla de proporciones (Tabla 3.9).

Tabla 3.9. Proporciones de varianza y desviaciones estándar para medias en un experimento con vitaminas.

Tratamiento	x	S_i^2	S_i	S_i^2/X	S_i/X
R-C	0.3	0.01147	0.107	0.04	0.36
R-V	0.4	0.00347	0.059	0.01	0.15
G-C	2.4	0.3467	0.589	0.14	0.24
G-V	2.9	0.2133	0.462	0.07	0.16
O-C	137.0	546.0	23.367	3.98	0.17
O-V	151.0	425.3	20.624	2.82	0.14

Puede advertirse que la proporción de varianzas para las medias se incrementa marcadamente al hacer las medias, mientras que la proporción de desviaciones estándar permanecen constantes. A propósito, si las varianzas y las medias no estuviesen relacionadas, cabría esperar que ambas razones disminuyan cuando las medias aumentan.

El supuesto que falta examinar la aditividad. Uno de los aspectos que notamos en los datos originales es que los efectos de bloque difieren ampliamente de especie a especie. Bajo el

supuesto de aditividad, sustrajimos el efecto de bloque promedio de todas las parcelas, para calcular los términos de error. Esta fue la razón principal para que en el análisis de varianza se registrara un término de error grande poco usual.

La prueba formal para la aditividad recibe el nombre de prueba de Tukey. Esta puede llevarse a cabo para probar la no aditividad de dos factores principales cualesquiera. Esto se ilustrará mediante la prueba de los efectos principales de especies y vitaminas. Primero, agrupamos los totales de cada combinación especies-vitaminas en una tabla (Tabla 3.10), y luego calculamos los principales efectos de especies y los efectos de vitamina en los márgenes.

Debemos tener en cuenta que cada casilla de la tabla es la suma de cuatro repeticiones; esto debe considerarse en el cálculo de las medias.

Tabla 3.10. Totales de tratamiento - combinaciones de especies.

Especies	**Control**	**Vitamina**	**Total**	$\overline{X}_{sp} = Totc$	$\overline{X}_{sp} - \overline{X} = S$
Ratones	1.2	1.6	2.8	.35	-48.65
Gallinas	9.6	11.6	21.2	2.65	46.35
Ovejas	548.0	604.0	1152.0	144.0	95.00
Total	558.8	617.2	1176.0		0
$\overline{X}v = Total/12$	46.567	51.433	X=49.0		
$\overline{X}v - \overline{X} = V_j$	-2.433	2.433	0		

Si el cálculo se llevó a cabo correctamente, las sumas, tanto de los efectos de especies como de los efectos de vitaminas deben ser iguales a cero. La media general se obtiene al dividir el total principal entre 24, y el número total de parcelas en el experimento. Debemos calcular ahora Q según la cual multiplicamos cada casilla de la tabla por los efectos de especies y de vitaminas correspondientes:

Q = [1.2 -< (-48.65) >; (-2.433)1 + ... + I 604 x 95.00 >; 2.4331 =12672.4.

La suma de cuadrados para la no aditividad se encuentra entonces como sigue:

SC no aditividad - (Q^2 x total de unidades experimentales/SCEp x SCV), donde SCEp es la suma de cuadrados para las especies, y SCV es la suma de cuadrados para las vitaminas en el análisis de varianza.

Aplicando esta ecuación obtenemos:

SC no aditividad = [(12 672.4)2x 24]/(108 321.16 x 142.11) = 250.375

Esta es una porción de la SC de C x V, de modo que puede probarse como sigue (Tabla 3.11).

Tabla 3.11. Análisis de varianza para la interacción C x V y la no aditividad.

FV	**Gl**	**SC**	**CM**	**F**
C x V	2	250.41		
No aditividad	1	250.375	250.375	7153.6
Residual	1	0.035	0.035	

*: Significativo al p<0.05 de probabilidad, **: Altamente significativo al p<0.01 de probabilidad.

El valor F incluso excede el valor F requerido de 4052 en el nivel de 1 % para 1 g de libertad de no aditividad y 1 grado de libertad del residual, de modo que existe considerable evidencia de que el supuesto de aditividad es incorrecto.

Tenemos ahora comprobados todos los supuestos del análisis de varianza y encontramos que nuestros datos no satisfacen a ninguno de los mismos. No nos debe extrañar que el análisis de varianza haya arrojado resultados desilusionantes.

Quizá la forma más sensible de analizar estos datos consiste en manejar por separado cada especie.

Los análisis son (Tabla 3.12):

Tabla 3.12. Análisis de varianza para especies y vitaminas.

Especies	FV	gl	SC	CM	F
Ratones	Bloques	3	0.0400	0.0133	8.31
	Vitaminas	1	0.0200	0.0200	12.50*
	Error	3	0.0048	0.0016	
Gallinas					
Bloques		3	1.64	0.547	41.00**
	Vitaminas	1	0.50	0.500	37.5**
	Error	3	0.04	0.013	
Ovejas	Bloques	3	2834.0	944.7	157.4**
	Vitaminas	1	392.0	392.0	66.3**
Error		3	18.0	6.0	

Estos resultados son, ciertamente, mucho más satisfactorios que el análisis de varianza general original. Estos análisis son válidos, ya que dentro de alguna especie los datos se ajustan bastante bien a los supuestos básicos. El único defecto de dichos análisis es que revelan muy poco acerca de si las diferentes especies reaccionan análogamente a las vitaminas. Quizá este no sea un aspecto demasiado importante, y en la práctica el investigador estaría satisfecho, sin lugar a dudas, de detenerse en este punto; sin embargo, seguiremos el otro procedimiento de transformación de los datos, a fin de mostrar los resultados notables que pueden obtenerse.

Transformación logarítmica

Debemos afrontar ahora el problema de cómo transformar los datos. Siempre que tengamos datos en los que las desviaciones estándar (no las varianzas) de las muestras sean aproximadamente proporcionales a las medias, la transformación más efectiva será la de tipo logarítmico. Otro criterio para la elección de esta transformación es la evidencia de efectos principales multiplicativos, en vez de aditivos. Ambos criterios se encuentran en los datos con los que estamos trabajando, de modo que intentaremos transformarlos en logaritmos y observar qué sucede.

Antes de empezar, hagamos algunas consideraciones acerca de la aplicación de esta transformación. Los datos con valores negativos no pueden transformarse en esta forma. Si existen ceros entre los datos, afrontaremos el problema de que el logaritmo de cero es menos infinito. Para evitar esta situación, se recomienda sumar 1 a cada dato antes de la transformación. Los datos que contienen un gran número de ceros probablemente se manejarían mejor mediante algún otro método. Se pueden utilizar logaritmos de cualquier base, pero los logaritmos comunes (de base 10) son generalmente los más sencillos. Antes de la transformación, es legítimo multiplicar todos los datos por una constante, puesto que la misma no ejerce ningún efecto sobre

el análisis subsecuente. Es una buena recomendación de que ninguno de los datos sea menor que 1, pues en esta forma se pueden evitar los logaritmos negativos.

En los datos con los que estamos trabajando no existen ceros, pero el menor valor es 0.18, de modo que multiplicaremos todos nuestros valores por 10, antes de obtener los logaritmos. Esto da la siguiente tabla de valores transformados (Tabla 3.13).

Tabla 3.13. Datos del experimento con vitaminas, transformados a log 10X.

Especies—tratamiento	Bloques				Total	Media
	I	I	111	IV		
Ratones- control	0.26	0.48	0.45	0.64	1.83	0.4575
Ratones-vitamina	0.51	0.60	0.62	0.66	2.39	0.5975
Subtotales	0.77	1.08	1.07	1.30	4.22	0.5275
Gallinas-control	1.30	1.48	1.26	1.45	5.49	1.3725
Gallinas-vitamina	1.40	1.52	1.40	1.52	5.84	1.4600
Subtotales	2.70	3.00	266	2.97	11.33	1.41625
Ovejas-control	3.03	3.15	3.13	3.22	12.53	3.1325
Ovejas-vitamina	3.10	3.18	3.17	3.25	12.70	3.1750
Subtotales	6.13	6.33	6.30	6.47	25.23	3.15375
Totales	9.60	10.41	10.03	10.74	40.74	
Medias	1.60	1.735	0.672	0.790		1.69917

El análisis de varianza es como se presenta a continuación (Tabla 3.14).

Tabla 3.14. Análisis de varianza para Vitaminas, Especies e interacción Vitamina x Especie (V x E).

FV	gl	SC	CM	F
Bloques	3	0.12075	0.04025	13.77**
Vitaminas	1	0.04860	0.04860	16.62**
Especies	2	28.54926	0.00476	4883.00**
V x E	2	0.009525	14.27463	1.63
Error	15	0.04385	0.00292	

*: Significativo al $p<0.05$ de probabilidad, **: Altamente significativo al $p<0.01$ de probabilidad.

Este es ciertamente un resultado más satisfactorio que el análisis de los datos originales hasta donde los resultados positivos están implicados. No hemos obtenido aún una interacción significativa entre especies y vitaminas, pero ahora estamos planteando la pregunta en forma diferente. Antes nos preguntamos: "¿Varia de especie a especie la cantidad de cambio de peso debido a la adición de vitaminas?". Ahora nos preguntamos: "¿Varía de especie a especie la proporción o el porcentaje de cambio de peso debido a las vitaminas?". ¿Obtuvimos un resultado más evidente esta vez, simplemente porque estuvimos "disponiendo las cosas" hasta alcanzar un resultado deseado?, ¿o fue justificada la transformación que utilizamos y es válido el nuevo análisis? Para estar seguros, comprobaremos los supuestos del análisis de varianza con los nuevos datos.

Como antes, construiremos una tabla de términos de error (Tabla 3.15), sustrayendo la media, los efectos de tratamiento y los efectos de bloque de cada casilla de la tabla.

Tabla 3.15. Componentes del error de los datos transformados.

Especies-vitaminas	Bloques			
	I	II	III	IV
R-C	-0.10	-0.01	0.2	0.9
R-V	0.01	-0.03	0.5	- 0.0.3
G-C	0.03	0.07	-0.8	-0.01
G-V	0.04	0.02	- 0.03	-0.03
0-C	0.00	-0.02	- 0.02	0.00
0-V	0.02	-0.03	0.02	-0.02

Estos términos de error parecen estar más aleatoria y normalmente distribuidos que aquellos de los datos originales.

Para probar la homogeneidad de la varianza, nuevamente llevamos a cabo la prueba de Bartlett (Tabla 3.16):

X^2=2.3026 ((18 x .9614) -(3 x 5.11)) =4.548

C = 1.13 como antes.

X^2 ajustada =4.548 1.13 =4.03, la cual de acuerdo con X^2 en la tabla A.6, se excedería por casualidad en más de 50% de las oportunidades.

Tabla 3.16. Prueba de Bartlett aplicada a los datos transformados del experimento con vitaminas.

Tratamiento	Media	S_i^2	codificado S_i^2	logaritmos de S_i^2 codificado
Ratón — control	0.4575	0.0243	24.3	1.39
Ratón—vitamina	0.5975	0.0040	4.0	0.60
Gallinas—control	1. 3725	0.0118	11.8	1.07
Gallinas-vitamina	1.4600	0.0048	4.8	0.68
Oveja-control	3.1325	0.0062	6.2	0.79
Oveja-vitamina	3.1750	0.0038	3.8	0.58
Totales			54.9	5.11
Media				9.15
Log de la media				0.9614

Una mirada a la próxima tabla revela que no existen indicaciones de ninguna relación entre las medias y las varianzas.

Realizando la prueba para la aditividad obtenemos, como antes, los siguientes resultados (Tabla 3.17).

Tabla 3.17. Análisis de varianza para no aditividad.

FV	Gl	SC	CM	F
S x T	2	0.009525		
Sin aditividad	1	0.009035	0.009035	18.44 ns
Residual	1	0.000490	0.000490	

*: Significativo al $p<0.05$ de probabilidad, **: Altamente significativo al $p<0.01$ de probabilidad.

El valor F no se aproxima al nivel de significación de $p<0.1$ de probabilidad para 1 gl de si aditividad y 1 gl del residual= 39.86.

Ahora nos sentimos confiados en que el nuevo análisis es válido, puesto que los datos transformados satisficieron todos los supuestos del análisis de varianza. Con los datos originales, ninguno de los supuestos fue verdadero.

Transformación de la raíz cuadrada

Siempre que estamos tratando con cómputos de acontecimientos poco comunes, los datos tienden a seguir una distribución especial, denominada distribución de **Poisson**. Entendemos por acontecimiento poco común aquel que tiene muy baja probabilidad de ocurrir en cualquier individuo; por ejemplo, supongamos (en una partida de semillas de lechuga, 0.1% de las semillas llevaban el virus de la enfermedad del mosaico. probabilidad de que cualquier semilla en particular contenga el mosaico es entonces de sólo 1/1000, éste es un acontecimiento poco común. Si tomamos 100 muestras de 1000 semillas de dicho lote, obtendremos aproximadamente estos resultados:

37 muestras contendrán "0" semillas infectadas

37 muestras contendrán 1 semilla infectada

18 muestras contendrán 2 semillas infectadas

6 muestras contendrán 3 semillas infectadas

2 muestras contendrán 4 semillas infectadas

Resulta obvio que esto se parece muy poco a una distribución normal. Esta distribución de Poisson tiene características muy interesantes: la varianza es igual a la media. En la práctica**, la varianza es generalmente algo mayor que la media**, debido a otros factores, además de la variación de muestreo, que afectan la ocurrencia de los acontecimientos objeto de cómputo. En cualquier proporción, la varianza tiende a ser proporcional a la media.

Cuando analizamos datos de este tipo, estamos violando diversos supuestos hechos en un análisis de varianza. Los errores no están normalmente distribuidos y las varianzas están relacionadas con las medias (siendo, por tanto, homogéneas).

Otro ejemplo de datos de este tipo se encuentra en el conteo de insectos, como el realizado a partir de números estándar de barridas con una malla. Aquí resulta bastante difícil definir qué se entiende por observación individual. Podemos considerarla como un sitio en particular sobre el cual podría hallarse un insecto. Al barrer con una malla, estamos haciendo un muestreo de miles de sitios semejantes y encontrando solamente algunos insectos. Entonces, la probabilidad de hallar un insecto en un punto particular, aleatoriamente escogido en un instante dado es, en realidad, un acontecimiento poco común.

Los datos de este tipo pueden hacerse más normales y al mismo tiempo las varianzas pueden hacerse relativamente independientes de las medias a través de su transformación en raíces cuadradas. En realidad, es mejor utilizar especialmente si existen conteos por debajo de 10.

Los datos presentados a continuación (Tabla 3.18), muestran el número de insectos *lygus* obtenidos en 50 barridas en cada parcela de un experimento para probar 10 insecticidas y un tratamiento de control, repetido cuatro veces en un diseño de bloques completamente aleatorios.

Tabla 3.18. Número de *lygus* por 50 barridas.

Tratamiento	Bloques				Total	Media	s_i^2
	1	II	III	IV			
A	7	5	4	1	17	4.25	6.25
B	6	1	2	1	10	2.50	5.67
C	6	2	1	0	9	2.25	6.92
D	0	1	2	0	3	0.75	0.92
E	1	0	1	2	4	1.00	0.67
F	5	14	9	15	43	10.75	21.58
G	8	6	3	6	23	5.75	4.25
H	3	0	5	9	17	4.25	14.25
1	4	10	13	5	32	3.00	18.00
J	6	11	5	2	24	6.00	14.00
K	8	11	2	6	27	6.75	14.25

El análisis de varianza resultante se observa en la Tabla 3.19

Tabla 3.19. Análisis de Varianza para tratamientos.

FV	gl	SC	CM	F
Bloques	3	12.25	4.08	0.40
Tratamientos	10	380.00	38.00	3.70**
Error	30	308.00	10.27	

*: Significativo al $p<0.05$ de probabilidad, **: Altamente significativo al $p<0.01$ de probabilidad.

Transformando los datos mediante la aplicación de raíz cuadrada, se obtiene la siguiente tabla para el análisis (Tabla 3.20).

Tabla 3.20. Datos transformados del *lygus*.

Tratamiento	Bloques				Total	Media	s_i^2
	I	II	III	IV			
A	2.74	2.35	2.12	1.22	8.43	2.11	0.41
B	2.55	1.22	1.58	1.22	6.57	1.65	0.39
C	2.55	1.58	1.22	0.71	6.06	1.52	0.60
D	0.71	1.22	1.58	0.71	4.22	1.06	0.18
E	1.22	0.71	1.22	1.58	4.73	1.18	0.13
F	2.35	3.81	3.08	3.94	13.18	3.29	0.54
G	2.92	2.55	1.87	2.55	9.89	2.45	0.19
H	1.87	0.71	2.35	3.08	8.01	2.00	0.99
1	2.12	3.24	3.67	2.35	11.38	2.84	0.53
J	2.55	3.39	2.35	1.58	9.87	2.47	0.55
K	2.92	3.39	1.58	2.55	10.44	2.61	0.59

El análisis de varianza de los datos anteriores arrojaría la siguiente tabla (Tabla 3.21).

Tabla 3.21. Análisis de varianza para bloque y tratamientos.

FV	gl	SC	CM	F
Bloques	3	0.532	0.177	0.36
Tratamientos	10	19.993	1.999	4.04**
Error	30	14.841	1492	

*: Significativo al $p<0.05$ de probabilidad, **: Altamente significativo al $p<0.01$ de probabilidad.

Los dos análisis no son demasiado diferentes, puesto que ambos muestran un efecto de tratamiento altamente significativo. El valor F es aproximadamente 10% mayor después de la

transformación. Se registrarán algunas diferencias importantes en la separación de medias (Tabla 3.22).

Tabla 3.22. Prueba de rango múltiple de Duncan, aplicada a datos sin transformar y transformados (nivel del 5%).

Separacion de medias de:	Tratamientos y medias										
	D	E	C	B	A	H	G	J	K	L	F
	0.75	1.00	2.25	2.25	4.25	4.25	5.75	6.00	6.75	8.00	10.00
Datos brutos											
Datos tranformados											

Notamos que en los datos transformados, G y D, G y E, J y D, fueron declarados significativamente diferentes, puesto que no se encontraban entre los datos no trasformados. El efecto general de la transformación es incrementar la precisión con la cual podemos medir las diferencias entre medias pequeñas. Esto es altamente deseable en el trabajo de control de insectos, ya que por lo general no estamos tan interesados en las diferencias entre dos tratamientos relativamente ineficaces como en comparar tratamientos que permitan un buen control.

Una mirada a las varianzas en las dos tablas mostrará que antes de la transformación existió una estrecha relación positiva entre medias y varianzas. El coeficiente de correlación lineal entre las mismas fue de 0.89, significativo al nivel de $p<0.01$ de probabilidad. Después de la transformación, la correlación fue de 0.37, ni siquiera significativa en el nivel de 10%. Por tanto, uno de los supuestos del análisis de varianza fue violado en los datos originales, lo cual se subsanó mediante la transformación.

En general, podemos decir que los datos que requieren la transformación de raíz cuadrada no violan los supuestos del análisis de varianza casi tan drásticamente como los datos que requieren una transformación logarítmica. Consecuentemente, los cambios en el análisis provocado por la transformación no son tan espectaculares.

Transformación angular (arco-seno)

Otro tipo de datos que pueden requerir transformación son los basados en conteos, que pueden expresarse en porcentajes o proporciones de la muestra total. Por regla general, tales datos tienen una **distribución binomial**, en vez de una distribución normal. Como sabemos las distribuciones binomiales se caracterizan porque la varianza es función de la media.

$$\text{Media } \mu = np \qquad \text{Varianza} = \sigma^2 = npq = \mu q$$

Una de las características de esta distribución es que las varianzas se hallan relacionadas con las medias, pero en forma bastante diferente a la de los tipos de datos que hemos considerando. Hasta el momento, los casos que hemos estudiado son aquellos en que medias grandes tienden a

tener varianzas grandes, y viceversa. En los datos binomiales, las varianzas tienden a ser pequeñas en los dos extremos de los rangos de valores (cercanos a cero y a 100%), pero mayores en el medio (alrededor del 50%). En realidad, esta es una idea bastante natural, incluso para quienes no son matemáticos. Tendemos a otorgarle más importancia a una diferencia entre cero y 6%, o entre 94% y 100%, que a una diferencia entre 47% y 53%, aunque todas ellas sean de la misma magnitud.

Para transformar los datos se obtiene el arco seno (inverso del seno) de la raíz de la proporción (arcoseno $\sqrt{p}$), siendo p es el valor proporcional de los datos originales (los porcentajes deben dividirse entre 100). Las unidades de los valores transformados son grados o radianes.

Expresada en notación matemática, ésta es arco seno $\sqrt{X}$ o seno $^{-1}$ $\sqrt{X}$. Se puede utilizar una tabla de arco seno para encontrar las transformaciones directamente de los porcentajes.

Los datos deben transformarse si el rango de porcentajes es mayor que 40. Por otro lado, esto apenas es necesario.

Por ejemplo, en la Tabla 3.23, se muestra la distribución del número de truchas como un porcentaje del total de presas encontradas en los estómagos de 246 individuos, antes y después de aplicarles la transformación angular.

Tabla 3.23. Distribución de la proporción de presas en el estómago de 246 truchas, para los datos originales (% presas) y transformados en el arcoseno (% de presas)/100.

% presas	Grados	N° truchas
0	0,000	5
10	18,435	25
20	26,565	59
30	33,211	60
40	39,232	45
50	45,000	20
60	50,768	12
70	56,789	8
80	63,435	6
90	71,565	4
100	90,000	2

La aplicación de la transformación a los % de las presas aproxima la distribución a una normal, tal como lo muestra la Figura 3.1.

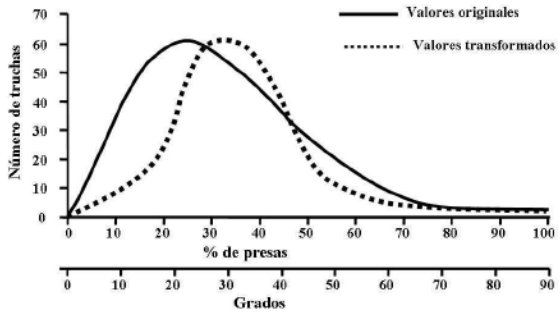


Figura 3.1. Distribución del número de presas (en % de presas ó en grados) en el contenido Estomacal de 246 truchas.

Por otra parte, los datos siguientes corresponden a un experimento en diseño completamente aleatorio (Tabla 3.24), con semillas de lechuga en el que se incluyen 24 tratamientos, cada uno de ellos repetido tres veces. Los tratamientos se encuentran dispuestos por el orden de la magnitud de sus medias (Tabla 3.24).

Tabla 3.24. Número de semillas de lechuga que germinan, en muestras de 50.

	Repeticiones					
Tratamiento	**1**	**2**	**3**	**Media**	**s_i^2**	**Log (10 x s_i^2**
1	0	0	1	0.33	0.33	0.519
2	0	1	0	0.33	0.33	0.519
3	0	0	1	0.33	0.33	0.519
4	0	2	0	0.67	1.33	1.124
5	2	0	0	0.67	1.33	1.124
6	0	2	3	1.67	2.33	1.367
7	7	10	7	8.00	3.00	1.477
8	11	12	15	12.67	4.33	1.637
9	13	18	18	16.33	8.33	1.921
10	22	16	13	17.00	21.00	2.322
11	24	13	18	18.33	30.33	2.482
12	23	21	29	24.33	17.33	2.239
13	24	29	29	27.33	8.33	1.921
14	37	28	27	30.67	30.33	2.482
15	42	41	40	41.00	1.00	1.000
16	39	41	45	41.67	9.33	1.970
17	41	45	40	42.00	7.00	1.845
18	47	41	43	43.67	9.33	1.970
19	45	42	48	45.00	9.00	1.954
20	46	42	48	45.33	9.33	1.970
21	49	46	48	47.67	2.33	1.367
22	48	49	48	48.33	0.33	0.519
23	50	49	48	49.00	1-.00	1.000
24	49	49	50	49.33	0.33	0.519
Totales	178.00	35.767				
10 x media	74.167					
Log (10 x media)	1.8702					

Nótese que hay una marcada tendencia, en las varianzas de los extremos, a ser más pequeñas que aquellas en la mitad de la distribución. Esto es clásico de los datos binomiales. Los logaritmos de las varianzas (codificados mediante la multiplicación por 10) se incluyeron de manera que la prueba de Bartlett puede llevarse a cabo.

X^2 = no ajustada = 2.3026 (log de la media X Σ log codificado S_i^2)

= 2.3026 (1872 x 48-2 x35.767)

C= 1 + (1/(muestras-1)) - (número de tratamientos/gl por tratamiento-1/ Σgl)

=1+1/3x23 (24/2-1/48) =1.1736

X^2 ajustada= X^2/C=35.78

Este sólo es significativo en el nivel de 5% (valor requerido 35.172), de modo que contamos con muy buena evidencia de que las varianzas no son homogéneas. Analizando los datos en bruto, obtenemos los siguientes resultados (Tabla 3.25):

Tabla 3.25. Análisis de varianza para tratamientos de datos no transformados.

FV	gl	SC	CM	F
Tratamientos	23	25266.0	1098.52	148.12**
Error	48	356.0	7.42	

*: Significativo al p<0.05 de probabilidad, **: Altamente significativo al p<0.01 de probabilidad.

Transformando los datos, ejercicio que dejaremos al estudiante, aparentemente eliminamos toda relación entre las varianzas y las medias. Una prueba de Bartlett sobre estas cifras transformadas da un valor de Chi - cuadrada, ajustado igual a 9.00, el cual puede ser excedido por casualidad en más de un 99% de las veces.

Un análisis de varianza de los datos transformados (Tabla 3.26), no parece conducirnos a una conclusión distinta de la del análisis de los datos en bruto.

Tabla 3.26. Análisis de varianza para tratamientos de datos transformados.

FV	gi	SC	CM	F
Tratamientos	23	60725.7	2640.25	102.37**
Error	48	1237.9	25.79	

*: Significativo al p<0.05 de probabilidad, **: Altamente significativo al p<0.01 de probabilidad.

La diferencia importante no se encuentra en el análisis total, sino en la separación de medias. La a prueba de rango múltiple de Duncan mostró los siguientes detalles:

1. Cinco diferencias fueron declaradas significativas antes de la transformación, y no después: 7—8, 8 - 11, 10 - 12, 11 - 12 y 12 - 14.
2. Cinco diferencias fueron declaradas significativas después, y no antes, de la transformación: 18 - 22, 19 - 23, 19 - 24, 20 - 23 y 20 - 24.

¿Qué conjunto de conclusiones debe ser aceptado por nosotros?. La respuesta es sencilla: debemos aceptar las conclusiones basadas en el análisis de mayor validez (en este caso, el análisis de los datos transformados).

Recuérdese que no se transforma los datos para obtener resultados agradables, sino que se transforman para que el análisis sea válido y las conclusiones correctas.

También es importantes mencionar que en los artificios matemáticos utilizados, lo único que varía son los datos de significancia, es decir cuando uno representa sea en tabla o en figura las medias, estas deben ser las originales y solo los niveles de comparación son provenientes de los datos transformados.

Escalas pre-transformadas

Frecuentemente sucede que nos gustaría expresar los datos en porcentajes, pero encontramos este procedimiento muy difícil y engorroso para hacer mediciones precisas. Consideremos, por ejemplo, el problema de evaluar la cantidad de cáscara tubérculos de papa. Una medida conveniente sería el porcentaje de área del tubérculo cubierta por la cáscara; no obstante, resulta muy difícil realizar esta medición con exactitud. Otro ejemplo, sería el porcentaje de área de la hoja cubierta por lesiones provocadas por la enfermedad.

Un último ejemplo, sería el porcentaje de control de la maleza obtenido mediante la aplicación diversos herbicidas. En todos estos casos podríamos realizar un gran esfuerzo para hacer las medidas de los porcentajes con bastante precisión; pero el trabajo contemplado en esta tarea

consumiría tal cantidad de tiempo que el número de parcelas se vería drásticamente limitado. Es una práctica común para realizar un mayor número; mediciones en un tiempo dado, hacer estimaciones visuales aproximativas de los porcentajes, en vez de efectuar mediciones precisas.

Por regla general, se construye una escala, como la escala de 0 a 10 comúnmente utilizada en el trabajo control de la maleza, donde el "0" representa ausencia de control y el 10 indica un 100% de control. La escala representa incrementos iguales de porcentajes, por lo que se recomienda hacer transformaciones angulares.

¿Por qué no pre-transformar nuestra escala? En otras palabras, podríamos escoger escalas de porcentajes tales que, al ser transformados por la transformación angular, resulten en escalas igualmente incrementados que podrían reducirse a enteros (Tabla 3.27).

Tabla 3.27. Escalas de clasificación pre-transformadas. Escala de cero a:

Clasificación	4	5	6	8	10	18	20	24
0	0	0	0	0	0	0	0	0
1	15	10	7	4	2.5	0.75	0.7	0.5
2	50	35	25	15	10	3	2.5	2
3	85	65	50	30	21	6.7	5.5	4
4	100	90	75	50	35	12	10	7
5		100	93	70	50	18	15	10
6			100	85	65	25	20	15
7				96	79	33	27	20
8				100	90	42	35	25
9					97.5	50	42	31
10					100	58	50	37
n						67	58	43
12						75	65	50
13						82	73	57
14						88	80	63
15						93.3	85	69
16						97	90	75
17						99.25	94.5	80
18						100	97.5	85
19							99.3	90
20							100	93
21								96
22								98
23								99.5
24								100

Dichas escalas aprovechan el hecho de que generalmente es más fácil advertir pequeñas diferencias entre 0 y 100% que alrededor del 50%. En realidad, algunas escalas se utilizaron en el pasado, donde fueron deliberada o subconscientemente designadas para ajustarse a estos porcentajes. Con el cultivo de papa (*Solanum tuberosum*) se empleó una escala de 0 a 10, basada en patrones fotográficos que representan los porcentajes aproximados. Con manzanas, se usó una clasificación de almidón que corresponde estrechamente a la escala de 0 a 8 (Tabla 3.27**).** En el trabajo con malezas, se utiliza una escala de 0 a 10, hay la tendencia a usar la clasificación 1, en vez de 10%, para un pequeño índice de control, y la clasificación 9 para un control casi total.

Al analizar datos basados en tales clasificaciones de escala, éstos no deben ser transformados. Debemos hacer hincapié en las clasificaciones de las parcelas de control. Hay una diferencia entre sí, éstas se incluyen en el experimento como un nivel cero de algún factor y se encuentran sujetas a la misma variación que la totalidad de los demás niveles de tratamiento, o si se incluyen

como parcelas de referencia con las cuales comparar las otras parcelas. En el último caso, éstas suelen clasificarse como cero, y las demás parcelas de un bloque se comparan con las mismas. Siendo éste el caso, los datos de las parcelas de control no deben incluirse en un análisis de varianza. Las parcelas de control, con valores de cero arbitrariamente asignados, no tienen varianza. Por tanto, su varianza difiere de la de otros tratamientos, de modo que el supuesto de homogeneidad de la varianza es automáticamente violado.

Ejemplo 3.1. Transformación de lecturas de severidad del tizón (*Phytophthora infestans*) utilizando arco-seno.

Se realizó un experimento de campo en la localidad de Chullchunq'ani (provincia Carrasco del departamento de Cochabamba) a 108 Km. de la carretera antigua Cochabamba - Santa Cruz. Geográficamente situada a 17 º 30´ de Latitud Sud y 65º 15´de Longitud oeste, a una altitud de 3200 msnm. Cuenta con una temperatura media anual de 15.5 ºC y una precipitación pluvial media anual de 629 mm. Es una zona que presenta condiciones favorables para el desarrollo del oomycete *Phytophthora infestans*, que causa la enfermedad del tizòn en papa y otras solanáceas.

```
Options ls=85;
Data campo;
Input rep bloq trat lec1 lec2 lec3 lec4 lec5 lec6 lec7 lec8 lec9;
lec1= (arsin((lec1/100)**0.5))*180/3.1416;
lec2= (arsin((lec2/100)**0.5))*180/3.1416;
lec3= (arsin((lec3/100)**0.5))*180/3.1416;
lec4= (arsin((lec4/100)**0.5))*180/3.1416;
lec5= (arsin((lec5/100)**0.5))*180/3.1416;
lec6= (arsin((lec6/100)**0.5))*180/3.1416;
lec7= (arsin((lec7/100)**0.5))*180/3.1416;
lec8= (arsin((lec8/100)**0.5))*180/3.1416;
lec9= (arsin((lec9/100)**0.5))*180/3.1416;
AUDPC=(lec1+lec2)/2*7+(lec2+lec3)/2*8+(lec3+lec4)/2*6+(lec4+lec5)/2*7+(lec5+lec6)/2*7+(lec6+lec7)/2*7+(lec7+lec8)/2*8+
(lec8+lec9)/2*13;
AUDPCr = AUDPC/900;
Cards;
 1  1  1   3     8    8   10   12  18  33  89  100
 1  1  2   3     8    8    8    8  11  25  70   96
 1  1  3   3     9    9   10   10  15  15  20   98
 1  1  4   2    15   15   18   19  24  27  36   97
 1  1  5   7     7    7    7    7  11  15  26   34
 1  1  6  15    15   15   18   18  25  27  73   99
 1  2  7   1     5    5    7    7  11  13  27   33
 1  2  8   3     3    4    4    4   4   5   9   17
 1  2  9  30    20   20   22   25  25  25  33  100
 1  2 10  17    18   18   18   18  20  25  38   99
 1  2 11   0     0    0    2    3   8   8  11   16
 1  2 12   0.6   6    6    7    8  10  18  39   68
 1  3 13   0     0    0    2    6   6   8   8   14
 1  3 14   4    11   11   11   12  15  16  26   98
 1  3 15   1.6   6    6    6    7   9  13  22   37
 1  3 16   6.6  20   20   25   27  34  36  46  100
 1  3 17  18    18   18   22   26  26  28  36   90
 1  3 18  25    25   25   25   26  30  34  45   99
 1  4 19   0     0    0    0    3   6   9  11   17
 1  4 20  30    30   30   33   35  42  52  66  100
 1  4 21   5     7    7    8    9   9  19  25   25
 1  4 22  15    15   15   16   16  23  27  37  100
 1  4 23  37    37   37   37   37  42  46  98  100
 1  4 24  24    24   24   25   27  29  40  65  100
 1  5 25   1     2    2    5    5  10  18  44  100
 1  5 26  45    45   48   85   94  95  97 100  100
 1  5 27  13    16   16   18   18  18  22  30   34
 1  5 28   2     2    2    2   14  16  24  25   31
 1  5 29  36    36   36   42   42  45  48  96  100
 1  5 30   2     6    6    6    7   9  10  26   28
 1  6 31   2     8    8   10   14  16  25  48  100
 1  6 32  34    34   34   37   44  52  55  63  100
 1  6 33   8     9    9   13   13  15  17  37   38
 1  6 34  10    16   16   23   24  28  35  65  100
 1  6 35   2.4 10     8   13   15  17  18  25   36
 1  6 36  15    15   15   18   23  27  35  43   72
 2  1 32  16    30   30   45   65  68  71  99  100
 2  1  4   4     8   15   28   30  33  35  45   60
 2  1 17  15    15   15   25   28  32  35  46   79
 2  1 27  14    14   17   22   22  23  32  35   47
 2  1 24  10    22   28   38   54  57  65  96  100
 2  1  6   9    13   13   13   14  14  18  44   93
 2  2 15   3    13   13   16    7  11  17  37   39
 2  2 20  33    40   43   46   46  48  49  72  100
 2  2 31   2    12   12   20   20  23  31  45  100
```

```
 2  2 26 45   45   45  46  46  48  48  48   98
 2  2 19  0    0    0   0  12  13  14  15   17
 2  2 10 60   60   60  63  63  67  94 100  100
 2  3 25  0    0    0  18  19  20  20  44   98
 2  3 18 23   23   25  30  32  34  36  56  100
 2  3  7  3    7    8  14  19  21  23  28   34
 2  3  8  5    6    6   7   7   9  13  23   27
 2  3 33 7.4   8   12  12  10  12  18  33   36
 2  3 30  4    7    7   9   9  11  12  25   29
 2  4  5  1    7    6   6   6   8  12  19   30
 2  4 11  0    0    0   0   3   5   9   9   17
 2  4  1 18   26   26  27  33  37  45  86  100
 2  4  9 35   35   35  37  37  37  37  68  100
 2  4 22 15   15   17  25  25  26  32  76  100
 2  4 12  3   15   19  23  26  36  39  55   99
 2  5 36  9    9    9   9  10  13  17  38   91
 2  5 13  0    0    0   0   3   5   5   9   11
 2  5 14 13   13   13  15  15  17  26  46   89
 2  5  2 12   17   17  20  18  20  36  94   93
 2  5 35  1    6    8  14  16  18  25  33   39
 2  5  3 20   20   20  25  25  29  29  40   85
 2  6 29 55   55   66  73  73  84  87  97  100
 2  6 34 11   22   22  25  27  29  35  76   99
 2  6 28  2    3    5  11  22  26  30  31   34
 2  6 21 12   15   15  18  18  19  19  28   38
 2  6 23 12   35   35  38  33  36  44  96  100
 2  6 16 19   35   35  40  43  48  52  93  100
 3  1 35  4   16   10  13  13  15  16  28   36
 3  1  5 12   12   12  13  13  13  17  29   37
 3  1  1 20   20   24  28  28  33  40  94   99
 3  1 13  0    0    0   0   3   5   5   9    9
 3  1 10  1    4    4  18  28  34  95 100  100
 3  1  3 10   23   23  32  32  35  35  43   96
 3  2 33  9   10   12  12  12  12  19  36   38
 3  2 15 10   12   12  15  15  15  25  37   38
 3  2 18 24   35   35  35  38  41  43  89  100
 3  2  9 27   40   42  42  45  47  67  94  100
 3  2 12  4   15   14  14  16  32  38  47   99
 3  2 11  0    0    0   0   3   4   5  10   17
 3  3 14 12   13   13  13  15  16  24  41  100
 3  3 31  4    9    9  14  16  18  26  74  100
 3  3 28  3    6    6  13  25  27  32  33   34
 3  3  8  7    7    7   7   7   7  14  26   28
 3  3 20 44   50   50  52  50  53  56 100  100
 3  3 30  6    9   10  12  12  12  14  26   31
 3  4 29 35   35   37  53  53  57  66  95  100
 3  4 21 25   25   25  26  26  28  28  33   39
 3  4  7 10   12   12  15  15  21  21  28   36
 3  4 17 15   15   15  17  18  25  32  36   43
 3  4 34  6   23   23  20  26  31  38  68   99
 3  4 26 80   80   80 100 100 100 100 100  100
 3  5  2 25   25   25  29  30  30  32  53   98
 3  5 36 15   15   17  17  20  22  26  44   94
 3  5 19  1    1    2   2   6   6  10  10   23
 3  5 22 17   22   22  24  24  24  27  44  100
 3  5  6 13   13   13  15  24  26  29  46  100
 3  5 16 25   37   39  39  67  70  75  92   99
 3  6 25  2    2    6   8   8  13  15  54  100
 3  6 27 27   27   27  27  24  26  28  33   42
 3  6 32 29   38   38  28  39  43  45  47   74
 3  6 24 25   25   25  28  32  36  40  65  100
 3  6 23 23   58   58  66  68  68  74  91  100
 3  6  4  9   25   27  29  29  29  33  42   87
;
Proc univariate plot normal;
Var AUDPCr;
proc discrim method=normal short pool=test;
class trat;
run;
```

En éste experimento, se hizo una transformación de las lecturas de severidad utilizando **arco seno**, para lograr un ajuste a una curva normal y homogeneidad de varianzas, que permita hacer una interpretación correcta del área bajo la curva de progreso de la enfermedad relativa (AUDPCr), que es una medida que relaciona el tiempo y la severidad (1% a 100%).

La salida del SAS University sería como sigue:

Procedimiento UNIVARIATE

Variable: AUDPCr

.Momentos			
N	108	**Sumar pesos**	108
Media	2.37526399	**Observ suma**	256.528511
Desviación std	0.97512158	**Varianza**	0.9508621
Asimetría	0.60615788	**Curtosis**	0.73360963
SC no corregida	711.065181	**SC corregida**	101.742245
Coef. variación	41.0531877	**Media error std**	0.09383112

Test para normalidad				
Test	**Estadístico**		**P valor**	
Shapiro-Wilk	**W**	0.973482	**Pr < W**	0.0294
Kolmogorov-Smirnov	**D**	0.062615	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.093065	**Pr > W-Sq**	0.1404
Anderson-Darling	**A-Sq**	0.583496	**Pr > A-Sq**	0.1310

Las pruebas claves o momentos son todos las de Shapiro-Wilk y Kosmororov-Smirnov, que indican un ajuste a una curva normal y no fueron altamente significativos al $p<0.01$ de probabilidad, con la prueba de Kolmogorov-Smirnov. Se debe indicar que la prueba de Shapiro - Wilk es más estricta.

El análisis de homogeneidad de varianzas mediante la prueba de Chi-cuadrada no mostró significancia al $p<0.05$ de probabilidad.

Procedimiento DISCRIM

Test de homogeneidad de las matrices de covarianza intra

Chi-cuadrado	DF	Pr > ChiSq
0.000000	3185	1.0000

Puesto que el valor de chi-cuadrado no es significante en el nivel 0.1, se usará una matriz de covarianza ponderada en la función discriminante. Referencia: Morrison, D.F. (1976) Métodos estadísticas multivariantes p252.

Ejemplo 3.2. Transformación a raíz cuadrada.

Este experimento fue realizado por la Ing, Gema Burgos López en Puerto La Boca, provincia Manabí, Ecuador, en el año 2019 (Burgos López 2019), con los objetivos de: i) determinar el peso y tamaño de fruto en los tratamientos con poda y sin poda del parental femenino tutorado y ii) determinar el número de semillas hibridas (F1) obtenidas por fruto. Para el desarrollo de esta investigación se implementó una parcela en un invernadero de 500 m^2. Los tratamientos fueron alojados en un diseño experimental de bloques completos aleatorizados con tres repeticiones y dos tratamientos (TP): A1: con poda y A2: sin poda. Las variables de respuesta fueron: altura de plantas (m) (AP), diámetro de tallo (mm) (DT), número de nudos antes del fruto (NDNAF), número de nudos después del fruto (NDNDF), volumen del fruto (cm^3) (VFr), peso de frutos (g) PF), número de semillas vivas (NSV), número de semillas muertas (NSM), peso de semillas con embrión (g) (PS),

El análisis con los datos originales detecto que las variables NDNDF, NSV y NSM no mostraron distribución normal y no había homogeneidad de varianzas, por lo que se hizo una transformación a raíz cuadrada ($X1 = \sqrt{(X + 0.5)}$ en el programa SAS, y así hacer una correcta interpretación de los resultados. Las demás variables tuvieron distribución normal y homogeneidad de varianzas.

```
Data gema;
input rep TP AP  DT NDNAF NDNDF LFr1 LFr2 LFr3 AFr1 AFr2 AFr3 HFr1 HFr2 HFr3 PF NSV NSM PS;
LFr=(LFr1+LFr2+LFr3/3; AFr=(AFr1+AFr2+AFr3/3;HFr=(HFr1+HFr2+HFr3/3;
NDNDF1 = sqrt (NDNDF + 0.05); NSV1 = sqrt (NSV + 0.05); NSM1 = sqrt (NSM + 0.05);
VFr = 4/3*3.1416*LFr*AFr*HFr;
cards;
1 1 1.50 4 14 10 70 75 102 75 80 99 75 80
  99 619 407 28 9
1 1 1.45 4 20 6 84 95 99 75 92 92 75 92
  92 615 362 39 8
1 1 1.50 2 13 11 90 90 91 87 90 93 87 90
  93 . . . .
1 1 1.80 3 26 4 83 93 95 73 95 95 73 95
  95 612 440 35 12
1 1 1.80 4 26 3 84 92 95 73 86 92 73 86
  92 580 423 30 11
1 1 1.40 4 18 5 95 97 100 90 93 95 90 93
  95 615 405 31 8
1 1 1.80 3 21 6 91 101 103 85 93 94 85 93
  94 633 462 38 12
1 1 1.80 4 19 8 95 101 102 87 92 92 87 92
  92 589 433 25 6
1 1 1.60 4 22 4 100 102 105 86 92 94 86 92
  94 597 281 25 8
1 1 1.70 3 19 5 86 91 92 80 89 90 80 89
  90 484 178 61 5
1 2 1.80 4 21 5 100 104 104 87 92 93 57 92
  93 623 280 35 6
1 2 1.60 3 18 6 40 80 82 38 75 76 38 75
  76 . . . .
1 2 2.00 4 26 6 47 83 90 35 73 88 35 73
  88 . . . .
1 2 1.70 3 20 6 100 104 105 86 91 92 86 91
  92 . . . .
1 2 1.70 3 19 5 94 93 95 80 80 80 80 80
  80 533 284 30 7
1 2 1.80 4 20 5 95 98 98 82 90 92 82 90
  92 . . . .
1 2 1.80 3 22 3 80 90 91 70 70 87 70 70
  87 375 75 16 1
```

1	2	1.80	3	18	8	60	80	92	50	69	76	50	69
	76	.	.	.	.								
1	2	1.70	3	12	11	91	93	93	80	86	87	80	86
	87	.	.	.	.								
1	2	1.50	5	7	22	80	84	85	75	81	83	75	81
	83	.	.	.	.								
2	1	1.68	2	23	3	68	81	85	63	73	80	63	73
	80	370	300	29	7								
2	1	1.65	3	19	5	80	84	86	76	84	86	76	84
	86	350	315	10	6								
2	1	1.40	2	21	3	74	78	80	70	77	78	70	77
	78	273	.	.	.								
2	1	1.70	3	19	6	70	73	83	70	83	86	70	83
	86	353	259	13	6								
2	1	1.65	3	21	3	83	85	86	71	72	80	71	72
	80	.	.	.	.								
2	1	1.50	2	16	8	60	73	77	63	72	73	63	72
	73	242	243	8	5								
2	1	1.50	2	19	6	70	73	.	66	67	.	66	67
	.	209	214	100	4								
2	1	1.40	3	25	3	56	64	65	66	71	76	66	71
	76	188	.	.	.								
2	1	1.70	4	28	3	83	95	100	90	97	98	90	97
	98	635	522	25	11								
2	2	1.80	2	23	3	70	57	78	75	85	85	75	85
	85	308	320	40	4								
2	2	1.80	2	18	3	87	89	90	76	85	90	76	85
	90	463	87	80	3								
2	2	2.00	3	25	2	80	89	89	76	85	85	76	85
	85	375	354	25	9								
2	2	2.00	3	28	3	60	67	.	65	82	.	65	82
	.	310	245	46	4								
2	2	2.10	3	24	4	83	86	88	72	82	82	72	82
	82	331	301	11	7								
2	2	2.00	3	25	3	83	95	95	80	87	90	80	87
	90	465	410	50	4								
2	2	1.90	3	22	5	75	82	85	75	84	85	75	84
	85	357	317	10	7								
2	2	1.80	3	19	7	83	89	92	82	91	91	82	91
	91	429	418	15	8								
2	2	1.70	3	18	6	72	83	86	70	80	83	70	80
	83	350	332	8	5								
2	2	1.80	3	19	8	64	80	93	55	73	93	55	73
	93	306	286	19	5								
3	1	1.40	2	18	3	84	87	90	72	79	83	72	79
	83	380	322	35	7								
3	1	1.40	3	19	3	82	89	89	75	83	85	75	83
	85	381	238	36	5								
3	1	1.40	4	16	2	78	84	85	75	82	85	75	82
	85	348	192	28	4								
3	1	1.70	3	20	3	90	96	97	80	89	92	80	89
	92	520	271	30	6								
3	1	1.70	4	17	5	94	98	100	86	94	93	86	94
	93	538	431	8	9								
3	1	1.60	4	17	6	95	101	101	83	89	89	83	89
	89	505	488	10	9								

```
3   1   1.90   3   17   7   92   97   98   83   90   90   83   90
    90  485    275 28   7
3   1   1.55   3   13   10  93   99   100  80   86   86   80   86
    86  .      .   .    .
3   1   1.50   3   16   7   90   92   93   80   82   96   80   82
    96  .      .   .    .
3   1   1.60   4   21   8   83   86   86   80   85   88   80   85
    88  371    .   .    .
3   2   1.90   5   15   9   70   80   82   65   73   73   65   73
    73  236    135 40   6
3   2   2.00   5   24   3   65   81   82   55   71   73   55   71
    73  351    132 43   .
3   2   2.00   3   24   6   70   83   83   60   79   80   60   79
    80  335    .   .    .
3   2   1.90   4   17   7   100  110  110  85   93   93   85   93
    93  616    238 30   6
;
Proc univariate plot normal;
Var AP DT NDNAF NDNDF VFr PF NSV NSM PS;
proc discrim method=normal short pool=test;
class TP;
proc print;
proc glm;
classes TP rep;
model AP DT NDNAF NDNDF VFr PF NSV NSM PS = rep TP;
means TP/duncan tukey;
run;
PROC CORR;
VAR AP DT NDNAF NDNDF VFr PF NSV NSM PS;
RUN;
```

Solución

Análisis de varianza

El análisis de varianza de las variables AP, DT y NDNAF, mostró que los coeficientes de variación (C.V. 8.76 a 20.21) están dentro de los rangos permitidos para este tipo de investigaciones. La variable AP mostró diferencias altamente significativas al Pr<0.01 de probabilidad, lo que estaría indicando que al menos uno de los tratamientos fue diferente (Tabla 3.2.1).

Tabla 3.2.1. Análisis de varianza de las variables altura de planta (AP), diámetro de tallo (DT) y número de nudos antes del fruto (NDNAF).

FV	**Gl**	**CM**		
		AP	**DT**	**NDNAF**
Total	52			
Repeticion	2	0.01ns	4.12**	55.46*
Tratamiento	1	0.76**	0.88ns	1.32ns
Error	49	0.02	0.47	15.95
C.V.(%)		8.76	21.21	20.21

**: Altamente significativo al Pr<0.01 de probabilidad, *: Significativo al Pr<0.05 de probabilidad, ns: no significativo.

El análisis de la variable NDNDF no fue significativo al Pr<0.05 de probabilidad (Tabla 3.2.2), se debe mencionar que para esta variable se hizo una transformación a raíz cuadrada para normalizar los datos, razón por la cual los C.V. se ajustaron los rangos permitidos para esta investigación (Tabla 3.2.2).

Tabla 3.2.2. Análisis de varianza para la variable número de nudos después del fruto (NDNDF).

Origen	gl	SC	CM	Valor F	Pr > F
Total	52				
Repetición	2	2.20	1.10	3.65	0.03
Tratamiento	1	0.18	0.18	0.62ns	0.43
Error	49		0.30		
C.V.(%)	22.64				

**: Altamente significativo al Pr<0.01 de probabilidad, *: Significativo al Pr<0.05 de probabilidad, ns: no significativo.

El análisis de varianza para la variable VFr mostró diferencias significativas al P<0.05 de probabilidad (Tabla 3.2.3).

Tabla 3.2.3. Análisis de varianza para la variable volumen del fruto (VFr).

Origen	gl	SC	CM	Valor F	Pr > F
Total	27				
Repetición	2	2.00	1.00	2.34	0.1183
Tratamiento	1	0.26	2.11	4.93*	0.0361
Error	24	2.11	0.43		
C.V.(%)	2.,01				

**: Altamente significativo al Pr<0,01 de probabilidad, *: Significativo al Pr<0,05 de probabilidad, ns: no significativo

El análisis de varianza para la variable PF no mostró diferencia significativa al Pr<0,05 de probabilidad para tratamientos (Tabla 3.3.4).

Tabla 3.2.4. Análisis de varianza para la variable peso del fruto (PF).

Origen	gl	SC	CM	Valor F	Pr > F
Total	22				
Repetición	2	218224.90	109112.45	10.16	0.0010
Tratamiento	1	25702.70	25702.70	2.39ns	0.13
Error	19		10734.46		
C.V. (%)	22.56				

**: Altamente significativo al Pr<0,01 de probabilidad, *: Significativo al Pr<0,05 de probabilidad, ns: no significativo.

El análisis de varianza para la variable NSV mostró diferencia significativa al Pr<0.05 de probabilidad para tratamientos, mientras el NSM no mostró significación Pr<0.05 (Tabla 3.2.4), se debe mencionar para estas variables se hizo una transformación de raíz de cuadrada para normalizar los datos, razón por la cual los C.V. se ajustaron los rangos permitidos para esta investigación.

Tabla 3.2.5. Análisis de varianza para la variable número de semillas vivas (NSV) y número de semillas muertas (NSM).

FV	**gl**		
		Cuadrados de medios	
		NSV	**NSM**
Total	37		
Repeticion	2	12.70ns	1.10*
Tratamiento	1	59.55 *	0.18ns
Error	34	9.95	0.30
C.V.(%)		18.31	22.64

**: Altamente significativo al Pr<0.01 de probabilidad, *: Significativo al Pr<0.05 de probabilidad, ns: no significativo.

El análisis de varianza para la variable PS no mostró diferencia significativa al Pr<0.05 de probabilidad para tratamientos (Tabla 3.2.6).

Tabla 3.2.6. Análisis de varianza para la variable peso de semillas (PS).

Origen	**gl**	**SC**	**CM**	**Valor F**	**Pr > F**
Total	30				
Repetición	2	8.987	4.493	0.79	0.4644
Tratamiento	1	22.26	22.261	3.91ns	0.0583
Error	27		5.694		
C.V.(%)					

**: Altamente significativo al Pr<0.01 de probabilidad, *: Significativo al Pr<0.05 de probabilidad, ns: no significativo

Análisis de medias

El análisis de medias, mediante la comparación múltiple de tukey para la variable AP mostró diferencias significativas al Pr<0,05 de probabilidad, indicando que el tratamiento sin poda (2), fue el que más altura mostró, lo cual es lógico, desde que se hizo un despunte de los ápices de las plantas con poda (Tabla 3.2.7).

Tabla 3.2.7. Prueba del rango múltiple de Tukey para altura de la planta (AP), número de semillas vivas (NSV), peso de semilla en gramos (PS) al P<0,05 de probabilidad.

Tratamiento	AP	NSV	PS
Con poda	1.59 b	303.14 a	7650 a
Sin poda	1.83 a	85.38 b	5727 b
DSH	0.08	95.21	1.84

Medias con la misma letra, no son significativamente diferentes al Pr<0.05 de probabilidad. DHS: Diferencias Significativamente Honesta.

Para el DT, NDNAF, NDNDF, VFr, PF y MSM, no mostraron diferencias significativas al Pr<0,05 de probabilidad entre los tratamientos en la comparación de sus medias. Indicando esto, que todas fueron iguales estadísticamente.

Para el NSV, se observó diferencias significativas al Pr<0.05 de probabilidad (Tabla 3.2.7), entre los tratamientos, siendo el tratamiento con poda el mejor, mostrando un promedio de 303 semillas/fruto, respecto de 85 semillas/fruto del tratamiento sin poda.

Para el PS, se observó diferencias significativas al Pr<0.05 de probabilidad (Tabla 3.2.7), entre los tratamientos, siendo el tratamiento con poda el mejor, que mostró un promedio de 7650 gramos de semillas, respecto de 5727 gramos de semillas del tratamiento sin poda.

PARTE II

USO DE PAQUETES ESTADISTICOS E INTERPRETACION DE DATOS

UNIDAD 4

INTRODUCCIÓN AL SAS, OTROS PAQUETES ESTADÍSTICOS Y SUS APLICACIONES

Julio Gabriel Ortega

Alfredo Valverde Lucio

Carlos Casto Piguave

Difiniciones

Desde sus orígenes a la fecha el paquete estadístico Statistical Analysys System (SAS) ha evolucionado convirtiéndose en un paquete versátil de amplia aplicación en varias disciplinas del saber científico y la enseñanza. El paquete SAS es un programa aplicado para el análisis y reportes escritos; entendiéndose por sistema un grupo de programas de computadora que interactuan entre sí (SAS 2004).

El empleo del paquete SAS estaba limitado en su uso a minicomputadoras y computadoras de gran capacidad; sin embargo, con el advenimiento de las computadoras de tipo personal (PC) y la implementación del SAS para las PC, no existe excusa para la no utilización de éste u otros paquetes de análisis de datos (SAS 2004). Este avance tecnológico ha favorecido, sin duda, grandemente, a los centros de investigación y enseñanza de provincia.

La necesidad de procesar información por medio de la computación electrónica ha creado la necesidad de conocer los paquetes disponibles. El objetivo de este manual es producir un manual, en español, que sirva de guía para el análisis de los datos resultantes de diseños experimentales o estudios descriptivos utilizando estadísticas paramétricas o noparamétricas.

Este manual abarca sólo aquellos procedimientos de uso más común en el análisis de datos. Para el caso de procedimientos más específicos, no incluidos en este manual se recomienda la consulta de las ediciones más recientes de los libros de SAS.

Para el uso de esta guía es indispensable que el sistema SAS este instalado y que se tengan algunos conocimientos básicos de computación (p.e. los comandos del sistema operativo).

El paquete SAS comprende un conjunto de programas de computadora útiles en el análisis estadístico de datos. Con el paquete SAS se pueden realizar diferentes tipos de trabajos como: Almacenar y recuperar información, modificar la información existente, manejo de archivos obtener diferentes tipos de estadísticas de los datos y análisis complejos de los mismos (SAS 2004).

El paquete SAS trabaja con grupos de datos, los cuales para poder ser manipulados deben de estar en un archivo tipo SAS. Un archivo SAS se compone de: Datos, Variables y Observaciones.

Se debe mencionar que actualmente esta disponible el **SAS University** en forma gratuita y se lo puede bajar e instalar en el siguiente enlace electrónico: https://www.sas.com/es_es/software/university-edition/download-software.html. Esta versión esta en español, pero aun los comandos para realizar los programas están en inglés.

Datos

Un dato es la unidad básica de información de un trabajo SAS. Está formada por un valor específico el cual mide una cierta característica en estudio: por ejemplo, la edad de un cerdo en particular (14 semanas), el peso de un animal (18 kg), el rendimiento de forraje en una parcela (2 ton/ha), la raza de un c erdo (Duroc), etc.

Variables

Una variable es el conjunto de valores (datos) que puede tomar una cierta característica (variable) en estudio en una población o muestra de esta. Por ejemplo, la variable edad puede tomar diferentes valores, los de todos aquellos individuos que intervienen en el estudio, la variable raza puede tomar también diferentes valores, los de aquellos animales que pertenecen a este grupo genético.

Observación

Una observación es un conjunto de valores asociados a cada una de las medidas tomadas de un objeto (individuo) de estudio. Este puede ser un cerdo al cual se le ha registrado su edad, peso, sexo, raza etc.

Data cerdos;

Input trat peso edad;

Cards;

A 14.5 22
A 13.9 20
A 13.5 18
A 14.2 20
B 14.3 21
B 14.5 21
B 13.8 19

Cómo iniciar

En la mayoría de los centros de cómputo, en donde se tiene instalado el paquete SAS, se comienza oprimiendo el icono SAS u oprimiendo los iconos: Inicio Programa Sas Arranca.

Después de unos segundos, el paquete SAS se instala en la memoria de la computadora

y está listo para ser utilizado. En el monitor aparecen tres ventanas con las leyendas OUTPUT, LOG y PROGRAM EDITOR (Figura 1).

La ventana PROGRAM EDITOR permite capturar los datos de los trabajos de investigación o escribir los comandos para ejecutar los programas de SAS.

La ventana LOG es la ventana de comunicación del SAS. En esta ventana aparecen los mensajes de advertencia o notas al usuario sobre posibles errores en los datos o al escribir los procedimientos del SAS. Para tener acceso a esta ventana presione la tecla F3 del tablero.

La ventana OUTPUT muestra los resultados de los procedimientos requeridos por los

usuarios al sistema. Para moverse a esta ventana apriete la tecla F4 o escriba OUT en la línea de comandos.

Figura 4.1. Ventanas SAS

OUTPUT
Command ==>
LOG
Command ==>
EDITOR
Command ==>
00001
00002
00003
00004

Para moverse de una ventana a otra, se escriben los comandos en la línea de comandos de cualquier ventana y se presiona la tecla ENTER, o presionando las teclas función previamente definidas, para ejecutar el comando.

Para conocer las definiciones de las otras teclas función presione la tecla F2. Las definiciones, por default, de las teclas función se presentan en La Tabla 4.1. Para salir de la ventana de las definiciones de las teclas función o de cualquier otra ventana de auxilio escriba END en la línea de comandos y presione la tecla ENTER o RETURN.

Tabla 4.1. Definición de las teclas función

TECLA	
1	AUXILIO
F2	DEFINICION DE LAS TECLAS FUNCION
F3	VENTANA LOG
F4	VENTANA OUT
F5	SIGUIENTE VENTANA
F6	VENTANA PROGRAM EDITOR
F7	AGRANDAR LAS VENTANAS (ZOOM)
F8	SIN DEFINICION
F9	RECUPERAR LO EDITADO EN LA VENTANA PROGRAM EDITOR
F10	EJECUTAR LOS PROCEDIMIENTOS DEL SAS

Existen otras ventanas con propósitos especiales que pueden ser llamadas utilizando comandos del Display Manager.

Cómo capturar información

Para capturar información en SAS es necesario estar en la ventana PROGRAM EDITOR.

Si está iniciando con SAS entonces el cursor se encontrará en la línea de comando del PROGRAM EDITOR. En caso de encontrarse en cualquiera de las otras dos ventanas (LOG o OUT) presione la tecla F6 y el cursor se colocará en la línea de comandos. Presione la tecla ENTER para colocar el cursor en la línea 00001 y comience a capturar la información.

Cómo trabaja en el SAS

Para poder trabajar, SAS necesita crear un conjunto de datos (DATA SET) que sean

reconocidos por el sistema. Para ello el sistema utiliza las sentencia siguientes, DATA "XXX", indica a SAS que cree un conjunto de datos para su uso posterior y que lo denomine XXX, en donde XXX puede ser cualquier nombre no mayor de ocho letras; p.e., DATA FRIJOL, DATA PEPE etc.

INPUT, indica a SAS como se capturó la información en las líneas de datos y qué nombres se han dado a las variables; p.e., INPUT Trat 1-2 Bloque$ 4 Rend 6-8 Peso 10-12; significa que los valores o identificaciones para Trat (tratamiento) se localizan en la columnas uno y dos, la identificación para Bloques en la columna cuatro, los datos de rendimiento (Rend) en las columnas seis a ocho y los datos de peso vivo en las columnas 10, 11 y 12. El signo de pesos ($) después del nombre bloque indica SAS que esta variable (factor) contine caracteres alfanuméricos.

CARDS o DATALINES, esta sentencia indica a SAS que a continuación siguen los datos en el orden indicado en el INPUT.

El cursor se mueve conlas teclas hasta el lugar de la corrección, esto es posible dado que las sentencias no se ejecutan hasta que se oprima la tecla F10. Cada sentencia de SAS debe terminar en PUNTO Y COMA.

Veamos un ejemplo:

PROGRAM EDITOR

Command line

```
DATA cerdos;
INPUT trat 1-2 Bloque 4 Rend 6-8 Peso 10-12;
CARDS;
10      1       240     3.9
10      2       250     4.2
10      3       247     4.1
15      1       253     4.2
15      2       259     4.6
15      3       263     4.9
20      1       265     5.1
20      2       269     5.5
20      3       274     5.8
;
PROC MEANS;
VAR REND PESO;
RUN;
```

RUN indica el SAS que ejecute las sentencias anteriores.

COMO SALVAR UN ARCHIVO. File “a: nombre del archivo”

COMO EJECUTAR EL PROGRAMA. Se presiona la tecla F10 o el icono donde hay una persona corriendo.

COMO TERMINAR UNA SESION SAS. Salir de windows.

Otros paquetes estadísticos para sofware

SPSS

Es un programa estadístico informático muy usado en las ciencias exactas, sociales y aplicadas, además de las empresas de investigación de mercado. Originalmente SPSS fue creado como el

acrónimo de *Statistical Package for the Social Sciences* aunque también se ha referido como "Statistical Product and Service Solutions" (Pardo y Ruiz 2002). Sin embargo, en la actualidad la parte SPSS del nombre completo del software (IBM - SPSS) no es acrónimo de nada.

Es uno de los programas estadísticos más conocidos teniendo en cuenta su capacidad para trabajar con grandes bases de datos y un sencillo interface para la mayoría de los análisis. En la versión 12 de SPSS se podían realizar análisis con 2 millones de registros y 250.000 variables. El programa consiste en un módulo base y módulos anexos que se han ido actualizando constantemente con nuevos procedimientos estadísticos. Cada uno de estos módulos se compra por separado. Por ejemplo SPSS puede ser utilizado para evaluar cuestiones educativas.

Actualmente, compite no sólo con softwares licenciados como lo son SAS, MATLAB, Statistica, Stata, sino también con software de código abierto y libre, de los cuales el más destacado es el Lenguaje R. Recientemente ha sido desarrollado un paquete libre llamado PSPP, con una interfaz llamada PSPPire que ha sido compilada para diversos sistemas operativos como Linux, además de versiones para Windows y OS X. Este último paquete pretende ser un clon de código abierto que emule todas las posibilidades del SPSS (Pardo y Ruiz 2002).

R

R es un software para el análisis estadístico de datos, considerado como uno de los más interesantes. Apoyan esta opinión la vasta variedad de métodos estadísticos que cubre, las capacidades gráficas que ofrece y, también muy importante, el hecho de ser un software libre, es decir, gratuito (Muñoz 2007).

El mayor inconveniente que podría presentar frente al software más utilizado en nuestro medio es el hecho de funcionar mediante comandos, lo que para algunos usuarios puede resultar engorroso. Para solventar esta dificultad existe un paquete llamado **R Commander** que permite utilizar R sin tener que escribir los comandos, es decir, con la sola utilización del ratón (Muñoz 2007).

Una vez realizado el registro para esta actividad, tendrá acceso a materiales que permiten aprender a: descargar e instalar el programa, manipular ficheros de datos, hacer representaciones gráficas y realizar análisis estadísticos de mayor o menor complejidad. De todas formas, es importante destacar que éste no es un curso de estadística, tan solo pretende que el usuario aprenda a manejar R mediante R Commander (Muñoz 2007).

STATISTICA

Es un paquete estadístico usado en investigación, minería de datos y en el ámbito empresarial. Lo creó StatSoft, empresa que lo desarrolla y mantiene. StatSoft nació en 1984 de un acuerdo entre un grupo de profesores universitarios y científicos (Gómez 2017).

Sus primeros productos fueron los programas PsychoStat-2 y PsychoStat-3. Después desarrolló Statistical Supplement for Lotus 1-2-3, un complemento para las hojas de cálculo de Lotus. Finalmente, en 1991, lanzó al mercado la primera versión de STATISTICA para MS-DOS (Gómez 2017).

Actualmente compite con otros paquetes estadísticos tanto propietarios, como SPSS, SAS, Matlab o Stata, como libres, como R.

El programa consta de varios módulos. El principal de ellos es el Base, que implementa las técnicas estadísticas más comunes. Éste puede completarse con otros módulos específicos tales como:

Advanced: técnicas multivariantes y modelos avanzados de regresión lineal y no lineal.

QC: técnicas de control de calidad, análisis de procesos (distribuciones no normales,Gage R&R, Weibull) y diseño experimental.

Data Miner: minería de datos, análisis predictivos y redes neurales.

El paquete puede ser extendido a través de una interfaz con el lenguaje R. Además, se pueden modificar y añadir nuevas librerías usando el lenguaje NET.

MINITAB

Es un programa de computadora diseñado para ejecutar funciones estadísticas básicas y avanzadas. Combina lo amigable del uso de Microsoft Excel con la capacidad de ejecución de análisis estadísticos. En 1972, instructores del programa de análisis estadísticos de la Universidad Estatal de Pensilvania (Pennsylvania State University) desarrollaron MINITAB como una versión ligera de OMNITAB, un programa de análisis estadístico del Instituto Nacional de Estándares y Tecnología (NIST) de los Estados Unidos (Gómez 2017).

Este programa es un paquete estadístico que abarca todos los aspectos necesarios para el aprendizaje y la aplicación de la estadística en general. El programa incorpora opciones vinculadas a las principales técnicas de análisis estadísticos (análisis descriptivo, contrastes de hipótesis, regresión lineal y no lineal, series temporales, análisis de tiempos de fallo, control de calidad, análisis factorial, ANOVA, análisis cluster, etc), además de proporcionar un potente grafico y de ofrecer total compatibilidad con los editores de texto, hojas de cálculo y bases de datos más usuales (Gómez 2017).

MATLAB

Nace como una solución a la necesidad de mejores y más poderosas herramientas de cálculo para resolver problemas de cálculo complejos en los que es necesario aprovechas las amplias capacidades de proceso de datos de grandes computadores (Gómez 2017).

Éste es un entorno de computación y desarrollo de aplicaciones totalmente integrado orientado para llevar a cabo proyectos en donde se encuentren implicados elevados cálculos matemáticos y la visualización gráfica de los mismos. MATLAB integra análisis numérico, cálculo matricial, proceso de señal y visualización gráfica en un entorno completo donde los problemas y sus soluciones son expresados del mismo modo en que se escribirían tradicionalmente, sin necesidad de hacer uso de la programación tradicional.

Está dirigido a ingenieros y científicos, éste requiere que el operador adquiera conocimientos en su lenguaje de programación, se ejecuta principalmente a través de una interfaz de línea de comandos y es más pesado al momento de instalarse, además requiere mayor capacidad en el disco duro, un equipo más rápido (mayor memoria RAM).

Para finalizar se puede decir que MATLAB es un lenguaje de alto nivel y un entorno interactivo para el cálculo numérico, visualización y programación. Usando MATLAB, puede analizar los datos, desarrollar algoritmos y crear modelos y aplicaciones. El lenguaje, las herramientas y funciones matemáticas integradas que permiten explorar múltiples enfoques y llegar a una

solución más rápida que con hojas de cálculo o lenguajes de programación tradicionales, como C / C + + o Java (Gómez 2017).

MSTAT-C

Análisis de datos con MSTAT-C

Introducción

MSTAT o MSTAT-C es un paquete de software estadístico basado en computadora desarrollado por el Departamento de Culturas y Ciencias del Suelo de la Universidad Estatal de Michigan de los Estados Unidos. El software fue escrito por el Dr. Russel Freed, profesor y director de cultivos y el Departamento de Ciencias del Suelo de la Universidad Estatal de Michigan. Este es un método ampliamente utilizado por Investigadores en el campo de las ciencias de la vida, especialmente para el análisis de investigación científica.

Principales características del software MSTAT

- MSTAT es un software de DOS (Disc Operating System) basado
- MSTAT es un programa práctico y no requiere un ordenador rica
- Se requiere una pequeña cantidad de espacios de los discos (Aproximadamente 1,2 Megabytes)
- Se plantea un ahorro automático de sistema que guarda automáticamente los datos putted
- Hay una serie de teclas de acceso rápido mediante el cual cualquier comando se puede forzar fácilmente con teclado
- Los archivos de datos de entrada y salida se pueden ver con cualquier software de procesamiento de textos (Por ejemplo, Microsoft Word)
- Los archivos de salida se pueden imprimir directamente

Ventajas de utilizar el software MSTAT

- MSTAT es muy fácil de usar
- En todos los casos se pide una confirmación YES / NO
- Es capaz de analizar una variable así como los datos multivariados
- Es capaz de analizar los datos de campo de un solo factor de experimento o de múltiples factores
- Se puede calcular la media, desviación estándar y rango de otro valor de dispersión
- Análisis de probabilidad también se puede realizar por software MSTAT
- Se puede calcular un ANOVA completo con probabilidad, así como datos de medias con su Coeficiente de Variación.
- ANOVE se puede comprobar antes del análisis
- ANOVA de varios parámetros se puede obtener a la vez
- Los datos obtenidos lado a otro de cualquier tipo de diseño - CRD, RCBD, cuadrado latino, de factorial puede ser fácilmente analizado por MSTAT-C
- La media de comparación de diferentes factores puede llevar a cabo con la correlación y análisis de regresión entre los parámetros.

Limitaciones de MSTAT-C

- Se trata de un software basado en DOS, por lo que los usuarios tienen que tener una orientación con el teclado
- Los datos no se pueden insertar directamente desde otro software de Windows
- Si un usuario comete un error, él / ella debe volver al menú anterior
- No tienen diferentes comandos para diferentes tareas, es más lento
- En caso de comparación de medias, tiene que introducir los parámetros por separado cada vez para los factores
- Los archivos de datos deben ser vistos con otro software de procesamiento de textos (por ejemplo, Microsoft Word)

Pasos de análisis de datos de diseño experimental con MSTAT-C

Paso 1: Apertura del programa.

Abra la carpeta MSTAT-C y ejecute el archivo MSTATc.exe y la primera pantalla será desplegado. En esta pantalla aparecerán 50 menús diferentes (1-50). Entonces debe seleccionar.

El menú 'ARCHIVOS' para hacer un archivo para analizar.

Paso 2: Creación de un archivo

Después de ejecutar el software, debería tener que crear un archivo. Primero, presione PATH para crear (Por ejemplo, C: \), luego vaya al menú MAKE e introduzca el nombre de archivo deseado (por ejemplo AGRO 516), Título (Análisis) y tamaño de entrada (100).

Paso 3: Entrada de datos

Vaya al menú SEDIT. En primer lugar inserte el caso y luego defina las variables. Después debe definir de nuevo las variables en el menú SEDIT y pulsando EDIT. A continuación, introduzca los datos.

Supongamos que hay un experimento con 2 factores a saber. Variedad (V1, V2 Y V3) y nitrógeno (N1, N2, N3) realizado con RCBD con 3 repeticiones. El acuerdo será como sigue:

Clase	Repetición	Variedad	Nitrógeno
1	1	1	1
2	1	1	2
3	1	1	3
4	1	2	1
5	1	2	2
6	1	2	3
7	1	3	1
8	1	3	2
9	1	3	3
10	2	1	1
11	2	1	2
12	2	1	3
13	2	2	1
14	2	2	2
15	2	2	3
16	2	3	1
17	2	3	2
18	2	3	3
19	3	1	1
20	3	1	2
21	3	1	3
22	3	2	1
23	3	2	2
24	3	2	3
25	3	3	1
26	3	3	2
27	3	3	3

El número de columnas para el parámetro será de acuerdo a sus experimentos.

Paso 4: Análisis del diseño

Después de ingresar los datos, vaya al menú principal y elija 19 FACTOR para hacer Análisis de varianza (ANOVA). Hay 35 paquetes de diseño diferentes. Elija la deseada. Por ejemplo 8. Factor RCBD2 (a). A continuación, introduzca los rangos de variables y finalice el proceso. Elija las variables del grupo (excepto 01. REPLICATION, 02 FACTOR-1 y 03 FACTOR-2)

Paso 5: Visualización del ANOVA y de la tabla media

Durante observar la tabla ANOVA y Mean no olvide anotar la Error Mean

Square (EMS) y Error.

Paso 6: Vuelva a comprobar los datos de entrada

Después de analizar nuevamente vaya al SEDIT y anote el primer caso de factores 1, primer caso del factor 2 y el primer caso de interacción. El primer caso significa que los factores vuelven a situarse en la misma línea horizontal

Paso 7: Comparación media

Vaya al menú 'RANGE' y pulse P (parámetro) y los valores de entrada en el campo. Existen varios Test de separación media en MSTAT-C a saber. LSD, gama múltiple de Duncan, Prueba de Tukey y prueba de Student-Newman-Keul.

Las medias pueden ser probadas con diferentes niveles significativos a saber. Nivel Alfa 0,01 (1% de niveles de Significancia), nivel Alfa 0,05 (niveles de significación del 5%) y nivel Alfa 0,1 (10% Niveles de significación).

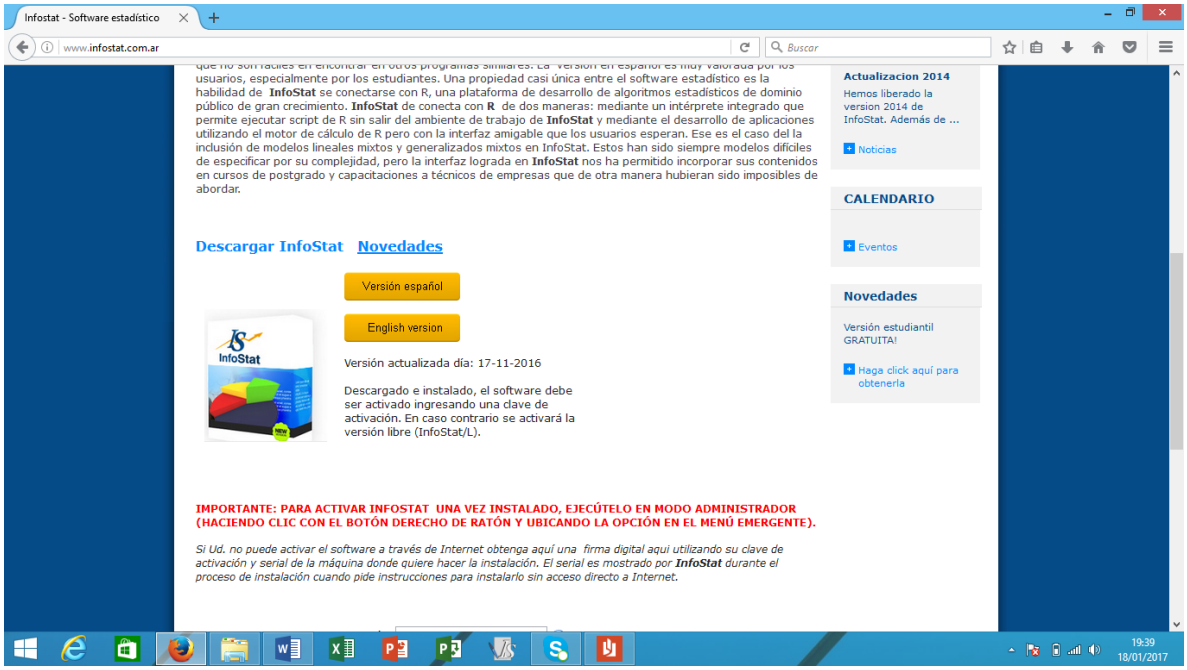
Después de terminar la comparación de un factor de nuevo al campo e ingresar de nuevo el siguiente Factores. (Hasanuzzaman, 2008) y (Russell, Everett, Weber, Eldor, & Isleib, 1993)

InfoStat

InfoStat es un software para análisis estadístico de aplicación general desarrollado bajo la plataforma Windows.

Cubre tanto las necesidades elementales para la obtención de estadísticas descriptivas y gráficos para el análisis exploratorio, como métodos avanzados de modelación estadística y análisis multivariado. Una de sus fortalezas es la sencillez de su interfaz combinada con capacidades profesionales para el análisis estadístico y el manejo de datos. Una propiedad casi única entre el software estadístico es la habilidad de InfoStat se conectarse con R, una plataforma de desarrollo de algoritmos estadísticos de dominio público de gran crecimiento. InfoStat se conecta con R de dos maneras: mediante un intérprete integrado que permite ejecutar script de R sin salir del ambiente de trabajo de InfoStat y mediante el desarrollo de aplicaciones utilizando el motor de cálculo de R pero con la interfaz amigable que los usuarios esperan (www.Infostat.com).

El InfoStat puede ser descargado en español de manera libre, es cuestión de seguir los pasos.



Ventajas

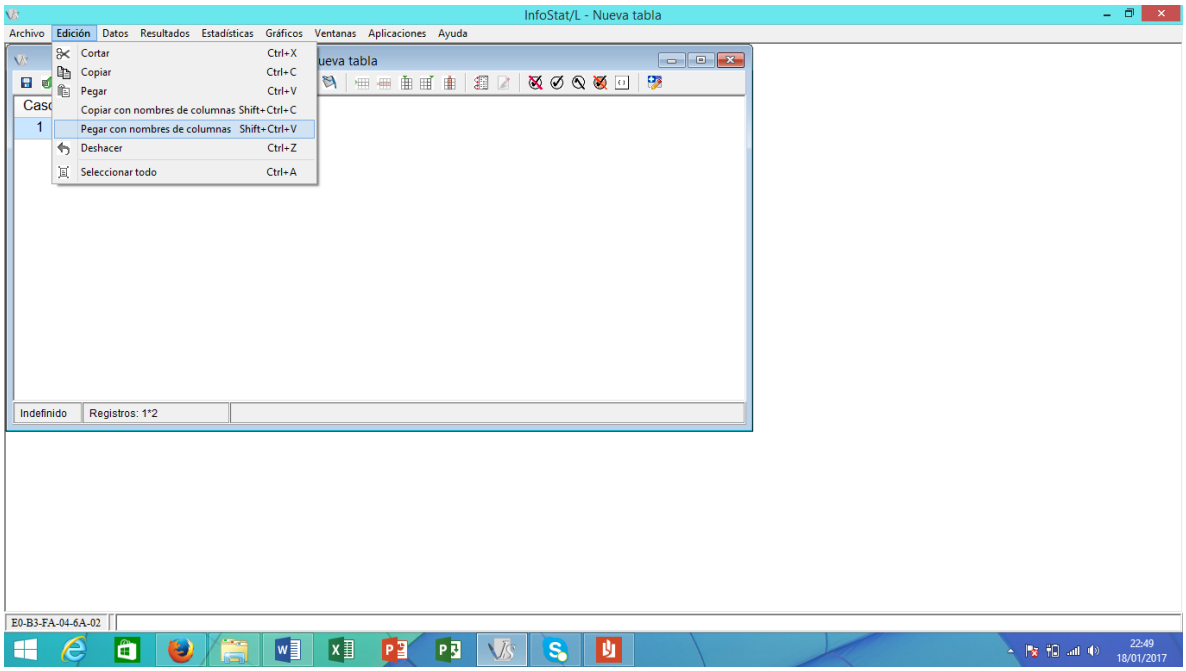
Gratuito, en español, interfaz intuitiva, fácil de usar, Loa datos se pueden cargar importando formatos, las tablas soportan las funciones de copiar y pegar, crea datos de manera sencilla, cubre necesidades para obtener estadística descriptiva y gráfica, conectividad con software más sofisticado “R”.

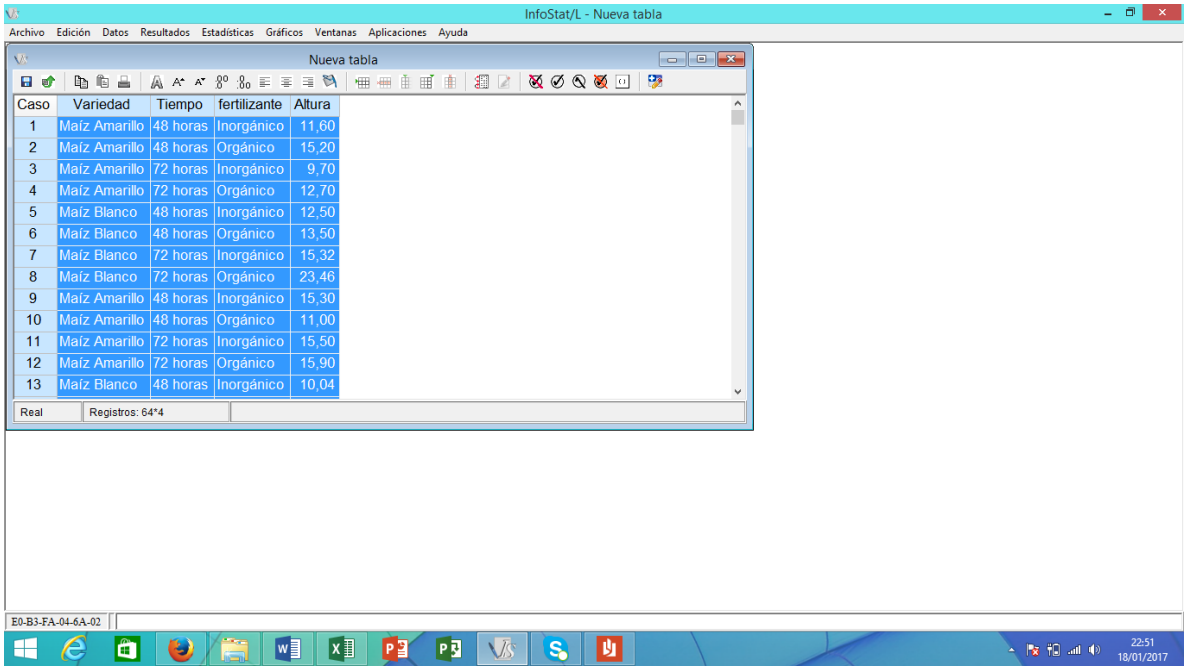
Desventajas

No disponible para otras plataformas (Linux, Mac), no es muy utilizado.

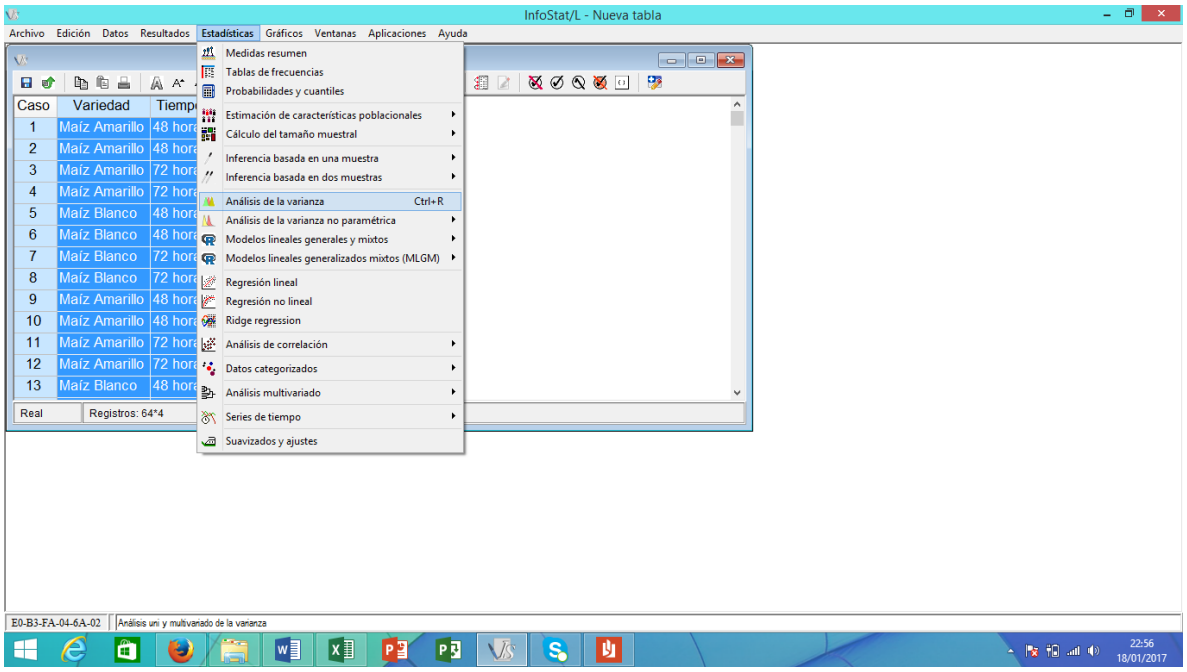
Aplicación del InfoStat en diseños experimentales

Como se indicó el InfoStat permite importar tablas del Excel, permitiendo copiar incluso con nombre de columnas, tal como se demuestra en los gráficos:

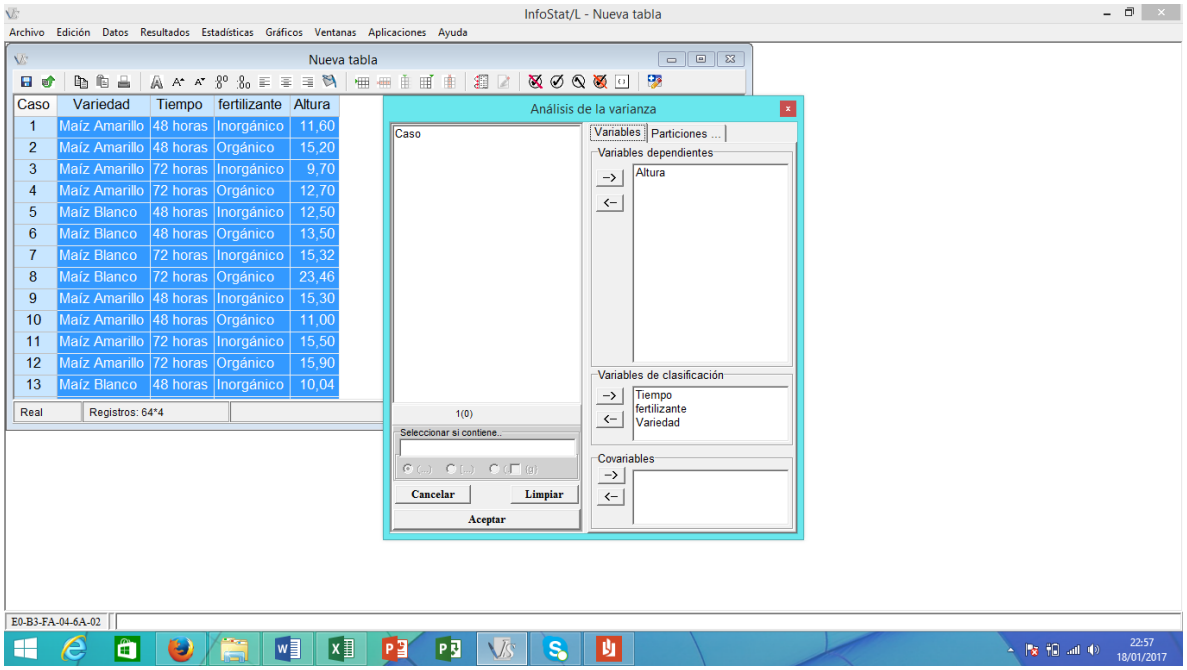




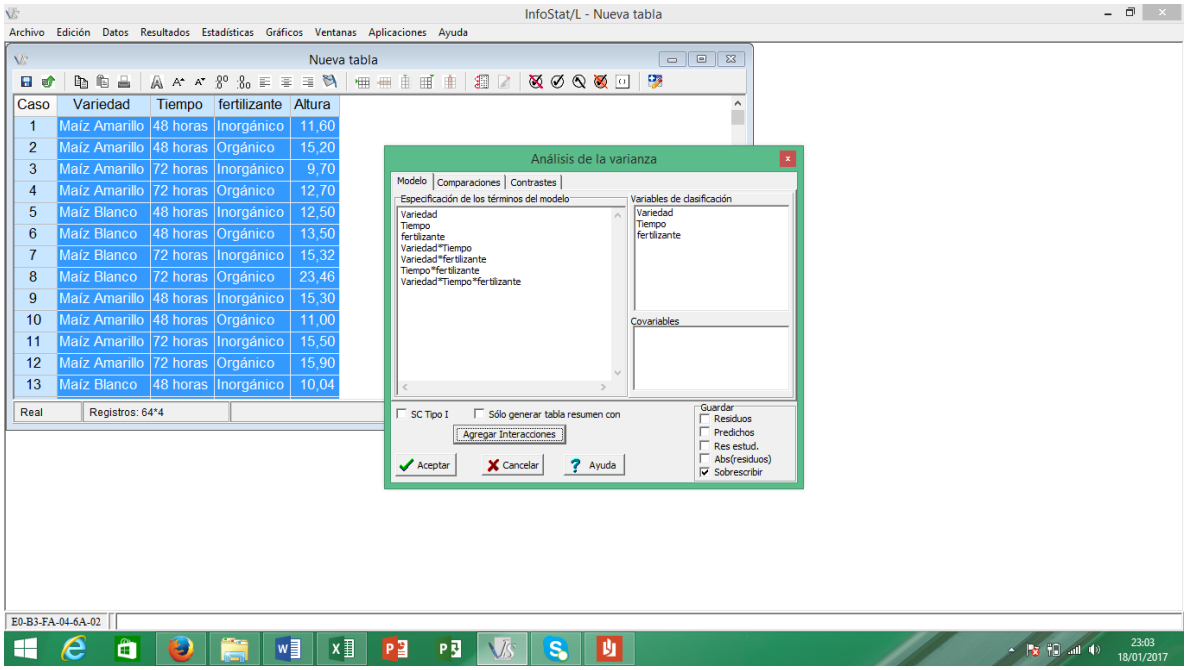
Con los datos copiados, se procede a aplicar según el tipo de diseño experimental establecido, para esto se necesario ir a estadística, y dentro de ella escoger análisis de varianza, tal como se demuestra a continuación:



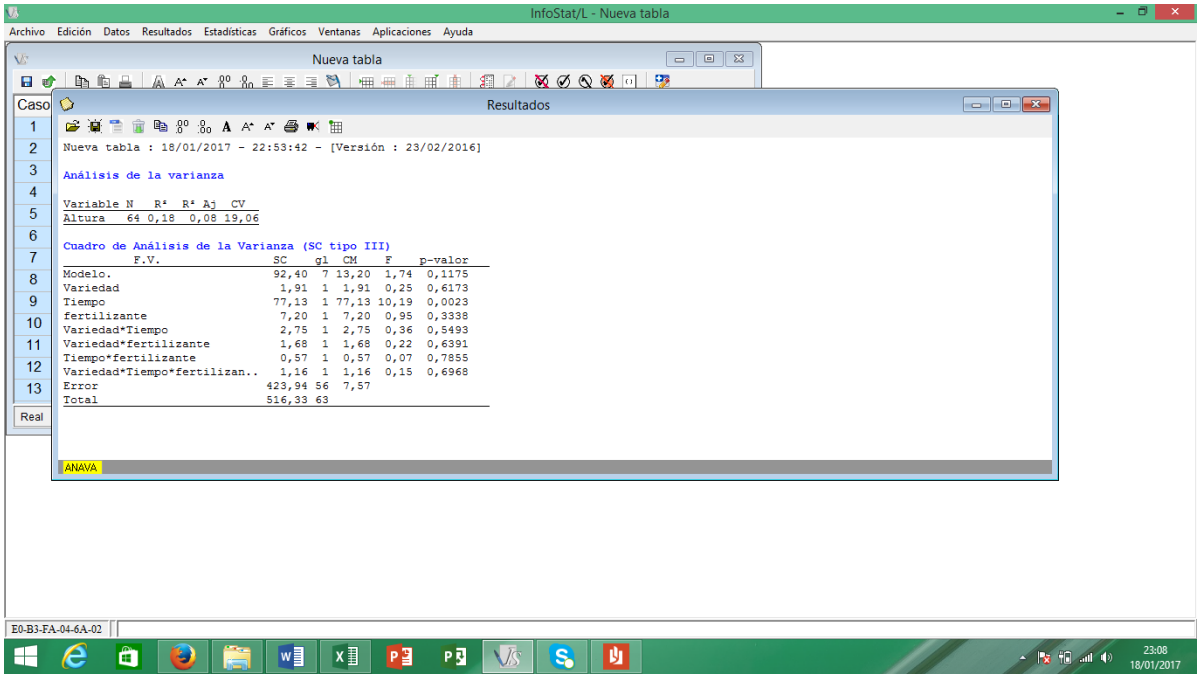
Se seleccionan la variable dependiente, así como las variables de clasificación. Como ejemplo podemos ver la variable altura (variable dependiente) y tiempo, fertilizantes y variedad (variables de clasificación).



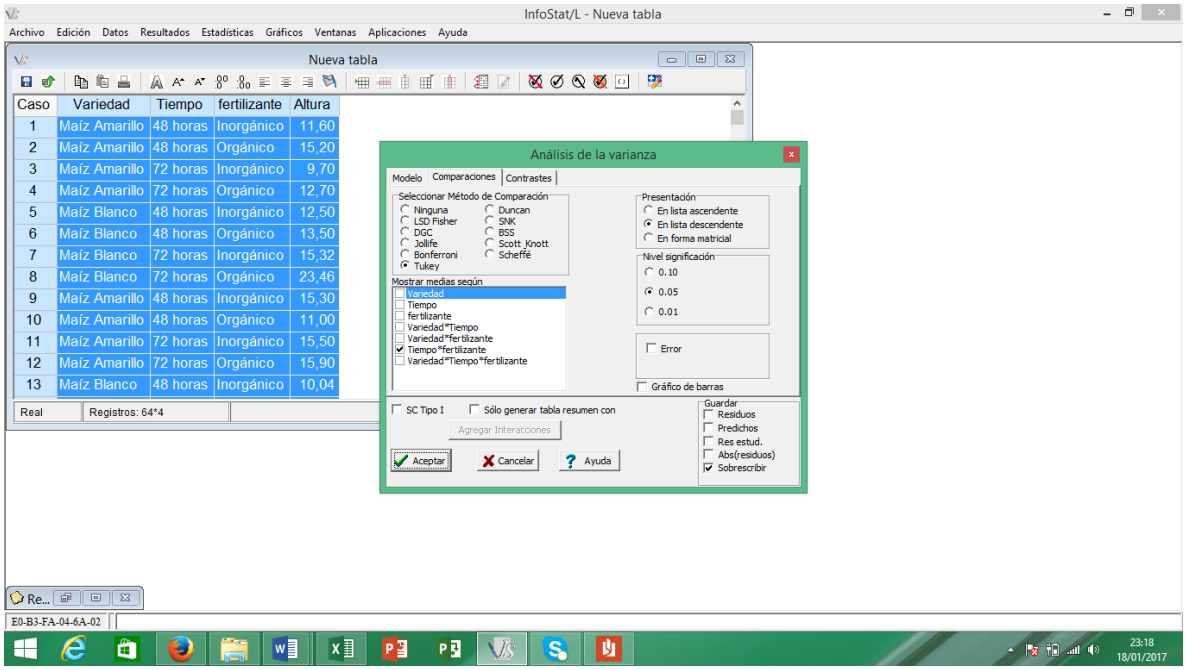
En la siguiente ventana, es necesario agregar interacciones. aclarando oportunamente, que tal acción solo es necesario cuando exiten arreglos factoriales como en el ejemplo que estamos citando (factorial 2x2x2).



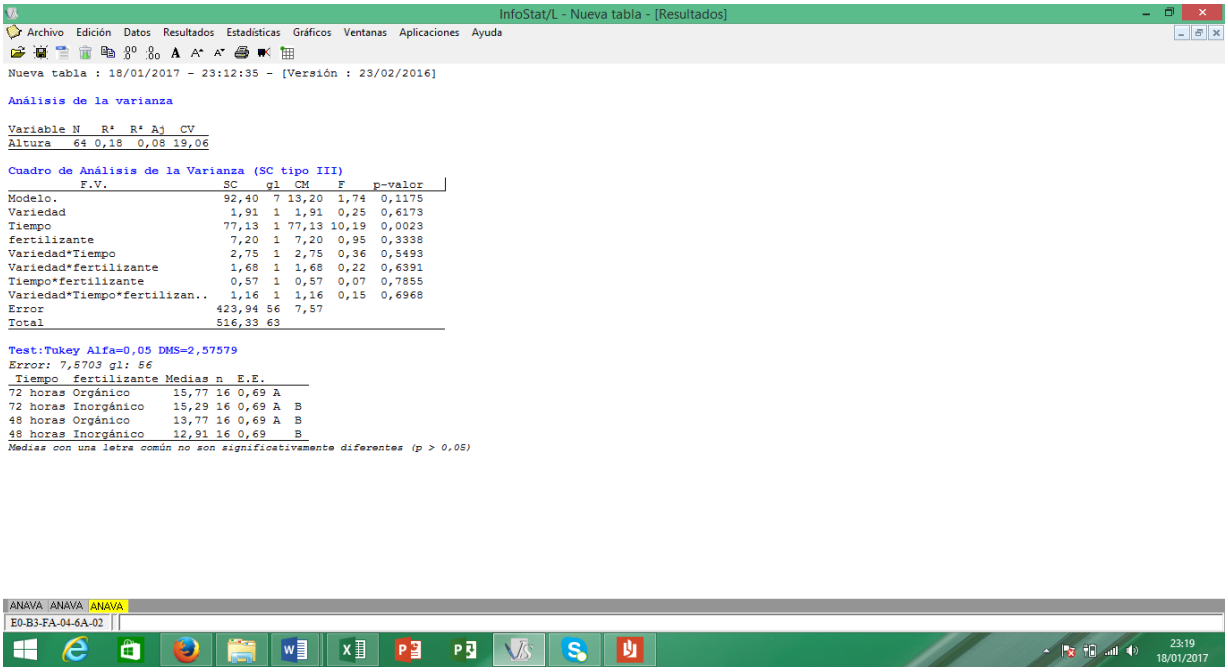
Realizada esta acción, a continuación el programa estadístico nos presenta el cuadro de análisis de varianza.



Finalmente, y solo por fines didácticos, es oportuno indicar que el InfoStad permite aplicar las pruebas de comparaciones o pruebas de rango múltiples, tal como se describe a continuación:



Sin embargo, es oportuno destacar que este paso se debe dar solo si el análisis de varianza determina significación entre las variables, de lo contrario no es necesario realizarlo. Se debe interpretar apropiadamente para tomar las decisiones adecuadas.



UNIDAD 5
INTERPRETACIÓN DE RESULTADOS DEL SAS

Julio Gabriel Ortega

Conceptos e interpretación de datos

Unidad experimental

- Es la unidad a la que se aplican los tratamientos

Variable: Es una característica medible de la unidad experimental.

- **Discreta:** (discontinua) toman un numero finito o contable de valores (Ej. Porcentaje de germinación, numero de caras que aparecen cuando se arroja dos veces una moneda).
- **Continua**: toman un número infinito no contable de valores (Ej. Peso de un animal, la altura de un árbol).

Muestra: Es un conjunto de mediciones que constituye parte de una población

Conceptos básicos para la interpretación

- **Distribución normal**

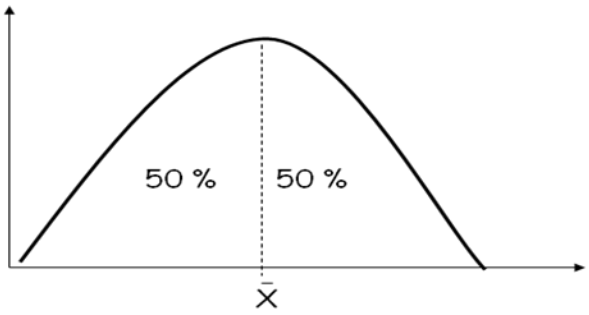


Figura 5.1. Los valores de un variable aleatoria su promedio o media aritmética parte a la "campana" en dos partes con poca o mucha dispersión de datos.

Medidas de tendencia Central

- **Media**: es el valor que se obtiene multiplicando cada valor de la variable por su probabilidad y sumando
- **Moda**: Es el valor de la variable que tiene máxima probabilidad, ósea es el valor mas frecuente
- **Mediana**: Es todo valor X que deja una probabilidad acumulada ≤ 0.5 a la derecha y a la izquierda. La mediana parte la distribución al medio dejando a ambos lados dos colas de aproximadamente el mismo peso.

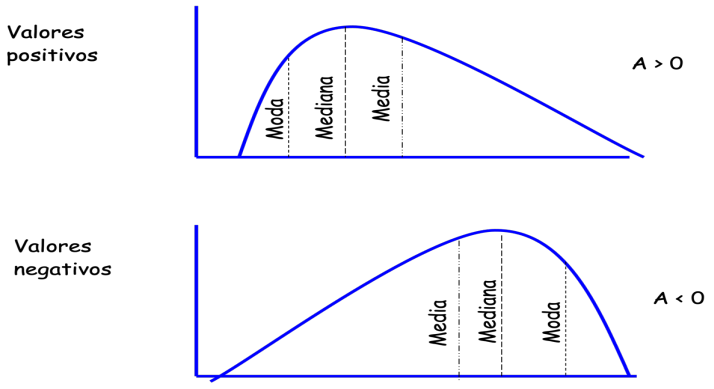


Figura 5.2. Medidas de tendencia central con valores positivos y negativos.

Medidas de dispersión o variación

- Varianza:

$$\sigma^2 = \frac{\sum (Xi - \mu)^2}{N} \qquad s^2 = \frac{\sum (Xi - \bar{X})^2}{n-1}$$

- Desviación estandar:

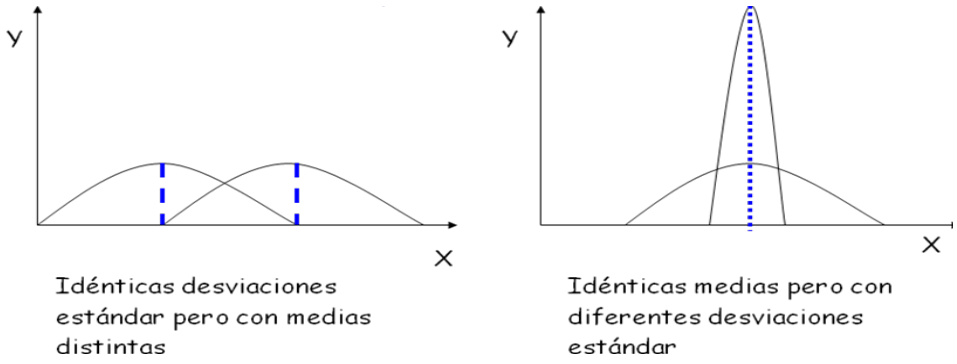
$$\sigma = \frac{\sum (Xi - \mu)^2}{N} \qquad s = \frac{\sum (Xi - \bar{X})^2}{n-1}$$

Determina el grado de amplitud o dispersión entre los elemento

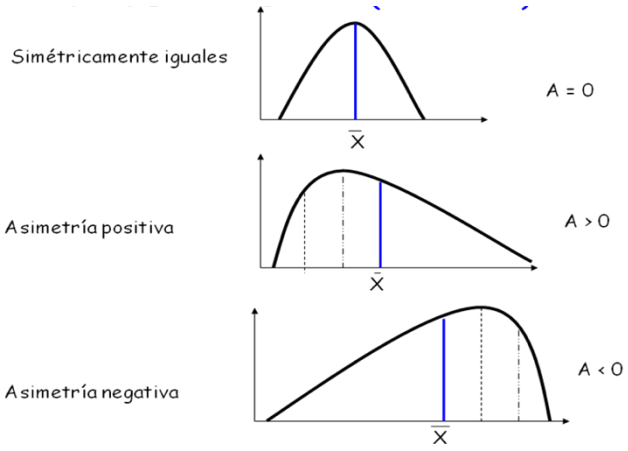
Peso de conejos

Kg/conejo X	$(X - \bar{X})$	$(X - \bar{X})^2$
3	0	0
4	1	1
5	2	4
2	-1	1
1	-2	4
$\sum X = 15$	$\bar{X} = 3$	$\sum (X - \bar{X})^2 = 10$

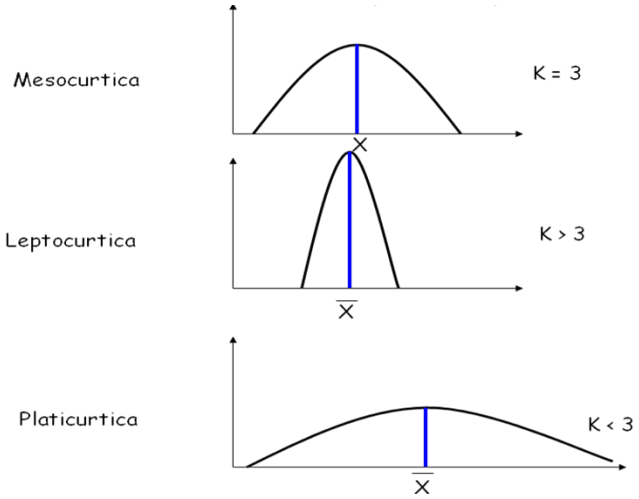
Distribución normal y otras distribuciones



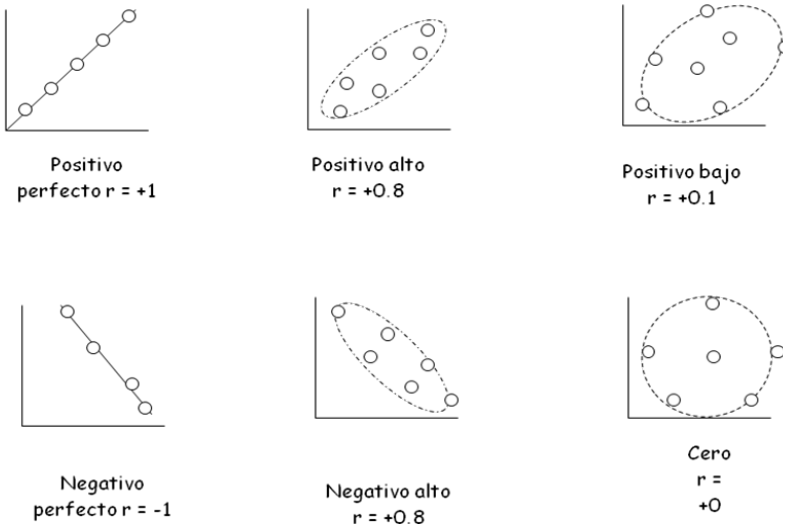
Grado de asimetria (Skewness)



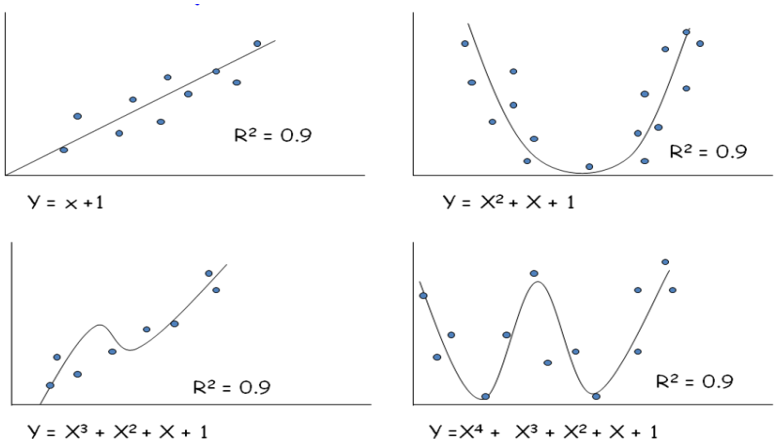
Grado de empinamiento (Kurtosis)



Coeficiente de correlación



Ecuaciones polinómicas



Análisis de varianza

Ejemplo: Diseño completamente aleatorio

- Es el diseño más simple
- Recibe el nombre, por qué los tratamientos son aplicados en forma aleatoria y sin restricción.
- Aplicable cuando las unidades experimentales son homogéneas y la administración es uniforme.
- Es el diseño que tiene el mayor número de grados de libertad asignados al error experimental.

Modelo aditivo lineal

$Yin = m + tj + eij$
$i = 1, \ldots..3$, b = Tratamientos
$j = 1, ..., 14$, c = Repeticiones

Donde:

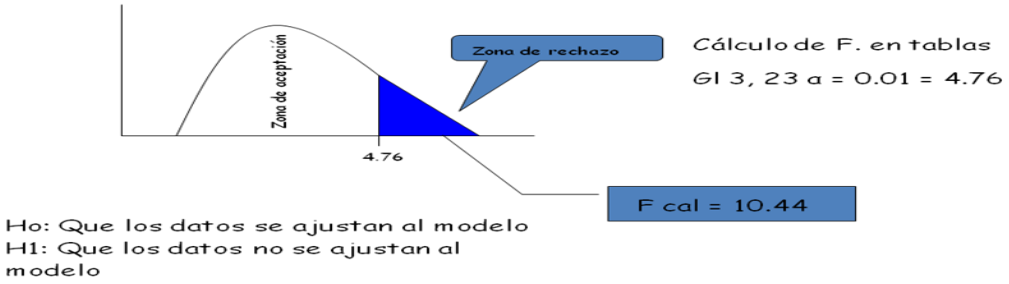
Yij = Valor observado de una variable de respuesta, en el i-esimo bloque, que recibe el j-ésimo cultivar.
m = Media general del ensayo
tj = Efecto fijo del j-ésimo cultivar.
eij =Efecto aleatorio de los residuales; eij ~ NIID(0,σ2e)

Pasos de análisis

1. Se calcula el factor de corrección C
2. C=G2/ Σ ri
3. Se calcula la suma de cuadrados debido al total
4. SCTOTAL = Yij2 – C
5. Se calcula la suma de cuadrados debido a los tratamientos de parcela grande
6. SCT = ΣTij2/ ri - C
7. Análisis de varianza

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F
Tratamiento	t - 1	SCT	CMT = SCT/t - 1	CMT/SCE
Error Experimental	$\sum r_i$ - t	SCE	CME = SCE/tr-a = s^2	
Total	$\sum_{ri}$ - 1	SCTOTAL		

Prueba de significancia del modelo



Conclusión: Como la Fcal > Ftablas se rechaza la hipótesis nula y se acepta la hipótesis alternativa. Se concluye indicando que los datos no se ajustan al modelo

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: PESO

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 27 MSE= 0.687815
Critical Value of Studentized Range= 4.864
Minimum Significant Difference= 2.0172

Means with the same letter are not significantly different.

Tukey Grouping			Mean	N	TRAT
	A		3.8500	4	C
B	A		3.5000	4	G
B	A	C	2.8250	4	H
B D	A	C	2.6250	4	F
B D	A	C	2.1750	4	B
B D	A	C	2.0500	4	I
B D	A	C	2.0500	4	D
B D		C	1.5000	4	A
D		C	1.4750	4	J
D			0.7000	4	E

Análisis de normalidad y homogeneidad de varianzas

Univariate Procedure

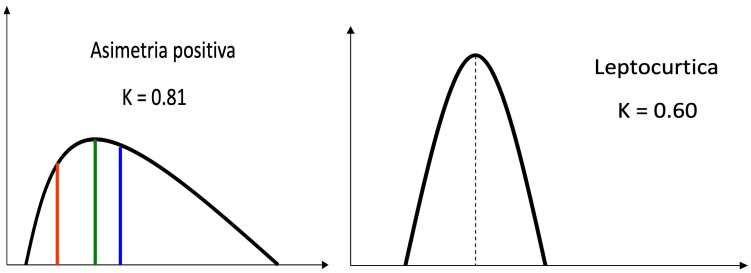
Variable=AUDPC

Moments

N	108	Sum Wgts	108
Mean	2036.108	Sum	219899.7
Std Dev	1172.156	Variance	1373949
Skewness	0.81804	Kurtosis	0.601519
USS	5.9475E8	CSS	1.4701E8
CV	57.56844	Std Mean	112.7907
T:Mean=0	18.05209	Pr>|T|	0.0001
Num ^= 0	108	Num > 0	108
M(Sign)	54	Pr>=|M|	0.0001
Sgn Rank	2943	Pr>=|S|	0.0001
W:Normal	0.943766	Pr<W	0.0003

Conclusión: El valor de la probabilidad de normalidad es 0.0001. Este valor es menor a $\alpha = 0.05$, lo lo tanto se concluye indicando que los datos no se distribuyen normalmente.

Asimetría y empinamiento



El valor de **Skewness (asimetrìa)** es 0.81 mayor que cero (A>0) por tanto tiene una asimetria positiva con una cola a la izquierda.

El valor de **Kurtosis (empinamiento)** es o.60 mayor que K>3 por tanto la gran mayoría de los datos están cercanos a la media y por tanto adquieren una forma leptocurtica.

100% Max	65	99%	65
75% Q3	32.7	95%	48.5
50% Med	**25.0**	**90%**	**45.2**
25% Q1	20.2	10%	19.5
0% Min	16	5%	18.5
		1%	16
Range	49.0		
Q3-Q1	12.5		
Mode	19.7		

Conclusión: El valor medio (mediana) en cuanto a las cuantiles es 45.2, por tanto el 90% de los datos de protombina están cercanos a la media.

Tallo y hoja

Extremes

Lowest	Obs	Highest	Obs
16 (	26)	38.4 (	6)
18.5(	25)	40.5 (	7)
19.5(	18)	45.2 (	10)
19.7(	22)	48.5 (	5)
19.7(	21)	65 (	3)

Ubicación de los Valores extremos en la muestra

Stem	Leaf	#	Boxplot
6	5	1	0
6			
5			
5			
4	58	2	|
4	0	1	|
3	8	1	|
3	134	3	+-----+
2	567789	6	*--+--*
2	00000112223	11	+-----+
1	68	2	|

----+----+----+----+

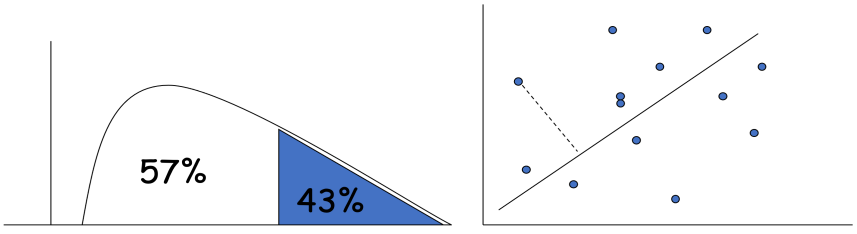
Multiply Stem.Leaf by 10**+1

Conclusión: LA PRUEBA DE TALLO Y HOJA demuestra que la cola esta a la izquierda por tanto tiene una asimetría positiva A>0.

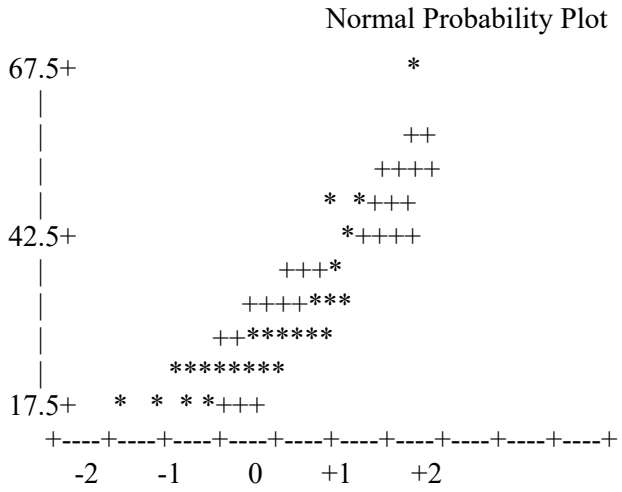
La prueba de boxplot demuestra que la media esta casi cerca de la mediana: Cuanto más cerca están + y * más se asemeja a la curva de normalidad.

Coeficiente de regresión (R^2)

R^2 = 0.57. Significa que **e**l 57% de los datos se ajustan al modelo, mientras 43% están distantes de la línea de regresión y son debidos a efectos ambientales. Mayores detalles sobre el coeficiente R^2, son discutidos en la unidad 11 del capítulo V.



Análisis de normalidad



Grado de dispersión de los datos

Análisis de homogeneidad de varianzas

Discriminant Analysis Test of Homogeneity of Within Covariance Matrices

Notation: K = Number of Groups

P = Number of Variables

N = Total Number of Observations - Number of Groups

N(i) = Number of Observations in the i'th Group - 1

$$V = \frac{\prod |\text{Within SS Matrix}(i)|^{N(i)/2}}{|\text{Pooled SS Matrix}|^{N/2}}$$

$$RHO = 1.0 - \left(\sum \frac{1}{N(i)} - \frac{1}{N} \right) \frac{2P^2 + 3P - 1}{6(P+1)(K-1)}$$

$$DF = .5(K-1)P(P+1)$$

Under null hypothesis: $-2\ RHO \ln \left| \frac{N^{PN/2}}{\prod N(i)^{PN(i)/2}} V \right|$

is distributed approximately as chi-square(DF)

Test Chi-Square Value = 18.140513 with 9 DF Prob > Chi-Sq = 0.0336

Conclusión: El valor de Chi-Sq = 0.0336. Este valor no es mayor que $\alpha = 0.01$ por tanto no existe homogeniedad de varianzas. Mientras bajo el mismo valor a $\alpha = 0.05$ muestra que existe homogenidad de varianzas.

PARTE III

DISEÑOS EXPERIMENTALES SENCILLOS Y APLICADOS

UNIDAD 6

DISEÑO COMPLETAMENTE ALEATORIO (DCA)

Julio Gabriel Ortega

Carlos Castro Piguave

Blanca indacochea Ganchozo

Definiciones

El diseño completamente aleatorio es el más simple y utilizado de todos. Es aplicable cuando las unidades experimentales son homogéneas y la administración del experimento es uniforme para todas ellas. Al concluir el experimento las unidades experimentales mostrarán diferentes resultados atribuibles en forma exclusiva a los tratamientos aplicados.

Este diseño es muy utilizado en experimentos de laboratorio, invernadero, almácigo y establos, en los que el material experimental (macetas, bandejas, almácigos, animales, etc.) es muy homogéneo por prepararse en forma provisional, y porque el experimento se conduce en condiciones ambientales controladas y uniformes para todas las unidades experimentales.

Ventajas

Los tratamientos, en experimentos unifactoriales, pueden tener igual o diferente número de repeticiones. Sin embargo, cuando se estudia más de un factor es recomendable tener el mismo número de repeticiones.

Es el diseño que tiene el mayor número de grados de libertad asignados al error experimental, beneficiando la estimación del mismo y logrando una mayor precisión en la estimación de las medias de los tratamientos y de las diferencias de éstas.

Tratamientos

En este diseño, como en cualquier otro, es muy importante seleccionar adecuadamente los tratamientos de estudio. Tratamiento es aquello que se aplica como se dijo antes a las unidades experimentales o la forma en que éstas son administradas. Ahí son tratamientos razas de animales, dosis de vacunas, dosis de fertilización, diferentes pesticidas, herbicidas, formas de riego, etc.; así como las diferentes combinaciones que pueden hacerse con los niveles de estos factores.

Los tratamientos pueden ser escogidos o fijados por el investigador. El primer caso, llamado *modelo al azar*, es relativamente frecuente en investigaciones de mejoramiento genético, para la evaluación de poblaciones o familias. Estudios sobre las dosis, pesticidas, sistemas de riego constituyen el segundo caso, llamado *modelo fijo*. La forma de seleccionar los tratamientos no afecta el cálculo del análisis de varianza, pero sí los cuadrados medios esperados y las pruebas de comparación de medias.

Cualquiera que sea la forma de seleccionar los tratamientos, estos deben corresponder a los objetivos del experimento y a las hipótesis a ser probadas.

Modelo aditivo lineal

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

i = 1, …..3, b = Tratamientos
j = 1, ..., 14, c = Repeticiones

Donde:

Y_{ij} = Valor observado de una variable de respuesta, en el i-esimo bloque, que recibe el j-ésimo cultivar.

μ = Media general del ensayo

τ_j = Efecto fijo del j-ésimo cultivar.

ε_{ij} =Efecto aleatorio de los residuales; $\varepsilon_{ij} \sim NIID(0,\sigma^2_e)$

Pasos para el análisis

1. Se calcula el factor de corrección C
   $C=G^2/ \Sigma r_i$
2. Se calcula la suma de cuadrados debido al total
   $SCTOTAL = Y_{ij}^2 - C$
3. Se calcula la suma de cuadrados debido a los tratamientos de parcela grande
   $SCT = \Sigma T_{ij}^2/ r_i - C$
4. Análisis de varianza

Este análisis de varianza permite determinar la variabilidad debida al material experimental y la variabilidad ocasionada por los tratamientos. Estas variaciones son importantes para estimar cuál es el efecto de los tratamientos y cuál es la diferencia entre ellos.

La variación se mide a través del Cuadrado Medio, que es la división de la suma de cuadrados entre los grados de libertad. Las sumas de cuadrados del análisis de varianza pueden deducirse a partir del modelo lineal (Tabla 6.1)

Tabla 6.1. Análisis de varianza.

Fuente de variación	Grados de libertad	Suma de Cuadrados	Cuadrados medios	F
Tratamientos (T)	t-1	SCT	CMT=SCT/t-1	CMT/SCE
Error experimental	$\Sigma r_i - t$	SCE	CME=SCE/ tr-1=$\mathbf{s^2}$	
Total	Σr_{i-1}	SCTOTAL		

Prueba de hipótesis:
Ho: t1=t2=t3=…tn
H1: t1≠t2≠t3=…tn

Estadístico de prueba F:
F α (t-1, gl error)
Se rechaza Ho sí: Fcal ≥ F α (t-1, gl error)
Se acepta Ho sí: Fcal ≤ F α (t-1, gl error)

Separación de medias

Si la prueba de F rechaza la $H_0=\tau_1= \tau_j =$..... τ_j, la conclusión alternativa de que no todas las τ_j son iguales, en general es insuficiente para el experimento, quien requerirá los métodos adicionales de inferencia con respecto a los efectos de tratamientos.

El método Tukey proporciona una respuesta a esta clase de problemas, permitiendo al investigador descubrir cuales contrastes son significativamente importantes, cuando este se interesa por comparaciones simples de la forma $\tau_j - \tau_k$.

Otra técnica debido a Sheffé (1959), es más en general en su aplicación, permite probar la significancia de cualquier contraste.

Es muy común el empleo de la diferencia mínima significativa DMS, para comparaciones simples entre medias; está técnica es sólo una aproximación a una solución correcta al problema de las comparaciones entre medias, y debe usarse con precaución, y solo es válida en el caso de comparación de medias planeadas.

1. Diferencia Mínima significativa (DMS) o prueba de Duncan

$$DMS = t_{\alpha,\eta}\sqrt{\frac{2S^2}{r}}$$

Para su aplicación se hace uso de la distribución de T de Student} $\tau_j \neq \tau_k$

Donde:
$T_{\alpha,\eta}$ =t de student que separa α/2 del área en las colas de la distribución de t
H= (r-1) (t-1) = gl_E

Para decidir si $\tau_j \neq \tau_k$

$\left|\overline{Y_j} - \overline{Y_k}\right|$ ≥ DMS si considera $\tau_j \neq \tau_k$

$\left|\overline{Y_j} - \overline{Y_k}\right|$ < DMS entonces $\tau_i = \tau_k$

2. Método de Tukey (comparaciones múltiples)

Es valido solo para experimentos donde los tratamientos son igualmente repetidos. En este caso se cálcula la estadística de la diferencia significativa Honesta (DSH).

$$DSH = q_{\alpha,t,\eta}\sqrt{\frac{S^2}{r}}$$

Donde:
α= prueba de significación
η= gl Error =(r-1)(t-1)
t=# de medias que se compara

Si $\left|\overline{Y_j} - \overline{Y_k}\right|$ ≥ DSH se concluye que $\tau_j = \tau_k$

Si $\left|\overline{Y_j} - \overline{Y_k}\right|$ ≥ DSH se concluye que $\tau_j \neq \tau_k$

Cuando ỹ proviene de r_j y $\tilde{y}_k$ proviened de r_j observaciones, la estadística de prueba de Tukey aproximada es:

$$DSH = q_{\alpha,t,\eta}\sqrt{\frac{S^2}{2}\left(\frac{1}{rj}+\frac{1}{rk}\right)}$$

Empleado para comparaciones para pares de medias con rep. Variable

3. Método de Sheffé

Es capaz de considerar la prueba simultanea de q contrastes independientes al mismo tiempo. Cuando q=1 y se va a probar:

H0: $\psi = \psi_0$ Vs. H_1: $\psi \neq \psi_0$ } $\psi_0 = 0$

$$\psi = \sum_{j1}^{t} \lambda_j \tau_j \quad ; \qquad \psi = \sum_{j=1}^{t} \lambda_j \tau_j$$

$$\varepsilon = \sqrt{F_{\alpha;1,\eta}}\, x \sqrt{Var(\psi)} = \sqrt{F_{\alpha;(1,\eta)}}\, x \sqrt{\frac{S^2}{r} \sum_{j=1}^{t} \lambda_j{}^2}$$

Nivel de significancia gl error

Si $|\psi - \psi_0|$ ≥ ε ∴Ψ es significativamente diferente de Ψ_0

Si $|\psi - \psi_0|$ < ε ∴Ψ = Ψ_0

Estadística de prueba

$$\varepsilon = \sqrt{(t-1)F_{\alpha;t-1,\eta}}\, x \sqrt{\frac{2S^2}{r}}$$

H_o: $\tau_j = \tau_k$ Vs H_1: $\tau_j \neq \tau_k$

Si $|\overline{Y_j} - \overline{Y_k}|$ ≥ε se rechaza H_0

Si $\left|\overline{Y_j} - \overline{Y_k}\right|$ $< \varepsilon$ se acepta H_0

Normalmente se usa el 10 % de prueba (α=0.10)

Programa SAS

```
Data DCA;
Input  trat y;
Cards;
……
……
……
……
Proc glm;
Classes trat;
Model y = trat;
Means trat/duncan tukey;
Run;
```

Otra alternativa es hacer **comparaciones ortogonales** (LSMeans), para lo cual se aplica el siguiente programa:

```
Data DCA;
Input  trat y;
Cards;
……
……
……
Proc glm;
Class trat;
Model y = trat/ss3;
Lsmeans trat/pdiff;
Run;
```

Ejemplo 6.1. Elaboración de ensilaje de cáscara de banano

En una investigación realizada en la Universidad Estatal del Sur de Manabí, Jipijapa, se realizó un estudio de elaboración de ensilaje de cáscara de banano (*Musa paradisiaca*), utilizando microorganismos eficientes. La investigación tuvo como objetivo determinar el ensilaje y el microorganismo eficiente (ME) que presente mayor palatabilidad en el ganado bovino (consumo de alimento). Se utilizó un diseño experimental completamente aleatorio (DCA), donde los tratamientos utilizados fueron: testigo (melaza), Starlite (Ácidos húmicos), Microcompostic (bacterias descomponedoras) y Humiling 25 plus (Ácidos húmicos). Se midieron cinco variables de laboratorio (Humedad, ceniza, grasa, materia seca, proteína) y una variable de campo como el consumo de alimento. Los datos se reflejan en la Tabla 6.1.1.

Tabla 6.1.1. Datos de variables cinco variables evaluadas en la elaboración de ensilaje de cáscara de banano.

0	Tratamientos	% Humedad	Cenizas	Grasas	Proteína	MS	Consumo alimento Kg
1	Testigo	87.58	38.06	5.10	16.61	12.42	10
2	Testigo	85.07	25.92	5.57	16.25	14.93	10
3	Testigo	82,82	37,63	5,27	16,41	17,18	10
4	Testigo	85,72	29,17	5,60	16,32	14,28	10
1	Starley	84,18	29,19	5,34	18,97	15,82	9
2	Starley	85,55	30,85	5,48	18,89	14,45	10
3	Starley	83,18	38,19	5,61	18,88	16,82	10
4	Starley	86,72	33,91	5,18	18,98	13,28	9
1	Humiling	76,70	31,53	5,26	17,20	23,3	2
2	Humiling	80.13	24.96	5.27	17.12	19.87	2
3	Humiling	84.18	34.32	6.00	17.24	15.82	3
4	Humiling	86.41	24.81	5.79	17.20	13.59	4
1	Microcompo	77.05	25.89	5.15	17.69	22.95	4
2	Microcompo	82.88	29.58	5.26	17.84	17.12	3
3	Microcompo	77.37	29.95	5.88	17.66	22.63	4
4	Microcompo	82.70	28.94	5.08	17.70	17.3	3

Para el análisis de este experimento se deben seguir los siguientes pasos:

1. Realizar el análisis de normalidad
2. Realizar el análisis de homogeneidad de varianzas
3. Hacer el Análisis de varianza
4. Realizar la comparación de las medias de los tratamientos mediante la prueba múltiple de Tukey al $p<0.05$ de probabilidad
5. Interpretar los resultados

Con fines puramente didácticos el análisis manual de este experimento lo haremos considerando una sola variable. Las demás variables pueden ser analizadas de la misma manera por los lectores con fines de aplicación.

En el caso que nos toca analizar supondremos que los datos de la variable analizada, tiene una distribución normal y homogeneidad de varianzas. Por lo que seguiremos el análisis con los siguientes pasos.

1. Se debe crear y ordenar una nueva tabla (Tabla 6.1.2) en base a la Tabla 6.1.1., para el análisis. En este caso usaremos la variable porcentaje de humedad

Tabla 6.1.2. Tabla ordenada para la variable % de humedad.

Rep	Testigo	Starley	Humiling	Microcompo	Total	ΣY^2
1	87.58	84.18	76.70	77.05	325.51	105956.76
2	85.07	85.55	80.13	82.88	333.63	111308.977
3	82.82	83.18	84.18	77.37	327.55	107289.003
4	85.72	86.72	86.41	82.70	341.55	116656.403
Total	341.19	339.63	327.42	320.00	1328.24	
ΣX^2	116410.62	115348.54	107203.86	102400.00	441363.01	

2. Determinamos el coeficiente de corrección (C)

$C = (1328,24)^2/Sri = 441363.01/\mathbf{16} = 110263.84$

3. Suma de cuadrados totales (SCTotales)

$SCT = (87.58)^2 + \ldots + (82.70)^2 - C = \mathbf{182.61}$

4. Suma de cuadrados de tratamientos (SCT)

$SCT = \{[(341.19)^2 + (339.63)^2 + (327.42)^2 + (320.00)^2]/ ri\} - C = \{[(341.19)^2 + (339.63)^2 + (327.42)^2 + (320.00)^2]/4\} - C = \mathbf{76.91}$

5. Suma de Cuadrados del Error o Residual (SCE)

$SCE = SC\ Totales - SCT = 182.61 - 76.91 = \mathbf{105.70}$

6. Con los datos obtenidos hacer la Tabla de Análisis de Varianza (ANVA)

Tabla 6.1.3. Análisis de varianza para porcentaje de humedad.

FV	gl	SC	CM	Fc
Tratamiento	3 (t-1)	76.91	25.64	2.91ns
Error	12 (Σri-t)	105.70	8.81	
Total	15 (Σri-1)	182.61		

7. La Hipótesis de prueba es: Ho: T1 = T2 = T3 = T4

Se debe determinar cuál es el valor tabulado de $F_{\alpha.\ gl\ trat,\ 12gl\ error}$ y comparar este valor con la Fc

$F_{\alpha,\ 3,\ 12} = F_{0.05,\ 3,\ 12} = 3.490$

$F_{\alpha,\ 3,\ 12} = F_{0.01,\ 3,\ 12} = 5.953$

Conclusión:

Se acepta la Ho donde los tratamientos no son significativamente diferentes al $p<0.05$, lo que indica que todos los tratamientos tienen igual comportamiento estadísticamente en elaboración de ensilaje de cáscara de banano (*Musa paradisiaca*).

Sin embargo, un investigador experimentado, no quedará satisfecho con este resultado, debido a que el ANVA hace un análisis global utilizando un estadístico F, que no necesariamente discrimina los valores pequeños de diferencia. Por esta razón es recomendable hacer un análisis de comparación de medias, utilizando la prueba de Duncan u otra más estricta como la prueba múltiple de Tukey.

Con fines didácticos haremos el análisis de medias a través de la prueba de Tukey.

1. Lo primero que debemos hacer es obtener las medias de tratamiento. En el caso que estamos desarrollando serían los siguientes:

Tabla 6.1.4. Medias de tratamiento

Testigo	Starley	Humiling	Microcompo
85.30	84.91	81.86	80.00

2. Una vez obtenida las medias, debemos determinar la Diferencia Significativa Honesta (DSH), con la siguiente fórmula:

$$DSH = q_{\alpha,t,n}\sqrt{\frac{S^2}{r}}$$

Donde:

$q_{\alpha,\tau,\eta}$= Es un coeficiente obtenido de talas de q = 4.2

α = 0,05 o 0,01

τ = grado de libertad del tratamiento = 3

η = grado de libertad del error = 12

S^2 = CME = 8.81

r = Repetición = 4

DSH = 4.2 √ (8.81/4) = **6.23**

Para determinar si hay diferencias significativas entre los tratamientos, debemos determinar las diferencias entre las medias de tratamiento y comparar con el valor de DSH determinado.

Tabla 6.1.5. Comparación de las diferencias de medias de tratamiento.

		Testigo	Starley	Humiling	Microcompo
		85.30	**84.91**	**81.86**	**80.00**
Testigo	**85.30**	-	0.39ns	3.44ns	5.30ns
Starley	**84.91**		-	3.05ns	4.91ns
Humiling	**81.86**			-	1.85ns
Microcompo	**80.00**				-

Conclusión: Comparando el DSH con las medias, ninguno supera el valor de DSH=6.23. Por lo que no hubo diferencias significativas entre los tratamientos. Por lo que se concluye que todos los tratamientos tuvieron igual efecto estadísticamente en la elaboración de ensilaje de cáscara de banano (*Musa paradisiaca*).

Análisis en SAS

```
Data DCA;
Input trat $1-11 humedad 12-16;
Testigo  87.58
Testigo  85.07
Testigo  82.82
Testigo  85.72
.
.
.
Microcompo  77.05
Microcompo  82.88
Microcompo  77.37
Microcompo  82.70
;
Proc univariate plot normal;
Var humedad;
Proc discrim method = normal short pool = test;
Class trat;
Proc print;
Proc glm;
Classes  trat;
Model humedad = trat;
Means trat/ duncan tukey;
Run;
Proc glm;
Class trat;
Model humedad = trat/ss3;
Lsmeans trat/pdiff;
Run;
```

Ejemplo 6.2.

En el siguiente experimento se compararon los tiempos de coagulación de la sangre en cuatro grupos de ardillas con diferentes niveles de protrombina (un componente del plasma, necesario para la formación de coágulos). En los tratamientos, 100% es la sangre normal.

Niveles de protrombina			
20%	**30%**	**50%**	**100%**
34.4	25.0	20.9	19.7
27.3	23.2	22.2	21.7
65.0	45.2	27.8	21.1
31.3	26.4	19.5	18.5
48.5	26.8	20.1	16.0
38.4	32.7	22.1	20.2
40.5	28.8	19.7	

Análisis en SAS

```
Data DCA1;
Input trat niveles;
Cards;
1	34.4
1	27.3
1	65.0
1	31.3
1	48.5
1	38.4
1	40.5
2	25.0
2	23.2
2	45.2
2	26.4
2	26.8
2	32.7
2	28.8
3	20.9
3	22.2
3	27.8
3	19.5
3	20.1
3	22.1
3	19.7
4	19.7
4	21.7
4	21.1
4	18.5
4	16.0
4	20.2
;
Proc univariate plot normal;
Var niveles;
Proc discrim method = normal short pool = test;
Class trat;
Proc print;
Proc glm;
Classes  trat;
Model Niveles=trat;
Means trat/ duncan tukey;
Run;
Proc glm;
Class trat;
Model y = trat/ss3;
Lsmeans trat/pdiff;
Run;
```

Ejemplo 6.3.

Un Psicólogo estudió el efecto del sonido sobre el comportamiento de ciertos monos, el sonido fue aplicado a cuatro diferentes niveles de intensidad, cada uno de los niveles de sonido fueron aplicados a seis monos aleatoriamente y cada uno de ellos fueron colocados en jaulas independientes con alimento, cuando el mono alcanzaba el aliento el sonido fue aplicado y el tiempo en segundos hasta que el mono alcance nuevamente el alimento fue registrado.

Nivel	1	2	3	4	5	6	Total
1	6.1	7.3	5.4	6.3	6.6	7.2	
2	7.3	9.1	7.6	8.4	9.3	8.3	
3	10.5	12.4	10.6	13.0	10.5	9.0	
4	13.7	14.8	13.6	14.4	14.9	7.9	

Analisis en SAS

```
Data DCA2;
Input Nivel Tiempo;
Cards;
1	6.1
1	7.3
1	5.4
1	6.3
1	6.6
1	7.2
2	7.3
2	9.1
2	7.6
2	8.4
2	9.3
2	8.3
3	10.5
3	12.4
3	10.6
3	13.0
3	10.5
3	9.0
4	13.7
4	14.8
4	13.6
4	14.4
4	14.9
4	7.9
;
Proc univariate plot normal;
Var Nivel;
Proc discrim method = normal short pool = test;
Class Nivel;
Proc print;
Proc glm;
Classes  Nivel;
Model Tiempo=Nivel;
Means Nivel/ duncan tukey;
Run;
Proc glm;
Class Nivel;
Model y = Nivel/ss3;
Lsmeans Nivel/pdiff;
Run;
```

Ejemplo 6.3.

El experimento se llevó a cabo en las instalaciones de la Escuela de Medicina Veterinaria y Zootecnia de la Universidad Michoacana de San Nicolás de Hidalgo, ubicada en el km 10 de la carretera Morelia Zinapécuaro. Veintisiete toretes cebú, se distribuyeron en un diseño completamente al azar, a tres tratamientos que consistieron en la inclusión del 15 (P15), 25 (P25) y 35 % (P35) de pollinaza en base seca como ingrediente de la dieta. La unidad experimental fue el corral, siendo tres en cada tratamiento.

La melaza se diluyó en un 10 % de agua y se mezcló diariamente con el alimento ofrecido. La duración total del trabajo fue de 118 días, de los cuales 20 fueron para la adaptación a las dietas, corrales y para el manejo de los animales que consistió en vacunación, desparasitación, identificación y aplicación de vitaminas A, D y E por vía intramuscular. Las variaciones en el peso de los animales se registraron en períodos catorcenales. Los resultados obtenidos se sometieron a un análisis de varianza y a la prueba de diferencia mínima significativa para comparación de medias.

Variables de respuesta

Comparar Ganancia de peso (Kg)

Datos del experimento

	Corral 1					Corral 2					Corral 3				
Trat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Peso Inicial															
P15	121	210	200	222	209	315	312	311	309	300	120	123	219	118	216
P25	112	116	112	221	211	310	314	301	305	307	225	115	117	122	321
P35	205	215	318	213	219	316	309	314	316	300	223	319	218	316	120
Peso final															
1	447	446	445	446	444	432	435	436	432	431	417	411	415	416	419
2	446	445	447	444	443	433	433	430	432	430	415	413	412	411	418
3	442	446	448	449	446	434	435	437	438	435	416	417	416	415	413

Programa SAS

```
Data DCA3;
Input trat Pi Pf; GP=Pf-Pi;
Cards;
;
 1          121        447
 1          210        446
 1          200        445
 1          222        446
 1          209        444
 1          315        432
 1          312        435
 1          311        436
 1          309        432
 1          300        431
 1          120        417
 1          123        411
 1          219        415
 1          118        416
 1          216        419
 2          112        446
 2          116        445
 2          112        447
 2          221        444
 2          211        443
 2          310        433
 2          314        433
 2          301        430
 2          305        432
 2          307        430
 2          225        415
 2          115        413
 2          117        412
 2          122        411
 2          321        418
 3          205        442
 3          215        446
 3          318        448
 3          213        449
 3          219        446
 3          316        434
 3          309        435
 3          314        437
 3          316        438
 3          300        435
 3          223        416
 3          319        417
 3          218        416
 3          316        415
 3          120        413

Proc univariate plot normal;
Var gp;
Proc discrim method = normal short pool = test;
Class trat;
Proc print
Proc glm;
Classes trat;
Model gp = trat;
Means trat/duncan tukey;
Run;
```

Otra alternativa es hacer **comparaciones ortogonales** (lsmeans) , para lo cual se aplica el siguiente programa:

```
Proc glm;
Class trat;
Model y = trat/ss3;
Lsmeans trat/pdiff;
Run;
```

UNIDAD 7

DISEÑO DE BLOQUES COMPLETAMENTE ALEATORIOS (DBCA)

Julio Gabriel Ortega

Alfredo Valverde Lucio

Máximo Vera Tumbaco

José Alcívar Cobeña

Descripción

Es uno de los diseños más utilizados en experimentación agrícola y pecuario. Este diseño es utilizado cuando el material experimental, campo agrícola, invernadero, calmas de almacigo, animales etc., presentan una fuente de variabilidad conocida, factible de evaluar y de deducir el error experimental. Con ello se logra disminuir el error experimental, lo que incrementa la precisión en la comparación entre tratamientos. Recibe el nombre de bloque completo al azar, porque el material experimental se fracciona en bloques o en estrato uniformes dentro de si pero diferente entre sí. Todos los tratamientos están presentes y distribuidos al azar en cada uno de los bloques. La distribución de los tratamientos al azar se realiza independientemente en cada bloque. Aunque por lo general los tratamientos se presentan una sola vez en cada bloque, es posible que un tratamiento de interés este duplicado en cada bloque.

Durante la conducción del experimento se debe tener especial cuidado para que no se pierdan unidades experimentales. Ya que toda pérdida de información es irreparable. Pero si pese a todas las precauciones ello ocurre, el análisis estadístico no se mayormente afectado. Para el efecto se estima el valor de la unidad pérdida y luego se realiza el análisis estadístico en forma regular. Sin embargo, la estimación del valor perdido. Solo tiene el propósito de resolver el cálculo estadístico sin lograr recuperar la información pérdida.

El número de tratamientos a tener en un bloque depende de la homogeneidad existente dentro del bloque. La práctica se recomienda que no sea muy elevado. Si se observa demasiada variabilidad puede utilizarse el diseño de bloques incompletos.

La forma del bloque vario de acuerda a las particularidades del área experimental; sin embargo, es recomendable que los bloques tengan forma cuadrada, mientras que las parcelas deben tener forma rectangular y paralela a la fuente de variabilidad. Los bloques pueden estar juntos o separados. Lo importante es que exista homogeneidad dentro de cada bloque, aunque sean diferentes entre sí.

El número recomendable entre tratamientos por bloque varia de 6 a 24 y el número recomendable de bloques varía de 3 a 8.

Caracteristiscas

- En un bloque completo si cada tratamiento aparece representado en el bloque
- Cualquier bloque en el que se omite al menos un tratamiento se llama bloque incompleto.

- Todos los bloques son del mismo tamaño y comprenden t unidades experimentales ensayando cada tratamiento una vez dentro de cada bloque.
- Se proyectan r bloques completos donde hay rt unidades experimentales.
- Los tratamientos se alojan al azar en forma independiente dentro de cada bloque.

Cuando aplicar este diseño

- Se utiliza cuando el material experimental comprende un componente de variación sistemática.
- Unidades experimentales se agrupan en r conjuntos más o menos homogéneos dentro de ellos, pero heterogéneos entre sí.

(-)

Pendiente

R1t1	R1t2	R1t3	R1t4	R1t5	R1t6	R1t7	R1t8	R1t9	Bloque 1
R2t9	R2t3	R2t7	R2t2	R2t8	R2t6	R2t5	R2t1	R2t4	Bloque 2
R3t3	R3t1	R4t5	R8t8	R7t7	R3t9	R3t4	R5t6	R3t2	Bloque 3

(+)

Modelo aditivo lineal

$$Y_{ijk} = \mu m + \beta_i + \tau_j + \varepsilon_{ijk}$$

Donde:
I= 1,2,.....r
R=1,2,.....t
$E(\varepsilon_{ij})=o$, $E(\varepsilon_{ij}{}^2) =\sigma^2$, $E(\varepsilon_{ij}\,\varepsilon_{i'j'})=0$

Donde:
Y_{ijk} = Características del experimento
β_i= Efecto del Bloque i
τ_j= Efecto de tratamiento j
ε_{ij}= Error experimental

Bloques y tratamientos son factores ortogonales

$$MEll(\sum_{j=i}^{t} \lambda_j \tau_j) = \sum_{j=1}^{t} \lambda_j \overline{y_i}$$

2. Calculo del factor de corrección

C=G2/rt } û=C

3. Se obtiene la suma de cuadrados totales corregida por la media

$$SCtotal = \sum_{i=1}^{b} \sum_{j=1}^{t} Y_{ij}{}^2 - C$$

Sé obtiene la suma de cuadrados debido a los bloques

$$SCB = \frac{\sum_{i=1}^{t} B_i^{\,2}}{t} - C$$

4. Se obtiene la suma de cuadrados debido a los tratamientos
5. Se obtiene la SCE

   SCE=Sctotal – SCB – SCT
6. Comparación de medias

$$SCT = \frac{\sum_{j=1}^{b}}{r} - C$$

Prueba de Tukey (Diferencia significativamente honesta DSH)

$$DSH = q_{0.05,t-1,\eta}\sqrt{\frac{S^2}{r}}$$

Aplicación en el sas

```
Data CA;
Input trat y;
Cards;
1    Y1
1    Y2
.
.
.
.
n    Yn
proc print;
proc anova;
classes trat;
model y=trat;
means trat/tukey;
run;
```

Eficiencia de los diseños experimentales

En la experimentación agrícola, emplear un diseño experimental sin evaluar su eficiencia relativa con respecto a otro menos complejo se manifiesta como un problema recurrente Ruiz-Ramírez (2010).

Para evaluar si el diseño experimental empleado es el correcto, se sugiere calcular la eficiencia relativa (ER). Para ello se prueba la ER de ese diseño experimental con respecto a uno menos complejo. En el cálculo de la ER se utiliza la varianza del experimento, conocida como el cuadrado medio del error, obtenido mediante el análisis de varianza, donde se registra la variabilidad de factores no controlados, como son la calidad de las semillas y los factores físicos en las unidades experimentales, cuando éstas se establecen en el campo o sitio experimental.

Una manera de verificar la calidad de un experimento es a través del coeficiente de variación (CV), del que se espera no exceda de 20% para experimentos uniformes que se realizan en cultivos tales como maíz, trigo y caña de azúcar (Ruiz-Ramírez 2010); sin embargo, existen otros cultivos con mucha variabilidad reportada principalmente para el rendimiento agrícola, como ocurre con las hortalizas, los árboles frutales y los maderables, donde es normal que el CV exceda el 50%.

No obstante, tanto al CV como a la ER se les ha otorgado poca importancia, y los escasos estudios sobre la ER únicamente informan el valor de la probabilidad para probar el efecto significativo de los tratamientos sobre la variable dependiente o de respuesta.

Para calcular la ER con respecto a los 2 diseños experimentales (DBA contra DCA) más utilizados, se estima el cuadrado medio del error en un diseño completamente al azar, usando la información del DBA.

$$\hat{CM}_E(DCA) = \frac{f_b CM_{Bloq} + (f_t + f_e) CM_E}{f_b + f_t + f_e}$$

donde:
f_b = grados de libertad de bloques
f_t = grados de libertad de los tratamientos
f_e = f grados de libertad del error experimental
CM_{Bloq} = cuadrado medio de bloques
CM_E = cuadrado medio del error experimental

Una vez obtenido el cuadrado medio del error del diseño completamente al azar, el cálculo de la eficiencia relativa está dada por:

$$ER = \frac{(f_1+1)(f_2+3)\ \mathrm{CM}_E(DCA)}{(f_2-1)(f_1+3)\ CM_E(DBA)} x100$$

donde:

$fe = f1$

$f_2 = f_b + f_e$

El criterio que se sigue para determinar que el DBA fue el apropiado, ocurre cuando la ER es mayor al 100%, en caso contrario, el diseño completamente al azar (DCA) fue más eficiente y se requiere analizar nuevamente el experimento con un DCA.

Un indicador de la calidad de los experimentos es el CV, y se utiliza en los cultivos de algodón y de la caña de azúcar.

En el Cuadro 2 se presenta la clasificación de la calidad aplicada a experimentos de cultivos anuales y se utilizó para describir la calidad de experimentos del cultivo de caña de azúcar. Para ello se elaboró una tabla de frecuencias, útil en la representación de los resultados.

Tabla 7.1. Precisión de la calidad de los experimentos en el cultivo anual (Ruíz-Ramirez 2010)

Rango de coeficientes de variación (CV)	**Precisión**
$5 < CV \leq 10$	Muy buena
$10 < CV \leq 15$	Buena
$15 < CV \leq 20$	Regular
$20 < CV \leq 25$	Mala
$CV > 25$	Muy mala

Ejemplo 7.1.

Para la ilustración didáctica del diseño de Bloques completamente aleatorios o bloques completo al azar (DBCA), tomaremos el ejemplo 6.1. de este documento, pero analizaremos otra variable como el contenido de ceniza en porcentaje.

1. Se debe crear y ordenar una tabla para análisis. En este caso usaremos la variable porcentaje de ceniza

Tabla 7.1.1. Tabla ordenada para la variable % de ceniza.

Bloque	Testigo	Starley	Humiling	Microcompo	Total	ΣY^2
1	38.06	29.19	31.53	25.89	124.67	15542.846
2	25.92	30.85	24.96	29.58	111.30	12388.1166
3	37.63	38.19	34.32	29.95	140.09	19626.517
4	29.17	33.91	24.81	28.94	116.83	13649.7569
Total	**130.78**	**132.14**	**115.62**	**114.36**	**492.90**	
ΣX^2	**17103.76**	**17459.92**	**13368.30**	**13078.44**	**61010.42**	

2. Determinamos el coeficiente de corrección (C)

$C = (492.90)^2/rxt = 61010.42/4x4 =$ **15184.38**

3. Suma de cuadrados totales (SCTotal)

$SCTotal = (38.06)^2 + \ldots + (28.94)^2 - C =$ **305.38**

4. Suma de cuadrados de tratamientos (SCT)

$SCT = \{[(130.78)^2 + (132.14)^2 + (115.62)^2 + (114.36)^2]/ r\} - C = \{[(341.19)^2 + (339.63)^2 + (327.42)^2 + (320.00)^2]/4\} - C =$ **68.22**

5. Suma de cuadrados de bloques (SCB)

$SCB = \{[(124.67)^2 + (111.30)^2 + (140.09)^2 + (116.83)^2]/t\} - C = \{[(124.67)^2 + (111.30)^2 + (140.09)^2 + (116.83)^2]/4\} - C =$ **117.43**

6. Suma de Cuadrados del Error o Residual (SCE)

$SCE = SCTotal - SCT - SCB = 305.38 - 68.22 - 117.43 =$ **237.16**

7. Con los datos obtenidos hacer la Tabla de Análisis de Varianza (ANVA)

Tabla 7.1.2. Análisis de varianza para porcentaje de ceniza.

FV	gl	SC	CM	Fc
Tratamiento	3 (t-1)	68.22	22.74	0.86ns
Bloque	3 (r-1)	117.43	39.14	1.49ns
Error	9 (t-1)(r-1)	237.16	26.35	
Total	15 (rt-1)	305.38		

8. La Hipótesis de prueba es: Ho: T1 = T2 = T3 = T4

Se debe determinar cuál es el valor tabulado de $F_{\alpha.\ gl\ trat,\ 12gl\ error}$ y comparar este valor con la Fc

$F_{\alpha,\ 3,\ 12} = F_{0.05,\ 3,\ 9} = 3.863$

$F_{\alpha,\ 3,\ 12} = F_{0.01,\ 3,\ 9} = 6.992$

Conclusión:

Se acepta la Ho los tratamientos no son significativamente diferentes al $p<0.05$, lo que indica que todos los tratamientos tienen igual comportamiento estadísticamente en elaboración de ensilaje de cáscara de banano (*Musa paradisiaca*).

Un investigador experimentado, no quedará satisfecho con este resultado, debido a que el ANVA hace un análisis utilizando un estadístico F, que no necesariamente discrimina los valores pequeños de diferencia. Por esta razón es recomendable hacer un análisis de comparación de medias, utilizando la prueba de Duncan u otra más estricta como la prueba múltiple de Tukey.

Con fines didácticos haremos el análisis de medias.

1. Lo primero que debemos hacer es obtener las medias de tratamiento. En el caso que estamos desarrollando serían los siguientes:

Tabla 7.1.3. Medias de tratamiento

Testigo	Starley	Humiling	Microcompo
32.70	33.03	28.91	28.59

2. Una vez obtenida las medias, debemos determinar la Diferencia Significativa Honesta (DSH), con la siguiente fórmula:

$$DSH = q_{\alpha,t,\eta}\sqrt{\frac{S^2}{r}}$$

Donde:
$q_{\alpha,\tau,\eta}$= Es un coeficiente obtenido de talas de q = **6.42**
α = 0.05 o 0.01
τ = grado de libertad del tratamiento = 3
η = grado de libertad del error = 9
S^2 = CME =26.35
r = Repetición = 4

DSH = 6.42 √ (26.35/4) = **10.78**

Para determinar si hay diferencias significativas entre los tratamientos, debemos determinar las diferencias entre las medias de tratamiento y comparar con el valor de DSH determinado.

Tabla 6.1.5. Comparación de las diferencias de medias de tratamiento.

		Testigo	Starley	Humiling	Microcompo
		32.70	**33.03**	**28.91**	**28.59**
Testigo	**32.70**	-	-0.33ns	3.79ns	4.11ns
Starley	**33.03**		-	4.12ns	4.44ns
Humiling	**28.91**			-	0.32ns
Microcompo	**28.59**				-

Conclusión: Comparando el DSH con las medias, ninguno supera el valor de DSH=10.78. Por lo que No hubo diferencias significativas entre los tratamientos. Por lo que se concluye que todos los tratamientos tuvieron igual efecto estadísticamente en la elaboración de ensilaje de cáscara de banano (*Musa paradisiaca*).

Análisis mediante el SAS University

```
Data DBCA;
Input blo 1-2 trat $3-13 ceniza 14-18;
1 Testigo  38.06
2 Testigo  25.92
3 Testigo  37.63
4 Testigo  29.17
.
.
.
1 Microcompo  77.05
2 Microcompo  82.88
3 Microcompo  77.37
4 Microcompo  82.70
;
Proc univariate plot normal;
Var humedad;
Proc discrim method = normal short pool = test;
Class ceniza;
Proc print;
Proc glm;
Classes  blo trat;
Model ceniza  = trat blo;
Means trat/ duncan tukey;
Run;
Proc glm;
Class blo trat;
Model ceniza = trat/ss3;
Lsmeans trat/pdiff;
Run;
```

Ejemplo 7.1.

Se van a probar diez raciones respecto a sus diferencias en la engorda de novillos. Se dispon de 40 novillos para la experimentación, que se distribuyen en 4 bloques (10 novillos por bloque) cn base en sus pesos al iniciar la prueba de engorda. Los novillos más pesados se agruparon en un bloque, en el otro se agruparon los 10 siguientes más pesados y así sucesivamente. Los 10 tratamientos (raciones) se asignaron al azar dentro de cada bloque. A continuación se presentan los resultados del experimento (Tabla 7.1):

Tabla 7.1. Incrementos de peso en Kg de 40 novillos alimentados con diferentes raciones.

Tratamiento	**Bloques**			
	1	**2**	**3**	**4**
A	0.9	1.4	1.4	2.3
B	2.3	1.8	2.3	2.3
C	3.6	3.2	4.5	4.1
D	2.7	2.3	2.3	0.9
E	0.5	0.9	0.5	0.9
F	1.4	2.3	3.2	3.6
G	3.6	3.6	3.2	3.6
H	2.7	5.4	0.9	2.3
I	1.8	2.3	2.7	1.4
J	1.8	1.8	0.9	1.4

```
Data DBCA;
Input Trat Rep Peso;
Cards;
A       1       0.9
B       1       2.3
C       1       3.6
D       1       2.7
E       1       0.5
F       1       1.4
G       1       3.6
H       1       2.7
I       1       1.8
J       1       1.8
A       2       1.4
B       2       1.8
C       2       3.2
D       2       2.3
E       2       0.9
F       2       2.3
G       2       3.6
H       2       5.4
I       2       2.3
J       2       1.8
A       3       1.4
B       3       2.3
C       3       4.5
D       3       2.3
E       3       0.5
F       3       3.2
G       3       3.2
H       3       0.9
I       3       2.7
J       3       0.9
A       4       2.3
B       4       2.3
C       4       4.1
D       4       0.9
E       4       0.9
F       4       3.6
G       4       3.6
H       4       2.3
I       4       1.4
J       4       1.4
;
```

```
Proc univariate plot normal;
Var peso;
Proc discrim method = normal short pool = test;
Class trat;
Proc print;
Proc glm;
Classes Trat Rep;
Model peso = rep trat;;
Means trat/duncan tukey;
Run;
Proc glm;
Class trat;
Model peso = rep trat/ss3;
Lsmeans rep trat/pdiff;
Run;
```

UNIDAD 7.1
DISEÑO DE BLOQUES COMPLETAMENTE ALEATORIOS CON MUESTREO

Julio Gabriel Ortega

Introducción

La situación más práctica es aquella donde se forman los elementos muéstrales sobre cada unidad experimental.

Modelo aditivo lineal

$Y_{ijk} = \mu + \beta_i + \tau_j + \rho_{jj} + \eta_{ijk}$

$i = 1,2,.....,r$

$j = 1,2,.....,t$

$k = 1,2,.....,s$

ρ_j= Error experimental

η_{ijk} = Error de muestreo

$E(e_{ij})=o, \quad E(e_{ij}^2)=\sigma^2, \quad E(e_{ij}\ e_{i'j'}) = 0$

$E(\eta_{ijk})=o, \quad E(\eta_{ijk})^2=\sigma^2, \quad E(\eta_{ijk}\ \eta_{i'j'k'}) = 0$

Pasos para el análisis

1. Cálculo para el factor de corrección, C

   $C=G^2/rts$

2. Cálculo de la suma de cuadrados total, SCtotal

$$SCtotal = \sum_{ijk} y_{ijk}^{\ 2} - C$$

3. Cálculo de la suma de cuadrados debido a bloques, SCB

$$SCB = \frac{\sum_{t=1}^{r} B_i^{\ 2}}{ts} - C$$

$$SCT = \frac{\sum_{j=1}^{t} T_j^{\ 2}}{s} - C$$

4. Cálculo de la suma de cuadrados debido a tratamientos, SCT

5. Cálculo de la suma de cuadrados debido al error experimental, SCE

$$SCE = \frac{\sum_{i=1}^{r}\sum_{j=1}^{t} Y_{ij}^{2}}{\sum_{j=1}^{s}} - C - SCB - SCT$$

6. Cálculo de la suma de cuadrados debido a submuestreo, SCS

SCS= SCtotal – SCB – SCT – SCE

7. Construcción de la tabla de análisis de varianza

F. V.	**gl**	**SC**	**CM**	**Fcal**
Bloques	r-1	$SCB = \sum_{i=1}^{r} \frac{B_i^{2}}{ts} - \frac{G^{2}}{rts}$	CMB=SCB/r-1	CMB/S^2
Tratamientos	t-1	$SCT = \sum_{j=1}^{t} \frac{T_j^{2}}{rs} - \frac{G^{2}}{rts}$	CMT=SCT/t-1	CMT/S^2
Error experimental	(r-1)(t-1)	SCE	CME=SCE/(r-1)(t-1)=S^2	
Error de muestreo	rt(s-1)	SCS	CMS= SCS/ rt(s-1)	
Total	rts-1	$\sum_{ijk}^{t} Y_{ijk}^{2} - \frac{G^{2}}{rts}$		

Si Fcal= $CMT/S^2 > F\alpha$ (t-1,)η Se rechaza la H_0

Si Fcal= $CMT/S^2 < F\alpha$ (t-1,)η Se acepta la H_0

Comparación de medias

Diferencia mínima significativa (DMS)

$$DMS = t_{\alpha,\eta}\sqrt{\frac{2S^2}{r}}$$

T de student que separa $\alpha/2$ en el área de las colas.

η= (r-1)(t-1) grados de libertadad del error experimental
S^2 = CME
R= # repeticiones

Si $\left|\overline{Y_j - Y_K}\right| > DMS \Rightarrow \therefore \left|\tau_j - \tau_k\right|$ Es significativamente de cero

Si $\left|\overline{Y_j - Y_K}\right| < DMS \Rightarrow se-acepta\left|\tau_j - \tau_k\right| = 0$

Método de Tukey

En lugar de utilizar t como base de comparación, se usa la distribución de rango estandarizado (tabla). Para su empleo se calcula la DSH (Diferencia Estadística honesta)
Donde:
$q_{\alpha;t;\eta=}$ Valor tabulado

$$DSH = q_{\alpha;t;\eta}\sqrt{\frac{S^2}{r}}$$

α= Nivel de significancia
η= (r-1)(t-1) gl de S^2
S^2= CME
r=Bloques completos

$H_o : \tau_j = \tau_k$

$H_1 : \tau_j \neq \tau_k$

$\left|\overline{Y_j - Y}\right|$ $\geq$ DSH $\therefore$ $\tau j \neq \tau k$

Si

Método de Scheffé

Se basa en la distribución F basta calcular la estadística

$$\varepsilon = \sqrt{(t-1)F\alpha; t-1; \eta x \sqrt{\frac{2S^2}{r}}}$$

- En general el método de Sheffé es más riguroso que el método de Tukey.
- Scheffé se utiliza al 10% de significancia.
- Experimentos de mejoramiento de plantas $\Longrightarrow$ Tukey
- Experimentos de fertilizantes $\Longrightarrow$ Scheffé

Se rechaza la hipótesis nula

Ejemplo 7.1.1.

Los datos en la siguiente clasificación de dos factores son tiempos, en segundos, de varios corredores en una distancia de 1.5 millas (Tabla 7.1.1.1). Los corredores se clasifican en tres grupos de medidas y tres categorías de estado físico, siendo estas últimas funciones de varias variables. Completen el análisis de varianza. Probar la hipótesis de que las medias de la población de los tiempos de carrera no dependen de las categorías de estado físico.

Tabla 7.1.1.1. Datos del estado físico y la edad de corredores en una distancia de 1.5 millas.

		Estado Físico (Rep)		
		Bajo	**Medio**	**Alto**
Edad (Trat)	**40**	669	602	527
		671	603	547
	50	775	684	571
		821	687	573
	60	1009	824	688
		1060	828	713

Programa en SAS

```
Data DBCAS;
Input rep trat M Y;
Cards;
1    1    1    669
1    1    2    671
1    2    1    775
1    2    2    821
1    3    1    1009
1    3    2    1060
2    1    1    602
2    1    2    603
2    2    1    684
2    2    2    687
2    3    1    824
2    3    2    828
3    1    1    527
3    1    2    547
3    2    1    571
3    2    2    573
3    3    1    688
3    3    2    713
```

```
;
Proc univariate plot normal;
Var Y;
Proc discrim method = normal short pool = test;
Class trat;
Proc print;
Proc glm;
Classes rep trat M;
Model Y=rep trat rep*trat;
Means Trat/Duncan Tukey;
Run;
```

UNIDAD 8

DISEÑO EXPERIMENTAL CUADRADO LATINO (DCL)

Julio Gabriel Ortega

Carlos Castro Piguave

José Alcívar Cobeña

Características

- Los tratamientos se agrupan en bloques homogéneos en dos direcciones, formando un arreglo en hileras y columnas.
- El número total de tratamientos, t, es igual al número de hileras o columnas y es un entero igual y mayor que 2, siendo las unidades experimentales, un cuadrado perfecto: T^2
- Este diseño es característico porque un tratamiento cualquiera aparece representado exactamente una vez en la misma hilera o en la misma columna.
- La particularidad del diseño, de construir bloques completos en el sentido de las hileras y de las columnas, permite absorber e ambos sentidos, la variabilidad del material experimental.

Si t=5

A	B	C	D	E
E	A	B	C	D
D	E	A	B	C
C	D	E	A	B
B	C	D	E	A

Modelo aditivo lineal

$Yijk = \mu + \beta i + \rho j + \tau k\,(ij) + \xi ijk$

Donde:

i = 1, 2,… 10 filas
j = 1, 2…..8 columnas
k= 1, 2…..8 tratamientos

Yij= Valor de una variable de respuesta observada en el j-ésimo híbrido evaluado en el i-ésimo bloque
μ = Media general.
βi = Efecto fijo de la i- ésima fila
ρj = Efecto fijo de j- ésima columna
τk (ij) = Efecto fijo del k-ésimo tratamiento
$\xi_{ij}k$= Efecto aleatorio de los residuales $\xi ijk \sim$ NIID, (0, $\sigma e2$).

Ejemplo 8.1.

Un experimento con 3 novillos de un año en una vaquería se efectuó en un cuadrado latino. Los tratamientos fueron 3 raciones. Cada animal recibió las 3 raciones sucesivamente, con una semana cada una. La variable aquí es y= libras de materia seca consumidas por 100 lb de peso corporal. Los datos se presentan a continuación. Los números dentro paréntesis indican tratamientos. Los tratamientos fueron:

(1) Forma de alfalfa
(2) Maíz ensilado
(3) Pastillas de pasto azul

Tabla 8.1.1. Datos de tratamientos (raciones) y libras de materia seca.

	Animales			
Semana	**1**	**2**	**3**	**Hilera H.j**
1	2.7 (1)	2.6 (2)	1.9 (3)	7.2
2	2.1 (2)	0.2 (3)	2.3 (1)	4.6
3	1.9 (3)	2.1 (1)	2.4 (2)	6.4
Columna	6.7	4.9	6.6	G=18.2

Tabla 8.1.2. Resumen de los totales y promedio de tratamientos.

Alimento		Total	Promedio
Alfalfa	1	7.1	2.37
Maíz	2	7.1	2.37
Pastillas	3	4.0	1.33
Total		18.2	

Pasos para el análisis

1) Cálculo de la Suma de Cuadrados de Columnas (SCC)

2) Cálculo de la Suma de Cuadrados de Hileras (SCH)

$$SCC = \frac{\sum_{i=1} C_{i.}^{\ 2}}{t} - C = \frac{6.7^2 + 4.9^2 + 6.6^2}{3} - C = 34.49 - 36.80 = 0.69$$

3) Cálculo de la Suma de Cuadrados Total (SCT)

$$SCH = \frac{\sum_{j=1} H_{.j}^{\ 2}}{t} - C = \frac{7.2^2 + 4.6^2 + 6.4^2}{3} - C = \frac{113.96}{3} - C = 37.99 - 36.80 = 1.19$$

$$SCT = \frac{\sum_{k=1} T_k^{\ 2}}{t} - C = \frac{7.1^2 + 7.1^2 + 4.0^2}{3} - C = 38.94 - C = 37.99 - 36.80 = 2.14$$

4) Cálculo de la suma de cuadrados del Error (SCE)

$$SCE = SCTotal - SCC - SCH - SCT = 4.38 - 0.69 - 1.19 - 2.14 = 0.36$$

5) Análisis de varianza

Tabla 8.1.3. Análisis de varianza para materia seca.

FV	gl	SC	CM	Fc
Hileras	t-1=3-1=2	1.19	0.591	3.15
Columnas	t-1=3-1=2	0.69	0.341	1.92
Tratamientos	t-1=3-1=2	2.14	1.070	5.69 NS
Error	(t-1)(t-2)=2	0.36	$0.18=S^2$	
Total	T^2-1=8	4.38		

$F_{\alpha,t-1,\eta}=F_{0.05,2,2}=19.00$
$F_{\alpha,t-1,\eta}=F_{0.01,2,2}=99.00$

Conclusión : No hay diferencias significativas entre hileras, columnas y tratamientos

Prueba de medias de (Tukey)

$DSH_{0,05}$ = 19.00 √ 0.18/3 = **4.65**

$DSH_{0,01}$ = 99.00 √ 0.18/3 = **24.25**

X	Pastillas Pasto Azul 1.33	Ensilaje Maíz 2.37	Alfalfa 2.37
1.33	-	1.04 ns	1.04 ns
2.37		-	0.00
2.37			-

Conclusión: No existen diferencias significativas entre pastillas de pasto azul y ensilaje de maíz. Lo mismo se observa entre pastillas de pasto azul y alfalfa.

Programa SAS

```
Data latice;
Input h c trat y;
Cards;
1    1    1    2.7
1    2    2    2.6
1    3    3    1.9
2    1    2    2.1
2    2    3    0.2
2    3    1    2.3
3    1    3    1.9
3    2    1    2.1
3    3    2    2.4
;
Proc univariate plot normal;
Var y;
Proc discrim method = normal short pool = test;
Class trat;
Proc print;
Proc glm;
Classes h c trat y;
Model y = h c trat;
Menas trat/duncan tukey; Run;
```

Ejemplo 8.2.

Se realizó un experimento para asegurar las resistencias relativas a la abrasión de 4 tipos de pieles (A, B, C, D). Se Uso una máquina en la cual se probaron las muestras en una cualquiera de 4 posiciones. Puesto que se conoce, que diferentes ejecuciones del experimento (reproducciones) dan resultados variables, se decidió hacer 4 ejecuciones del mismo. Se utilizó un diseño Cuadrado Latino y se obtuvieron los siguientes resultados. Analice e interprete los datos.

Ejecución (Hileras)	**Posición (Columnas)**			
	1	**2**	**3**	**4**
1	118 (B)	136 (D)	168 (A)	135 (C)
2	127 (D)	141 (B)	129 (C)	151 (A)
3	174 (A)	173 (C)	126 (B)	134 (D)
4	130 (C)	170 (A)	125 (D)	95 (B)

```
Data Latice;
Input H C Trat Resis;
Cards;
1    1    A    174
2    2    A    170
3    3    A    168
4    4    A    151
1    1    B    118
2    2    B    141
3    3    B    126
4    4    B    95
1    1    C    130
2    2    C    173
3    3    C    129
4    4    C    135
1    1    D    127
2    2    D    136
3    3    D    125
4    4    D    134
Proc print;
Proc glm;
Classes H C trat;
Model Resis=H C trat;
Means trat/duncan Tukey;
```

Ejemplo 8.3.

Una investigación se realizó para determinar el efecto de diferentes niveles de insulina sobre la cantidad de azúcar en la sangre de los ratones. Para esta investigación se disponía de 4 grupos de ratones con diferentes características y solo 4 tratamientos se podía aplicar por razones de tiempo y análisis por cada día. Cada uno de cuatro niveles de concentración de insulina (150, 300, 600, 1200) microgramos fueron aplicados a 6 ratones en cada uno de 4 días, el promedio de 6 ratones en el cambio del nivel de azúcar en la sangre fueron las siguientes.

Días	1	2	3	4	Yi..
1	300 (-4.5)	1200 (92.33)	600 (59.83)	150 (45.00)	102.66
2	600 (91.83)	150 (-48.33)	1200 (108.99)	300 (89.00)	301.49
3	1200 (86.16)	300 (78.16)	150 (24.17)	600 (101.0)	84.83
4	150 (-0.17)	600 (68.83)	300 (25.17)	1200 (177.17)	271.00
Y.J.	173.32	34.67	229.82	322.17	739.98

En el Sas

```
Data CL;
Input fil col trat azu;
Cards;

   1      1      300       -4.5
   1      2      1200      92.33
   1      3      600       59.83
   1      4      150       -45.00
   2      1      600       91.83
   2      2      150       -48.33
   2      3      1200      108.99
   2      4      300       89.00
   3      1      1200      86.16
   3      2      300       78.16
   3      3      150       24.17
   3      4      600       101.00
   4      1      150       -0.17
   4      2      600       68.83
   4      3      300       25.17
   4      4      1200      177.17

Proc glm;
Class fil col trat;
Model azu=fil col trat/ss3;
Random fil col/test;
Lsmeans trat/pdiff;
Run;
Proc glm;
Class fil  col;
Model azu=fil col trat    trat*trat    tra*trat*trat;
Random fil col;
Run;
```

UNIDAD 8.1.
DISEÑO DE FILA – COLUMNA

Julio Gabriel Ortega

Alfredo Valverde Lucio

Caracterìsticas

En algunas ocasiones se está interesado en estudiar la influencia de dos (o más) factores tratamiento, para ello se hace un diseño de filas - columnas. En este modelo es importante estudiar la posible interacción entre los dos factores. Si en cada casilla se tiene una única observación no es posible estudiar la interacción entre los dos factores, para hacerlo hay que replicar el modelo, esto es, obtener k observaciones en cada casilla, donde k es el número de réplicas. El modelo matemático de este diseño es:

$$y_{ijkl} = \mu + \lambda_i + \varphi_{j(i)} + \delta_{k(i)} + \tau_l + \gamma_{il} + \varepsilon_{jkl(i)}$$

i= 1, …, i Secciones.
j= 1, 2, …, f Filas por sección
k=1, 2, …, c Columnas por sección
l=1, …l tratamientos
γ_{ij} =Variable de respuesta observada en una maceta dela j-ésima fila y la k-ésima columna en la i-ésima sección donde se sembró el l-ésimo tratamiento
μ=Media general
λ_i=Efecto aleatorio de la i-ésima sección
λ_i~NIID $(0,\sigma^2_l)$
$\varphi_{j(i)}$=Efecto aleatorio dela j-ésima Fila en la i-ésima sección
$\varphi_{j(i)}$~NIID $(0,\sigma^2_f)$
$\delta_{k(i)}$=Efecto aleatorio dela k-ésima Columna en la i-ésima sección
$\delta_{k(i)}$~NIID $(0,\sigma^2_c)$
τ_l= Efecto fijo de la l-ésimo tratamiento
γ_{il}= Efecto aleatorio de la interacción entre la i-ésima sección y el l-ésimo tratamiento
γ_{il} ~NIID $(0,\sigma^2_c)$
$\varepsilon_{jkl(i)}$=Efecto aleatorio de los residuales
$\varepsilon_{jkl(i)}$ ~NIID $(0,\sigma^2_e)$

Generalizar los diseños completos a más de dos factores es relativamente sencillo desde un punto de vista matemático, pero en su aspecto práctico tiene el inconveniente de que al aumentar el número de factores aumenta muy rápidamente el número de observaciones necesario para estimar el modelo. En la práctica es muy raro utilizar diseños completos con más de dos factores.

Ejemplo 8.1.1.

Un ejemplo aplicado fue el desarrollado bajo condiciones de invernadero en la zona de Puerto La Boca, en Manabí, Ecuador (Erazo, 2018). En este experimento se evaluaron tres híbridos nuevos de pimiento en un invernadero de 1000 m^2 con el propósito de controlar dos fuentes de variación como es la luz y la fertilidad del suelo, se implementó el ensayo en un diseño experimental de fila – columna con 10 repeticiones. Las Variables de respuesta evaluadas fueron: Número de frutos por planta (NFP), Peso de frutos (PF), Largo de frutos (LFr), Ancho de frutos (AFr), Alto de frutos (HFr), Diámetro de tallo (DT), Altura de planta (AP) y severidad de oídium (SEV), Volumen de fruto (V).

El modelo Aditivo lineal correspondiente fue el siguiente:

$Yijk = \mu + \beta i + \rho j + \tau k\ (ij) + \xi ijk$

Donde:

i = 1, 2,… 10 filas
j = 1, 2…..8 columnas
k= 1, 2…..8 tratamientos
Yij= Valor de una variable de respuesta observada en el j-ésimo híbrido evaluado en el i-ésimo bloque
μ = Media general.
βi = Efecto fijo de la i- ésima fila
ρj = Efecto fijo de j- ésima columna
τk (ij) = Efecto fijo del k-ésimo tratamiento
ξij= Efecto aleatorio de los residuales ξij ~ NIID, (0, σe2).

Sobre la base en el modelo definido se realizaron análisis de varianza para probar hipótesis acerca de los efectos fijos, así como comparaciones de medias de los tratamientos mediante la prueba de tukey al Pr<0.05 de probabilidad. El análisis de varianza también sirvió para estimar los componentes de varianza para los efectos aleatorios. Los análisis indicados se realizaron utilizando el Proc GLM del SAS.

El programa en SAS fue el que se describe a continuación:

```
data pimiento;
input trat $1-10 H C NFP PF LFr AFr HFr DT AP SEV;
V = 4/3*3.1416*LFr*AFr*HFr;
CARDS;
Marcato     1    1   20   143.75  15   22.0  22.0  1.0  90   20
Marcato     1    1   18   143.75  13   18.5  18.5  1.2  75   15
Marcato     1    1   20   115.00  15   22.5  22.5  1.0  84   14
Marcato     1    1   18   115.00  13   18.5  18.5  1.5  66   10
Marcato     1    1   20   143.75  12   18.0  18.0  1.0  85   8
Tandara     1    2   25   86.25   20   17.0  17.0  1.4  82   35
Tandara     1    2   25   115.00  19   16.5  16.5  1.3  105  30
Tandara     1    2   20   86.25   17   16.0  16.0  1.0  120  27
Tandara     1    2   25   57.50   16   15.5  15.5  1.5  72   25
Tandara     1    2   20   115.00  14   15.0  15.0  1.2  110  20
E20L.30100       13  30   57.50   16   15.0  15.0  1.3  104  5
E20L.30100       13  25   86.25   16   15.0  15.0  1.4  70   4
E20L.30100       13  30   86.25   17   17.0  17.0  1.0  100  3
E20L.30100       13  25   57.50   18   17.5  17.5  1.0  76   2
E20L.30100       13  30   57.50   16   14.5  14.5  1.2  95   1
Marcato     2    2   18   143.75  15   22.0  22.0  1.2  75   10
Marcato     2    2   20   115.00  13   18.5  18.5  1.0  90   20
Marcato     2    2   18   143.75  15   22.5  22.5  1.5  66   15
Marcato     2    2   20   143.75  12   18.0  18.0  1.0  84   14
Marcato     2    2   18   115.00  13   18.5  18.5  1.0  85   8
Tandara     2    3   20   57.50   14   15.0  15.0  1.3  105  27
Tandara     2    3   20   86.25   17   16.0  16.0  1.4  82   30
Tandara     2    3   25   115.00  20   17.0  17.0  1.2  120  35
Tandara     2    3   20   86.25   16   15.5  15.5  1.0  110  20
Tandara     2    3   20   115.00  19   16.5  16.5  1.5  72   25
E20L.30100       21  25   86.25   16   15.0  15.0  1.4  104  4
E20L.30100       21  20   57.50   17   17.0  17.0  1.0  100  1
E20L.30100       21  35   86.25   16   15.0  15.0  1.3  76   2
```

E20L.30100		21	25	57.50	16	14.5	14.5	1.0	70	3
E20L.30100		21	30	57.50	18	17.5	17.5	1.2	95	5
Marcato	3	3	22	115.00	13	18.5	18.5	1.5	66	8
Marcato	3	3	25	143.75	15	22.5	22.5	1.0	75	10
Marcato	3	3	22	143.75	13	18.5	18.5	1.0	90	20
Marcato	3	3	18	115.00	12	18.0	18.0	1.2	85	15
Marcato	3	3	22	143.75	15	22.0	22.0	1.0	84	14
Tandara	3	1	20	115.00	19	16.5	16.5	1.5	72	20
Tandara	3	1	20	86.25	20	17.0	17.0	1.3	120	25
Tandara	3	1	20	115.00	16	15.5	15.5	1.4	82	30
Tandara	3	1	20	86.25	14	15.0	15.0	1.2	105	35
Tandara	3	1	22	57.50	17	16.0	16.0	1.0	110	27
E20L.30100		32	25	57.50	17	17.0	17.0	1.0	100	2
E20L.30100		32	30	86.25	16	15.0	15.0	1.4	76	4
E20L.30100		32	25	57.50	16	14.5	14.5	1.0	70	3
E20L.30100	3	2	25	86.25	18	17.5	17.5	1.3	95	5
E20L.30100	3	2	30	57.50	16	15.0	15.0	1.2	104	1
Marcato	4	1	20	143.75	15	22.5	22.5	1.0	85	10
Marcato	4	1	18	115.00	13	18.5	18.5	1.2	84	8
Marcato	4	1	20	143.75	15	22.0	22.0	1.0	75	14
Marcato	4	1	22	143.75	13	18.5	18.5	1.5	90	20
Marcato	4	1	22	115.00	12	18.0	18.0	1.0	66	15
Tandara	4	2	25	115.00	17	16.0	16.0	1.2	120	25
Tandara	4	2	25	86.25	16	15.5	15.5	1.0	105	27
Tandara	4	2	23	57.50	14	15.0	15.0	1.3	110	20
Tandara	4	2	23	115.00	19	16.5	16.5	1.4	82	30
Tandara	4	2	22	86.25	20	17.0	17.0	1.5	72	35
E20L.30100	4	3	25	86.25	16	14.5	14.5	1.0	100	1
E20L.30100	4	3	30	57.50	16	15.0	15.0	1.2	70	3
E20L.30100	4	3	35	86.25	16	15.0	15.0	1.4	95	5
E20L.30100	4	3	30	57.50	18	17.5	17.5	1.0	104	4
E20L.30100	4	3	25	57.50	17	17.0	17.0	1.3	76	2
Marcato	5	2	22	143.75	13	18.5	18.5	1.0	84	8
Marcato	5	2	20	143.75	15	22.0	22.0	1.0	66	14
Marcato	5	2	18	115.00	13	18.5	18.5	1.5	85	10
Marcato	5	2	20	115.00	12	18.0	18.0	1.0	75	15
Marcato	5	2	20	143.75	15	22.5	22.5	1.2	90	20
Tandara	5	3	25	57.50	16	15.5	15.5	1.0	105	27
Tandara	5	3	23	86.25	14	15.0	15.0	1.2	110	20
Tandara	5	3	22	115.00	17	16.0	16.0	1.5	72	25
Tandara	5	3	23	86.25	20	17.0	17.0	1.3	120	35
Tandara	5	3	25	115.00	19	16.5	16.5	1.4	82	30
E20L.30100	5	1	25	57.50	16	15.0	15.0	1.2	70	3
E20L.30100	5	1	30	86.25	16	15.0	15.0	1.0	100	5
E20L.30100	5	1	31	57.5	17	17.0	17.0	1.0	104	4
E20L.30100	5	1	34	86.25	16	14.5	14.5	1.3	76	2
E20L.30100	5	1	30	57.50	18	17.5	17.5	1.4	95	1
Marcato	6	3	20	115.00	13	18.5	18.5	1.2	66	14
Marcato	6	3	21	143.75	15	22.5	22.5	1.5	85	8
Marcato	6	3	23	143.75	13	18.5	18.5	1.0	75	15
Marcato	6	3	20	115.00	12	18.0	18.0	1.0	90	20
Marcato	6	3	20	143.75	15	22.0	22.0	1.0	84	10
Tandara	6	1	24	86.25	19	16.5	16.5	1.5	110	25
Tandara	6	1	24	115.00	16	15.5	15.5	1.0	120	20
Tandara	6	1	25	115.00	20	17.0	17.0	1.2	72	35
Tandara	6	1	23	86.25	14	15.0	15.0	1.4	82	30

Tandara	6	1	25	57.50	17	16.0	16.0	1.3	105	27
E20L.30100	6	2	28	57.50	17	17.0	17.0	1.0	100	5
E20L.30100	6	2	29	86.25	16	14.5	14.5	1.0	104	4
E20L.30100	6	2	28	57.50	16	15.0	15.0	1.3	76	3
E20L.30100	6	2	30	57.50	18	17.5	17.5	1.4	70	1
E20L.30100	6	2	34	86.25	16	15.0	15.0	1.2	95	2
Marcato	7	1	22	143.75	13	18.5	18.5	1.5	85	10
Marcato	7	1	23	115.00	12	18.0	18.0	1.0	75	15
Marcato	7	1	20	143.75	15	22.0	22.0	1.0	90	20
Marcato	7	1	18	143.75	15	22.5	22.5	1.2	84	14
Marcato	7	1	18	115.00	13	18.5	18.5	1.0	66	8
Tandara	7	2	25	86.25	17	16.0	16.0	1.2	105	20
Tandara	7	2	25	86.25	20	17.0	17.0	1.5	110	35
Tandara	7	2	20	57.50	16	15.5	15.5	1.4	82	30
Tandara	7	2	23	115.00	14	15.0	15.0	1.3	72	27
Tandara	7	2	22	115.00	19	16.5	16.5	1.0	120	25
E20L.30100	7	3	30	57.50	16	15.0	15.0	1.2	104	4
E20L.30100	7	3	29	86.25	16	15.0	15.0	1.3	70	5
E20L.30100	7	3	25	57.50	18	17.5	17.5	1.4	95	2
E20L.30100	7	3	30	86.25	16	14.5	14.5	1.0	100	3
E20L.30100	7	3	32	57.50	17	17.0	17.0	1.0	76	1
Marcato	8	2	20	115.00	12	18.0	18.0	1.0	75	15
Marcato	8	2	22	143.75	15	22.0	22.0	1.2	90	20
Marcato	8	2	23	115.00	13	18.5	18.5	1.0	84	8
Marcato	8	2	18	143.75	15	22.5	22.5	1.0	66	10
Marcato	8	2	19	143.75	13	18.5	18.5	1.5	85	14
Tandara	8	3	25	115.00	19	16.5	16.5	1.2	120	35
Tandara	8	3	23	57.50	14	15.0	15.0	1.4	82	30
Tandara	8	3	24	86.25	17	16.0	16.0	1.3	110	27
Tandara	8	3	20	86.25	16	15.5	15.5	1.0	105	25
Tandara	8	3	26	115.00	20	17.0	17.0	1.5	72	20
E20L.30100	8	1	30	57.50	17	17.0	17.0	1.3	70	2
E20L.30100	8	1	29	86.25	16	14.5	14.5	1.0	104	4
E20L.30100	8	1	28	57.50	16	15.0	15.0	1.0	100	5
E20L.30100	8	1	30	57.50	18	17.5	17.5	1.4	76	1
E20L.30100	8	1	34	57.50	16	15.0	15.0	1.2	95	3
Marcato	9	3	22	143.75	15	22.0	22.0	1.0	90	20
Marcato	9	3	18	115.00	12	18.0	18.0	1.5	75	15
Marcato	9	3	18	143.75	13	18.5	18.5	1.2	85	10
Marcato	9	3	20	143.75	15	22.5	22.5	1.0	66	14
Marcato	9	3	20	115.00	13	18.5	18.5	1.0	84	8
Tandara	9	1	25	86.25	14	15.0	15.0	1.4	82	30
Tandara	9	1	25	86.25	20	17.0	17.0	1.3	130	35
Tandara	9	1	20	115.00	16	15.5	15.5	1.0	120	25
Tandara	9	1	20	57.50	19	16.5	16.5	1.5	110	27
Tandara	9	1	23	115.00	17	16.0	16.0	1.2	105	20
E20L.30100	9	2	34	57.50	16	14.5	14.5	1.4	76	1
E20L.30100	9	2	30	57.50	16	15.0	15.0	1.3	95	3
E20L.30100	9	2	28	86.25	18	17.5	17.5	1.2	104	4
E20L.30100	9	2	28	57.50	16	15.0	15.0	1.0	70	5
E20L.30100	9	2	30	86.25	17	17.0	17.0	1.0	100	2
Marcato	10	1	20	115.00	13	18.5	18.5	1.0	84	14
Marcato	10	1	20	143.75	15	22.5	22.5	1.2	90	20
Marcato	10	1	18	143.75	13	18.5	18.5	1.0	75	15
Marcato	10	1	22	115.00	13	22.0	22.0	1.5	66	8
Marcato	10	1	22	143.75	15	18.0	18.0	1.0	85	10

```
Tandara         10  2  25  86.25   20  17.0  17.0  1.3  120  27
Tandara         10  2  24  57.50   16  15.5  15.5  1.4  82   30
Tandara         10  2  20  115.00  19  16.5  16.5  1.0  105  35
Tandara         10  2  20  115.00  17  16.0  16.0  1.2  72   20
Tandara         10  2  24  86.25   14  15.0  15.0  1.5  110  25
E20L.30100      10  3  35  57.50   17  17.0  17.0  1.0  70   3
E20L.30100      10  3  30  86.25   18  17.5  17.5  1.0  100  2
E20L.30100      10  3  33  57.50   16  15.0  15.0  1.3  95   1
E20L.30100      10  3  30  57.50   16  15.0  15.0  1.2  104  4
E20L.30100      10  3  34  86.25   16  14.5  14.5  1.4  76   5
;
Proc univariate plot normal;
Var NFP PF V DT AP SEV;
proc discrim method=normal short pool=test;
classes trat;
proc print;
proc glm;
classes H C trat;
model NFP PF V DT AP SEV=H C trat;
means trat/Duncan Tukey Alpha=0.05;
run;
proc corr;
var NFP PF V DT AP SEV;
run;
```

PARTE IV
DISEÑOS EXPERIMENTALES ESPECIALES Y DE TRATAMIENTOS

9

DISEÑO LÁTICE O DE BLOQUES INCOMPLETOS

Julio Gabriel Ortega

Alfredo Valverde Lucio

Características

Cuando se tiene un experimento donde los tratamientos se alojan al azar sobre las unidades experimentales, de modo que solo un conjunto de los mismos aparece representado por lo menos en uno de los bloques, se dice que el experimento es de bloques incompletos. Se distinguen dos grandes grupos de arreglos experimentales de este tipo, a saber: i) la clase experimentos genéricamente conocidos con el nombre de látices, seudofactoriales o cuasifactoriales, y ii) los diseños de bloques incompletos propiamente dichos.

Fundamentalmente los látices se caracterizan porque es posIble establecer una relación uno a uno entre los tratamientos en ensayo y las combianciones de tratamientos de un experiemento factorial. Tal correspondencia se explota de manera natural para producir el método de análisis de esta clase de experimentos. Para los diseños de bloques incompletos restantes; aun cuando es posible establecer una relación similar, esta es de poca o de ninguna utilidad para producir el método del análisis estadístico.

El diseño de bloques incompletos asegura, por un lado, un mejor control de la heterogeneidad del material experimental, pero por otro lado, pierde precisión al confundir parte de la información relevante.

Análiticamente, es posible demostrar que cualquier contraste entre los efectos reales de tratamientos, puede estimarse de dos maneras, en términos de los efectos factoriales estimados. Una de ellas es a partir de comparaciones dentro de bloques, y la otra, a partir de comparaciones entre bloques, suponiendo aleatorios los efestos de estos últimos. Tales estimaciones son independientes si los errores de las parcelas son además, de las suposiones usuales, normales y no correlacionados, con media 0 y varianza σ_b^2 en general desconcocida. Por consiguiente, presumiblemente se lograrian mejores estimaciones combinando la información obtenida a partir de comparaciones dentro de bloques (denominada información intra bloque) con la información obtenida a partir de comparaciones entre bloque denominada (informacion interbloque). Los látice hacen uso de esta técnica de estimación, conociendose el proceso como recuperacion de información interbloque. Sin embargo, el proceso tiene una debilidad: ni la varianza de los errores de la parcela (σ^2) , ni la varianza de los efectosa de bloques (σ_b^2), se conocen, u solo pueden derivarse estimaciones insesgadas para estos parámetros. Al insertar las estimaciones de las varianzas, en lugar de los valores verdaderos, cuando se combinan las informaciones intra e interbloques, es obvio que no se tienes los mejores estimadores lineales insesgados de los contrastes importantes para el experimentador; de hecho; no se sabe con exactitud que propiedades tiene los estimadores combinados, excepto que son insesgados y posiblemente más eficientes que las estimaciones intrabloques.

Ejemplo 9.1.

Se realizó un experimento de campo en la localidad de Chullchunq'ani (provincia Carrasco del departamento de Cochabamba) a 108 Km. de la carretera antigua Cochabamba - Santa Cruz. Geográficamente situada a 17 º 30´ de Latitud Sud y 65º 15´de Longitud oeste, a una altitud de 3,200 msnm. Cuenta con una temperatura media anual de 15.5 ºC y una precipitación pluvial media anual de 629 mm. Es una zona que presenta condiciones favorables para el desarrollo del oomycete *Phytophthora infestans*, que causa la enfermedad del tizòn en papa y otras solanáceas.

Para evaluar la resistencia al tizón en campo, se sembró 31 cultivares nativos y cinco cultivares testigos en un diseño de Látice simple 6 x 6, con tres repeticiones. Donde las unidades experimentales estaban formadas por un surco de 5 plantas, distanciados a 0.70 m entre surcos y a 0.30 m entre plantas.

Las variables evaluadas fueron: Área Bajo la Curva de Progreso de *Phytophthora infestans* (AUDPC)

Modelo aditivo lineal utilizado

$$Y_{ijk} = \mu + \beta_i + \delta_j (\beta_i) + \tau_k + \varepsilon_{ijk}$$

i = 1, 2, 3…..6, b = bloques incompletos

j = 1, ..., 3, r = Repeticiones

k = 1,2,3,.....36, c =cultivares

Donde:

Y_{ijk} = Valor observado de una variable de respuesta, en el i-ésimo bloque incompleto, del j-ésimo bloque, que recibe el k-ésimo cultivar.

μ = Media general del ensayo

β_i =Efecto aleatorio del i-ésimo bloque incompleto, $\beta_i \sim NIID(0,\sigma^2_e)$

$\delta_j (\beta_i)$ =Efecto aleatorio de la interacción entre la j-ésimo repetición r, dentro del i-ésimo bloque incompleto b. $\delta_j (\beta_i) \sim NIID(0,\sigma^2_e)$

τ_k = Efecto fijo del k-ésimo cultivar.

ε_{ijk} =Efecto aleatorio de los residuales $\varepsilon_{ij} \sim NIID(0,\sigma^2_e)$

Los rendimientos se analizaron empleando el diseño de bloques completos al azar, cuyo modelo lineal aditivo corresponde a:

$$Y_{ijk} = \mu + \beta_i + \tau_j + \varepsilon_{ijk}$$

i = 1, …..3, b = Bloques

j = 1, ..., 14, c = Cultivares

Donde:

Y_{ijk} = Valor observado de una variable de respuesta, en el i-esimo bloque, que recibe el j-ésimo cultivar.

μ = Media general del ensayo

β_i =Efecto aleatorio del i-ésimo bloque b; $\beta_i \sim NIID(0,\sigma^2_e)$

τ_j τ= Efecto fijo del j-ésimo cultivar.

ε_{ijk} =Efecto aleatorio de los residuales; $\varepsilon_{ij} \sim NIID(0,\sigma^2_e)$

```
options ls=85;
data campo;
input rep bloq trat lec1 lec2 lec3 lec4 lec5 lec6 lec7 lec8 lec9;
/*
lec1= (arsin((lec1/100)**0.5))*180/3.1416;
lec2= (arsin((lec2/100)**0.5))*180/3.1416;
lec3= (arsin((lec3/100)**0.5))*180/3.1416;
```

```
lec4= (arsin((lec4/100)**0.5))*180/3.1416;
lec5= (arsin((lec5/100)**0.5))*180/3.1416;
lec6= (arsin((lec6/100)**0.5))*180/3.1416;
lec7= (arsin((lec7/100)**0.5))*180/3.1416;
lec8= (arsin((lec8/100)**0.5))*180/3.1416;
lec9= (arsin((lec9/100)**0.5))*180/3.1416;
*/
AUDPC=(lec1+lec2)/2*7+(lec2+lec3)/2*8+(lec3+lec4)/2*6+(lec4+lec5)/2*7+(lec5+lec6)/2*7+(lec6+lec7)/2*7+(lec7+lec8)/2*8
+ (lec8+lec9)/2*13;
Cards;
 1  1  1  3    8   8  10  12  18  33  89  100
 1  1  2  3    8   8   8   8  11  25  70   96
 1  1  3  3    9   9  10  10  15  15  20   98
 1  1  4  2   15  15  18  19  24  27  36   97
 1  1  5  7    7   7   7   7  11  15  26   34
 1  1  6 15   15  15  18  18  25  27  73   99
 1  2  7  1    5   5   7   7  11  13  27   33
 1  2  8  3    3   4   4   4   4   5   9   17
 1  2  9 30   20  20  22  25  25  25  33  100
 1  2 10 17   18  18  18  18  20  25  38   99
 1  2 11  0    0   0   2   3   8   8  11   16
 1  2 12  0.6  6   6   7   8  10  18  39   68
 1  3 13  0    0   0   2   6   6   8   8   14
 1  3 14  4   11  11  11  12  15  16  26   98
 1  3 15  1.6  6   6   6   7   9  13  22   37
 1  3 16  6.6 20  20  25  27  34  36  46  100
 1  3 17 18   18  18  22  26  26  28  36   90
 1  3 18 25   25  25  25  26  30  34  45   99
 1  4 19  0    0   0   0   3   6   9  11   17
 1  4 20 30   30  30  33  35  42  52  66  100
 1  4 21  5    7   7   8   9   9  19  25   25
 1  4 22 15   15  15  16  16  23  27  37  100
 1  4 23 37   37  37  37  37  42  46  98  100
 1  4 24 24   24  24  25  27  29  40  65  100
 1  5 25  1    2   2   5   5  10  18  44  100
 1  5 26 45   45  48  85  94  95  97 100  100
 1  5 27 13   16  16  18  18  18  22  30   34
 1  5 28  2    2   2   2  14  16  24  25   31
 1  5 29 36   36  36  42  42  45  48  96  100
 1  5 30  2    6   6   6   7   9  10  26   28
 1  6 31  2    8   8  10  14  16  25  48  100
 1  6 32 34   34  34  37  44  52  55  63  100
 1  6 33  8    9   9  13  13  15  17  37   38
 1  6 34 10   16  16  23  24  28  35  65  100
 1  6 35  2.4 10   8  13  15  17  18  25   36
 1  6 36 15   15  15  18  23  27  35  43   72
 2  1 32 16   30  30  45  65  68  71  99  100
 2  1  4  4    8  15  28  30  33  35  45   60
 2  1 17 15   15  15  25  28  32  35  46   79
 2  1 27 14   14  17  22  22  23  32  35   47
 2  1 24 10   22  28  38  54  57  65  96  100
 2  1  6  9   13  13  13  14  14  18  44   93
 2  2 15  3   13  13  16   7  11  17  37   39
 2  2 20 33   40  43  46  46  48  49  72  100
 2  2 31  2   12  12  20  20  23  31  45  100
 2  2 26 45   45  45  46  46  48  48  48   98
 2  2 19  0    0   0   0  12  13  14  15   17
 2  2 10 60   60  60  63  63  67  94 100  100
 2  3 25  0    0   0  18  19  20  20  44   98
 2  3 18 23   23  25  30  32  34  36  56  100
 2  3  7  3    7   8  14  19  21  23  28   34
 2  3  8  5    6   6   7   7   9  13  23   27
 2  3 33  7.4  8  12  12  10  12  18  33   36
 2  3 30  4    7   7   9   9  11  12  25   29
 2  4  5  1    7   6   6   6   8  12  19   30
 2  4 11  0    0   0   0   0   3   5   9    9   17
 2  4  1 18   26  26  27  33  37  45  86  100
 2  4  9 35   35  35  37  37  37  37  68  100
 2  4 22 15   15  17  25  25  26  32  76  100
```

```
 2  4 12  3 15 19 23 26 36 39 55 99
 2  5 36  9  9  9  9 10 13 17 38 91
 2  5 13  0  0  0  0  3  5  5  9 11
 2  5 14 13 13 13 15 15 17 26 46 89
 2  5  2 12 17 17 20 18 20 36 94 93
 2  5 35  1  6  8 14 16 18 25 33 39
 2  5  3 20 20 20 25 25 29 29 40 85
 2  6 29 55 55 66 73 73 84 87 97 100
 2  6 34 11 22 22 25 27 29 35 76 99
 2  6 28  2  3  5 11 22 26 30 31 34
 2  6 21 12 15 15 18 18 19 19 28 38
 2  6 23 12 35 35 38 33 36 44 96 100
 2  6 16 19 35 35 40 43 48 52 93 100
 3  1 35  4 16 10 13 13 15 16 28 36
 3  1  5 12 12 12 13 13 13 17 29 37
 3  1  1 20 20 24 28 28 33 40 94 99
 3  1 13  0  0  0  0  3  5  5  9  9
 3  1 10  1  4  4 18 28 34 95 100 100
 3  1  3 10 23 23 32 32 35 35 43 96
 3  2 33  9 10 12 12 12 12 19 36 38
 3  2 15 10 12 12 15 15 15 25 37 38
 3  2 18 24 35 35 35 38 41 43 89 100
 3  2  9 27 40 42 42 45 47 67 94 100
 3  2 12  4 15 14 14 16 32 38 47 99
 3  2 11  0  0  0  0  3  4  5 10 17
 3  3 14 12 13 13 13 15 16 24 41 100
 3  3 31  4  9  9 14 16 18 26 74 100
 3  3 28  3  6  6 13 25 27 32 33 34
 3  3  8  7  7  7  7  7  7 14 26 28
 3  3 20 44 50 50 52 50 53 56 100 100
 3  3 30  6  9 10 12 12 12 14 26 31
 3  4 29 35 35 37 53 53 57 66 95 100
 3  4 21 25 25 25 26 26 28 28 33 39
 3  4  7 10 12 12 15 15 21 21 28 36
 3  4 17 15 15 15 17 18 25 32 36 43
 3  4 34  6 23 23 20 26 31 38 68 99
 3  4 26 80 80 80 100 100 100 100 100 100
 3  5  2 25 25 25 29 30 30 32 53 98
 3  5 36 15 15 17 17 20 22 26 44 94
 3  5 19  1  1  2  2  6  6 10 10 23
 3  5 22 17 22 22 24 24 24 27 44 100
 3  5  6 13 13 13 15 24 26 29 46 100
 3  5 16 25 37 39 39 67 70 75 92 99
 3  6 25  2  2  6  8  8 13 15 54 100
 3  6 27 27 27 27 27 24 26 28 33 42
 3  6 32 29 38 38 28 39 43 45 47 74
 3  6 24 25 25 25 28 32 36 40 65 100
 3  6 23 23 58 58 66 68 68 74 91 100
 3  6  4  9 25 27 29 29 29 33 42 87
;
proc print;
proc glm;
class rep bloq trat;
model AUDPC= rep bloq(rep) trat;
means trat/duncan tukey;
run;
```

UNIDAD 10

DISEÑO EN PARCELAS DIVIDIDAS (DPD)

Julio Gabriel Ortega, Carlos Castro Piguave, Alfredo Valverde Lucio

Características

Considere un diseño de bloques completos al azar, con *p* tratamientos *A* y *r* repeticiones para cada tratamiento. Si cada una de las parcelas básicas se divide en *q* subparcelas, y sobre éstas se colocan al azar *q* tratamientos *B*, se genera un diseño experimental especial conocido como diseño en parcelas divididas (DPD).

Esta clase de arreglo es útil cuando ciertos tratamientos requieren de parcelas grandes para su ensayo. Como ocurre con frecuencia con los métodos de riego, las distancias entre surco, etc.; al combinarse con fertilizantes, cultivares, etc. Estos factores se colocan sobre unidades experimentales menores. Es frecuente referirse a los tratamientos sobre las parcelas grandes como los tratamientos, y a los tratamientos sobre las parcelas chicas como los sub-tratamientos.

La figura siguiente ilustra mejor la situación (Figura 4). Supóngase que se ensayan tres distancias entre surcos, digamos 0.99 m, 1.20 m y 1.50 m, sobre cuatro cultivares de caña: 1,2,3,4. Ya que es impráctico para la máquina delinear parcelas pequeñas con distintas distancias entre surcos, este factor podría colocarse sobre parcelas grandes, las cuales se subdividirán en cuatro sub-parcelas cada una, para alojar las variedades sobre éstas últimas. Después de la aleatorización, un bloque completo se vería como esquema representado en la figura 4.

Surcos a 1.5 m	Surcos a 0.90 m	Surcos a 1.20 m
2	3	1
4	2	3
1	1	4
3	4	2

Figura 4. Diseño experimental en parcelas divididas

Si se observa cualquier parcela grande de un diseño de bloques completos al azar en parcelas divididas, puede concluirse de inmediato que los efectos del factor en las parcelas grandes (tratamientos), están completamente confundidos con los efectos del factor en las sub-parcelas (sub-tratamientos). Así, los efectos de los tratamientos son directamente estimables a partir de contrastes entre sub-parcelas dentro de parcelas grandes; de hecho, estos contrastes producen estimaciones de los efectos principales del sub-tratamiento. Los efectos del factor en las parcelas grandes se estiman contrastando los totales de parcelas grandes. Por esta característica de los diseños en parcelas divididas, es común decir que estos arreglos conducen a la confusión de algunos efectos principales.

Es muy frecuente que en la práctica se cometan graves errores en el diseño de esta clase de experimentos, el más común se refiere a la escasa precisión con que a veces se estiman los efectos del factor en las parcelas grandes.

Análisis de los diseños en parcelas divididas

Modelo lineal

Dado que en este diseño la aleatorización se realiza en dos etapas, el modelo aditivo lineal tendrá dos fuentes de error, una desde las unidades completas y otra desde las subunidades.

En el caso de que los niveles del factor que va en las unidades completas se distribuyan según un DCA, el modelo aditivo lineal estará dado por:

$$Yijk = \mu + \alpha_{ij} + \beta_{ijk} + (a\beta)ik + eijk$$

donde:

Yijk = es el valor o rendimiento observado con el i-ésimo nivel del factor A, j-ésima repetición, y k-ésimo nivel del factor B.
i = 1, ..., p (p = número de niveles del factor A)
j es el efecto del j-ésimo bloque
j = 1, ..., r (r = número de bloques)
k = 1, ..., q (q = número de niveles del factor B)
μ = es el efecto de la media general.
αij = es el efecto del i-ésimo nivel del factor A.Es el efecto del error experimental en parcelas [Error (a)]
ßijk= es el efecto del k-ésimo nivel del factor B.
(aß)ik = es el efecto de la interacción en el i-ésimo nivel del factor A y el k-ésimo nivel del factor B.
eijk= es el efecto del error experimental en subparcelas [Error (b)]

Ejemplo 10.1.

En una investigación realizada en la Finca Andil de la Universidad Estatal del Sur de Manabí, Jipijapa, se probaron tres cultivares de café (*Coffe arabiga*) en tres diferentes densidades de siembra. Este experimento fue implementado en diseño experimental de parcelas divididas. Los datos se muestran en la Tabla 10.1.1. Se pide hacer el análisis de varianza y la comparación de medias de los tratamientos.

Tabla 10.1.1. Datos de producción de café en cereza (kg/UE) y rendimiento de café oro (kg/ha).

Tratamiento	Densidad (A)	Cultivar (S)	Bloque	Producción de grano en cereza (kg/UE)	Rendimiento de grano en café oro (Kg/ha)
sarchimor 1669 2,50x1,00	1	1	1	806,20	649,30
sarchimor4260 2,00x1,00	1	2	1	1252,81	1078,44
acawa 2,50x1,00	1	3	1	788,50	600,75
sarchimor 1669 2,50x1,00	2	1	1	1141,40	1025,00
sarchimor4260 2,00x1,00	2	2	1	965,56	894,25
acawa 2,50x1,00	2	3	1	805,45	719,30
sarchimor 1669 2,50x1,00	3	1	1	1346,75	1225,55
sarchimor4260 2,00x1,00	3	2	1	1669,00	1499,94
acawa 2,50x1,00	3	3	1	563,45	502,35
sarchimor 1669 2,50x1,25	1	1	2	519,25	450,85
sarchimor4260 2,00x1,25	1	2	2	1337,06	1189,43
acawa 2,50x1,25	1	3	2	260,45	266,30
sarchimor 1669 2,50x1,25	2	1	2	823,35	668,80
sarchimor4260 2,00x1,25	2	2	2	1010,56	864,75
acawa 2,50x1,25	2	3	2	700,65	556,15
sarchimor 1669 2,50x1,25	3	1	2	1080,95	940,20
sarchimor4260 2,00x1,25	3	2	2	2052,41	1884,63
acawa 2,50x1,25	3	3	2	412,40	356,80
sarchimor 1669 2,50x1,50	1	1	3	688,60	569,20
sarchimor4260 2,00x1,50	1	2	3	739,31	610,88
acawa 2,50x1,50	1	3	3	869,00	674,45
sarchimor 1669 2,50x1,50	2	1	3	493,05	420,70
sarchimor4260 2,00x1,50	2	2	3	1234,56	1040,94
acawa 2,50x1,50	2	3	3	688,90	570,50
sarchimor 1669 2,50x1,50	3	1	3	1738,20	1587,75
sarchimor4260 2,00x1,50	3	2	3	1567,44	1421,69
acawa 2,50x1,50	3	3	3	810,00	719,50

Con propósitos de didactica, se hará el análisis para la producción de grano en cereza.

Lo primero que se debe hacer es generar las tablas apropiadas para el análisis. La Tabla 10.1.2., fue organizanizada timado como referencia la Tabla 10.1.1.

Tabla 10.1.1. Datos ordenados de producción de café en cereza (kg/UE),

		Rep			
den	trat	**1**	**2**	**3**	**Total**
	1	806.20	519.25	688.60	2014.05
1	2	1252.81	1337.06	739.31	3329.19
	3	788.50	260.45	869.00	1917.95
Subtotal		**2847.51**	**2116.76**	**2296.91**	**7261.19**
	1	1141.40	823.35	493.05	2457.80
2	2	965.56	1010.56	1234.56	3210.69
	3	805.45	700.65	688.90	2195.00
Subtotal		**2912.41**	**2534.56**	**2416.51**	**7863.49**
	1	1346.75	1080.95	1738.20	4165.90
3	2	1669.00	2052.41	1567.44	5288.85
	3	563.45	412.40	810.00	1785.85
Subtotal		**3579.20**	**3545.76**	**4115.64**	**11240.60**
Total		**9339.13**	**8197.09**	**8829.06**	**26365.28**

Luego se elaboró la Tabla 10.1.3., en base a la Tabla 10.1.2.

Tabla 10.1.3. Tabla de doble entrada de densidad y cultivar.

	den (A)			**Y.k**	**Prom Y.k**
var (B)	1	2	3		
1	2014.05	2457.80	4165.90	8637.75	**2879.25**
2	3329.19	3210.69	5288.85	11828.73	**3942.91**
3	1917.95	2195.00	1785.85	5898.80	**1966.27**
Yi.	**7261.19**	**7863.49**	**11240.60**	**26365.28**	
Prom Yi.	**2420.40**	**2621.16**	**3746.87**		

Método de análisis

1. Se calcula el factor de corrección C

$C=G^2/rts= 26365.28^2/3x3x3=$ **25745471.33**

2. Cálculo de la suma de cuadrados de bloques, SCB

$SCB = \Sigma Bi^2/pq – C = [(9339.13)^2+ \ldots + (8829.06)^2/3x3] – C =$ **72733.55**

3. Cálculo de la suma de cuadrados debido a los tratamientos de parcela grande (densidades), SC(A)

$SC(A) = \Sigma Yi.^2/rq – C = [(7261.19)^2 + (7863.49)^2 + (11240.60)^2/3x3] – C =$ **1022347.33**

4. Cálculo de la suma de cuadrados del error de A, SCError (A)

SCError (A) = SCsubtotal – SCA - SCBloques = 1231888.88 – 1022347.33 – 72733.55 = **136808.00**

5. Cálculo de la suma de cuadrados subtotal, SCsub-total

$SCsub\text{-}total = \Sigma Yij.^2/q – C = [(2847.51^2 + \ldots + 4115.64^2)/3] – C =$ **1231888.88**

6. Cálculo de la suma de cuadrados debido a sub-tratamientos (cultivares), **SCS**

SCS = $\Sigma S_k^2/rxp - C$ $\{[(8637.75)^2 + \ldots + (5898.80)^2]/3x3\}-C$ = **1957339.97**

7. Cálculo de la suma de cuadrados de la interacción, **SC(AS)**

SC(AS) = = $(\Sigma Yijk^2/4 - C) - SCA - SCS$ = $\{[(2014,05)^2 + \ldots + (1785,85)^2/3] - C\}$ - 1022347,33 - 1957339,97 = 3755080.82 – 1022347.33 – 1957339.97 = **775393.52**

8. Suma de cuadrados del error de sub-tratamiento, SC (Error S)

SC (Error S) = SCT – SC (A) – SC (S) –SC (AS) = 4914592.00 – 136808.00 – 1957339.97 – 775393.52 = **1159511.18**

9. Cálculo de la suma de cuadrados total, SCTotal

SCTotal = $[(806.20)^2 + \ldots + (810.00)^2]-C$ = **4914592.00**

10. Elaboración de la Tabla de Análisis de Varianza

Tabla 10.1.4. Fórmulas para el análisis de varianza

Fuente de variación	Grados de libertad	Suma de Cuadrados	Cuadrados medios	Espereranza de los cuadrados medios
Bloques (B)	r-1	$\sum_{j=1}^{r} \frac{B_i^2}{pq} - \frac{G^2}{rpq}$	CMB	
Tratamientos (T)	p-1	$\sum_{j=1}^{t} \frac{T_j^2}{r-q} - \frac{G^2}{rpq}$	CMT	$\sigma_s^2 + q\sigma_p^2 + rq\frac{\sum_j (\tau_j - \bar{\tau})^2}{p-1}$
Error en parcelas grandes (Ep)	(p-1) (r-1)	SCEp	CMEp	$\sigma_s^2 + q\sigma_p^2$
Subtotal	Rp-1	$\sum_{ij} \frac{Y_{ij.}^2}{q} - \frac{G^2}{rpq}$	CMS	
Subtratamientos (S)	q-1	$\sum_{k=1}^{q} \frac{S_k^2}{rp} - \frac{G^2}{rpq}$	CMS	$\sigma_s^2 + rq\frac{\sum_k (\delta_{kj} - \bar{\bar{\delta}})^2}{q-1}$
Tx S	(p-1) (q-1)	SCTS	CMT X S	$\sigma_s^2 + q\sigma_p^2 + rq\frac{\sum_j (\tau_j - \bar{\tau})^2}{p-1}$
Error en parcelas chicas (Es)	p(r-1)(q-1)	SCEs	CMEs	σ_s^2
Total	Rpq – 1	$\sum_{ijk} Y_{ijk}^2 - \frac{G^2}{rpq}$		

Tabla 10.1.5. Análisis de varianza para densidad y cultivares de café

FV	gl	SC	CM	F
Bloques	2 (b-1)	72733.55		
A	2 (t-1)	1022347.33	511173.67	14.95*
Error (a)	4 (b-1) (t-1)	136808.00	**34202.00**	
Sub-total	8 (bt-1)	1231888.88		
S	2 (q-1)	1957339.97	97866998	10.13**
AS	4 (t-1) (q-1)	775393.52	193848.38	2.01ns
Error (b)	12 t(b-1) (q-1)	115951118	**96625.93**	
Total	26 btq - 1	4914592.00		

Ft0,05, 2,4 = 6.94, Ft0,01, 2,4 = 18.00, Ft0,05, 2,12 = 3.89, Ft0,01, 2,12 = 6.93, Ft0,05, 4,12 = 3.26

Ft0,01, 4,12 = 5.41

Conclusión: Hubo diferencias significativas al p<0.05 de probabilidad en densidades y altamente significativas al p<0.01 entre cultivares. No hubo diferencias significativas en la interacción, lo que indica que las variables evaluadas son independientes.

11. Comparación de medias mediante la prueba de Tukey

a. Calculamos las medias de densidad

		Densidad		
		1	**2**	**3**
		2420.40	**2621.16**	**3746.87**
1	**2420.40**	-	-200.77ns	-1326.47**
2	**2621.16**		-	-1125.70**
3	**3746.87**			-

b. Determinamos la Diferencia Significativa Honesta (DSH)

$$DSH = q_{\alpha,t,n}\sqrt{\frac{S^2}{r}}$$

DSH = q $_{0.01,\ 2,4}$ √34202.00/3 = 6.51√34202.00/3 = **695.1**

Interpretación: La diferencias entre los tratamientos 1-3 y 2-3, son más altos que el valor calculado de DSH, lo que indica que hay diferencias altamente significativas al p<0.01 de probabilidad entre densidades.

c. Se hace el mismo proceso para cultivares

		Cultivares		
		1	**2**	**3**
		2879.25	**3942.91**	**3942.91**
1	**2879.25**	-	-1063.66**	-1063.66**
2	**3942.91**		-	0.00
3	**3942.91**			-

$$DSH = q_{\alpha,t,n}\sqrt{\frac{S^2}{r}}$$

DSH = q $_{0.01,\ 2,12}$ √96625.93/3 = **4.32**√96625.93/3 = **775.30**

Interpretación: La diferencias entre los tratamientos 1-2 y 1-3, son más altos que el valor calculado de DSH, lo que indica que hay diferencias altamente significativas al p<0.01 de probabilidad entre los cultivares.

d. Para la interacción se debe hacer una gráfica de doble entrada con los valores de densidad y cultivar

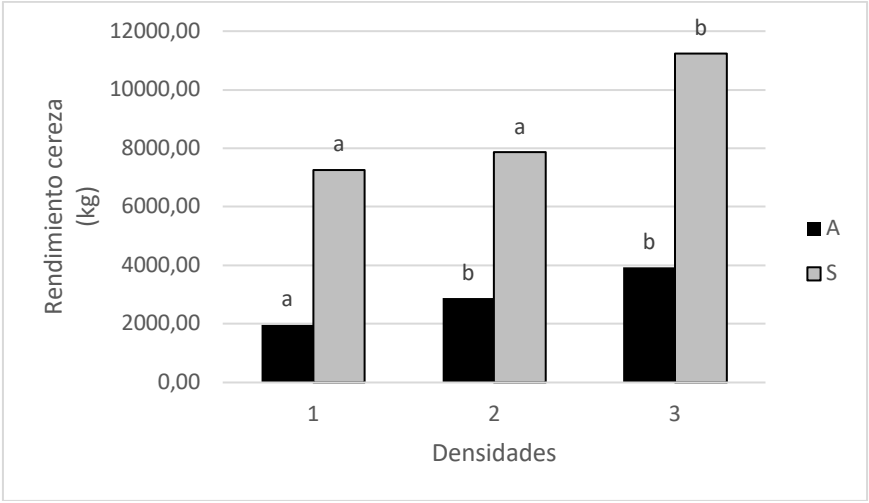


Figura 1. Análisis de la interacción AS

Es notorio según se observa en la figura que la densidad, es independiente al comportamiento de los cultivares. Hubo mejor respuesta de la densidad 3 y el cultivar 3.

Análisis en SAS.

Se usan los procedimientos anova, glm y mixed. El glm tiene la ventaja de presentar los resultados en un formato más familiar. La tabla resumen del análisis de varianza incluye todos sus componentes (fuentes de variación, grados de libertad, sumas de cuadrados, cuadrados medios, valores f y valores p), indicando cuál es el término del error usado en cada caso. Para la evaluación de los efectos principales se pueden usar diversas pruebas, entre ellas la de duncan – no disponible en el proc mixed. En este caso usaremos la prueba de Tukey.

Programa SAS

```
Data PD;
Input rep trat subt y;
Cards;
……
……
……
……
Proc anova;
Classes rep trat rep*trat subt trat*subt;
Test h=rep trat  e=rep*trat;
Means trat/ tukey  e=rep*trat  alpha= 0.01;
Means subt/ trat*subt/tukey;
Run;
```

Ejemplo 10.2.

El conjunto de datos a utilizar como ejemplo demostrativo (Tabla 10.2.1) proviene de Snedecor y Cochran (1971) donde se evaluó el efecto de cuatro fechas del último corte del año (F1, F2, F3 y F4) sobre la productividad de materia seca de tres variedades de alfalfa (Cossack, Ladak y Ranger). Cada variedad (parcela principal) se ubicó al azar dentro de cada bloque y posteriormente se subdividió aleatoriamente cada variedad en cuatro fechas de corte (sub-parcela), empleándose seis bloques en total.

Análisis estándar con PROC GLM

Cuando se analizan datos provenientes de un diseño de parcelas divididas, el diseño factorial y la estructura de aleatorización es acomodada utilizando errores separados para la construcción de las pruebas F. El error correcto a ser utilizado en la prueba F para el tratamiento aplicado a las parcelas principales es la interacción bloque*tratamiento de la parcela principal, mientras que el error residual es el apropiado para probar la hipótesis relacionada con el tratamiento aplicado a la sub-parcela y la interacción entre los tratamientos aplicados en la parcela principal y la subparcela.

Tabla 10.2.1. Datos del ejemplo utilizado en este trabajo. Los datos corresponden al rendimiento de tres variedades de alfalfa (t/acre) en cuatro fechas del último corte

Variedad	Fecha	Bloques					
		1	2	3	4	5	6
Ladak	1	2.17	1.88	1.62	2.34	1.58	1.66
	2	1.58	1.26	1.22	1.59	1.25	0.94
	3	2.29	1.60	1.67	1.91	1.39	1.12
	4	2.23	2.01	1.82	2.10	1.66	1.10
Cossack	1	2.33	2.01	1.70	1.78	1.42	1.35
	2	1.38	1.30	1.85	1.09	1.13	1.06
	3	1.86	1.70	1.81	1.54	1.67	0.88
	4	2.27	1.81	2.01	1.40	1.31	1.06
Ranger	1	1.75	1.95	2.13	1.78	1.31	1.30
	2	1.52	1.47	1.80	1.37	1.01	1.31
	3	1.55	1.61	1.82	1.56	1.23	1.13
	4	1.56	1.72	1.99	1.55	1.51	1.33

Fuente: Snedecor y Cochran (1971)

PROC GLM fue el principal procedimiento de SAS para analizar modelos mixtos antes del surgimiento del PROC MIXED, aun cuando los cálculos básicos del PROC GLM son para modelos de efectos fijos. Un análisis más detallado del PROC GLM puede encontrarse en Little *et al*. (1991). El siguiente programa lee los datos y ejecuta el análisis normal utilizando el PROC GLM:

```
data alfalfa;
input variedad $ fecha block rend;
cards;
- ver datos Cuadro 1 -
;
run;
```

```
title3 "Parcelas divididas con PROC GLM";
proc glm data=alfalfa;
class variedad fecha block;
model rend = block variedad block*variedad fecha fecha*variedad /ss3;
test h = variedad e=block*variedad;
run;
quit;
```

El término BLOCK*VARIEDAD se incluye en la instrucción de MODEL para establecer su utilización en la opción E en la instrucción TEST. Esto permite la construcción correcta de las pruebas para el factor parcelas principales (VARIEDAD).

Análisis con PROC MIXED para diseños de parcelas divididas

PROC MIXED está basado en estimaciones probabilísticas máximas o restringidas de modelos lineales estadísticos que contemplan términos fijos y aleatorios, tales como el modelo lineal en el que se basa el análisis de parcelas divididas, por lo que se presenta como la mejor opción para el análisis estadístico de este tipo de diseño experimental (Littell *et al*. 1996).

El programa básico para analizar diseños de parcelas divididas utilizando PROC MIXED es:

```
data alfalfa;
input variedad $ fecha block rend;
cards;
- Datos según Tabla 1 -
;
run;
proc mixed data=alfalfa method=reml;
class block variedad fecha;
model rend = block variedad fecha fecha*variedad /ddfm=satterth;
random block block*variedad;
run;
```

La opción METHOD=REML invoca la estimación por máxima verosimilitud restringida, el cual es el método predeterminado del PROC MIXED. La instrucción MODEL contiene los mismos términos que en PROC GLM, excepto por el factor BLOCK*VARIEDAD. Este término fue incluido en el PROC GLM para producir el cuadrado medio asociado con este componente aleatorio ya que constituye el denominador para las pruebas F que involucran a las parcelas principales. En el PROC MIXED este efecto del modelo es parte de la instrucción RANDOM. Esto indica que el factor BLOCK*VARIEDAD es un factor aleatorio y no fijo.

La opción DDFM=SATTERTH en la instrucción MODEL invoca el procedimiento Satterthwaite para obtener los grados de libertad correctos. Una vez incluida, esta opción controla el concepto de grados de libertad para la tabla "Test of Fixed Effects" (Prueba de Efectos Fijos) de la salida del programa y de las instrucciones LSMEANS, ESTIMATE y CONTRAST. En nuestro ejemplo, las pruebas que son afectadas por esta opción son las comparaciones de medias de parcelas principales a un nivel dado de sub-parcela.

Programa completo de SAS para analizar el ejemplo.

```
title3 ‘Análisis de parcelas divididas con PROC MIXED’;
data alfalfa;
input variedad $ fecha block rend;
cards;
lad 1 1 2.17
lad 1 2 1.88
lad 1 3 1.62
- continúan los datos Cuadro 1 -
;
run;
proc mixed data=alfalfa method=reml;
class block variedad fecha;
model rend = block variedad fecha fecha*variedad /ddfm=satterth;
random block*variedad;
/* comparacion entre medias de PP */
contrast ´cos vs lad´ variedad 1 -1 0;
/* comparación entre medias de SP */
contrast ´fecha1 vs fecha2´ fecha 1 -1 0 0;
/* comparación entre SP de una PP */
contrast ´fecha1 vs fecha 2 de Cos´ fecha 1 -1 0 0
fecha*variedad 1 -1 0 0 0 0 0 0 0 0 0 0;
/* comparación entre PP de una SP */
contrast ´Cos vs Lad en fecha1´ variedad 1 -1 0
fecha*variedad 1 0 0 0 -1 0 0 0 0 0 0 0;
/* calculando y comparando medias por LSMEANS */
lsmeans variedad fecha fecha *variedad / pdiff;
run;
```

UNIDAD 11
EXPERIMENTOS FACTORIALES

Julio Gabriel Ortega,

Alfredo Valverde Lucio

José Alcívar Cobeña

Caracteristicas

Los experimentos factoriales en realidad no son diseños experimentales, sino un diseño de tratamientos y nos permiten estudiar varios factores simultáneamiente con muy poco trabajo adicional; aumentan la precision, la cobertura y la utilidad de los resultados al proporcionar información sobre las interacciones entre los factores en prueba.

Un factor es un tipo de tratamiento y en un experimento factoral cualquier factor tendra varios tratamientos asociados (llamados niveles). Por ejemplo, si dieta es un factor entonces varias dietas diferentes seran estudiadas. El concepto de un experimento factorial se ilustra con el siguiente ejemplo.

Supongamos que se desea introducir vacas lecheras a un pais y no se sabe nada acerca de cual es la mejor raza lechera o el mejor forraje, para sacar buen rendimiento en calidad y cantidad de leche. Un procedimiento seria hacer dos experimentos independientes para determinar la mejor raza y el mejor forraje. Desafortunadamente no podemos hacer ningun estudio sobre calidad de forraje sin alimentar en alguna de las razas lecheras y si comparamos las razas lecheras, sin uso de diferentes forrajes no obtendremos la respuesta que buscamos. Podria suceder que el efecto de los forrajes en la diferentes razas sea completamente diferente. Por tanto, las conclusiones que han sido sacadas laboriosamente sobre el mejor tipo de forraje para una raza lechera pueden ser inaplicables a la raza que consideramos mejor al final del estudio.

La presencia o ausencia de efectos principales no dice nada acerca de la presencia o ausencia de la interacción o viceversa. Debemos considerarlos separadamente.

Una interacción significativa implica, que los efectos de los factores no son independientes entre si. En este caso no podemos concluir separadamente que una raza lechera es la mejor y que un tipo de forraje es el mejor sin estudiar mas a fondo como se comporta cada raza lechera con los diferentes tipos de forraje o los tipos de forraje con cada raza lechera.

La solución al problema es conducir un experimento factorial en el que se hace un diseño de tratamietos que contemplen todas las combinaciones de los factores en consideración.

Un arreglo mas simple es el 2^2 donde dos factores, A y B tienen dos niveles para cada uno (a_0,a_1.b_0,b_1), los que en combinanciones dan cuatro tratamientos: a_0b_0, a_0b_1, a_1b_0, a_1b_1. Por convención el nivel bajo para ambos factores se denomina como 1, el resto se expresa como a, b y ab. Los efectos principales de A y B pueden ser cálculados a partir del promedio de cada tratamiento dentro de una tabla de doble entrada para los niveles de A y B, asi:

$$A = \frac{a_1b_0 + a_1b_1}{2} - \frac{a_0b_0 + a_0b_1}{2}$$

$$B = \frac{a_1b_1 + a_1b_1}{2} - \frac{a_0b_0 + a_1b_1}{2}$$

La interaccion es el promedio de la diferencia entre el efecto de A al nivel alto de B y el efecto de A al nivel bajo de B, asi:

$$AB = \frac{1}{2n}[ab+(1)-a-b]$$ donde n = numero de repeticiones

Una forma de visualizar los efectos principales y observar si existe interacción es gráficando los resultados promedios obtenidos de cada factor con su respectivo nivel. Si las líneas son paralelas no existe interacción y la interpretación de los resultados debe ser realizada solo por efectos principales. En caso contrario, si hubiera interacción significativa, debe explicarse la interacción; aspecto que se dificulta cuando existen mas de dos factores.

A continuación se describen el arreglo de tratamientos de 2^3 con los efectos posibles de obtener. Los signos + y – ayudan a establecer los efectos posibles para cada factor. La tabla 11.1 se estructura considerando los niveles altos con el signo + y los de bajo nivel con el signo -, los demás son calculados por el producto algebraico de los signos en el cual pertenece.

Tabla 11.1. Arreglo de tratamientos de un experimento factorial 2^3.

Combinación de tratamientos	**Factores**							
	1	**A**	**B**	**AB**	**C**	**AC**	**BC**	**ABC**
(1)	+	-	-	+	-	+	+	-
A	+	+	-	-	-	-	+	+
B	+	-	+	-	-	+	-	+
Ab	+	+	+	+	-	-	-	-
C	+	-	-	+	+	-	-	+
Ac	+	+	-	-	+	+	-	-
Bc	+	-	+	-	+	-	+	-
Abc	+	+	+	+	+	+	+	+

Ejemplo 11.1. Diseño factorial 4 x 3 + 1

En un experimento realizado en la finca Andil de la UNESUM por la Ing. Gladys Holguin Flores (Holguín-Flores 2019), se estudio el desarrollo morfológico del cultivo de café en la etapa de crecimiento, se probaron dos factores. **Factor A**: cuatro tipos de fertilizantes (A1: Microrrizas, A2: Humus de lombriz, A3: yeso agrícola, A4: Micro esencial) y **Factor B**: tres diferentes dosis (B1, B2, B3). Todos los tratamientos fueron comparados a un testigo local (urea). La parcela de estudio fue implementada en un diseño experimental de bloques completamente aleatorios con tres repeticiones y los tratamientos fueron analizados en un diseño factorial 4 x 3 + 1. Los datos se muestran en la Tabla 11.1.1

Tabla 11.1. 1. Datos de campo de número de ramas en diseño factorial 4 x 3 + 1.

trat	rep	FACTOR A	FACTOR B	Números de ramas
1	1	micorriza	dosis 1	22.33
2	1	micorriza	dosis 2	22.33
3	1	micorriza	dosis 3	15.00
4	1	humus de lombriz	dosis 1	19.67
5	1	humus de lombriz	dosis 2	27.67
6	1	humus de lombriz	dosis 3	22.00
7	1	yeso agrícola	dosis 1	17.67
8	1	yeso agrícola	dosis 2	23.00
9	1	yeso agrícola	dosis 3	29.33
10	1	micro esencial	dosis 1	14.33
11	1	micro esencial	dosis 2	17.00
12	1	micro esencial	dosis 3	14.33
1	2	micorriza	dosis 1	28.00
2	2	micorriza	dosis 2	11.00
3	2	micorriza	dosis 3	18.67
4	2	humus de lombriz	dosis 1	24.00
5	2	humus de lombriz	dosis 2	31.67
6	2	humus de lombriz	dosis 3	19.33
7	2	yeso agrícola	dosis 1	29.67
8	2	yeso agricola	dosis 2	29.67
9	2	yeso agrícola	dosis 3	27.00
10	2	micro esencial	dosis 1	12.00
11	2	micro esencial	dosis 2	15.67
12	2	micro esencial	dosis 3	15.33
1	3	micorriza	dosis 1	19.00
2	3	micorriza	dosis 2	20.00
3	3	micorriza	dosis 3	15.00
4	3	humus de lombriz	dosis 1	27.00
5	3	humus de lombriz	dosis 2	31.67
6	3	humus de lombriz	dosis 3	21.67
7	3	yeso agrícola	dosis 1	30.67
8	3	yeso agrícola	dosis 2	31.33

9	3	yeso agricola	dosis 3	24.33
10	3	micro esencial	dosis 1	21.00
11	3	micro esencial	dosis 2	21.67
12	3	micro esencial	dosis 3	14.00
13	1	Testigo		12.33
13	2	Testigo		8.00
13	3	Testigo		9.33

Resolución manual del problema

1. Sobre la base de la tabla de datos, se debe elaborar las tablas de contingencia para hacer los cálculos.

Tabla 11.1.2. Datos en función de dósis y tipos de fertilizantes

Dosis	Fertilizantes				Suma
	1	2	3	4	
1	69	71	78	47	265
2	53	91	84	54	283
3	49	63	81	44	236
Suma	171	225	243	145	784
					813.67
	testigo	29.67			

2. Cálculo del factor de corrección (C). Existen dos factores que debemos calcular, por una parte debido al factorial y por otra al completo.

$C_f = (813)^2/rt = (813)^2/3x12 = (813)^2/39 =$ **17.04,** este calculo es debido al factorial

$C_c = (813)^2/rt = (813)^2/3x13 = (813)^2/39 =$ **16.98,** este cálculo es debido al completo

3. Cálculo de la suma de cuadrados totales (SCTotales)

$SCTotal = [(22.33)^2 + (22.33)^2 + \ldots + (9.33)^2] - C =$ **1715.5**

4. Cálculo de la suma de cuadrados de repeticiones (SCR)

Para determinar la SCR se debe elaborar una tabla, que permita determinar la producción promedio de las repeticiones.

Factorial +1	**Producción**	**(Producción)^2**
R1	257	66.05
R2	270	72.90
R3	287	82.18
Suma	813.67	221.13

Ahora estamos en condición de calcular la SCR

$SCR = \{[(257)^2 + [(270)^2 + [(287)^2]/t\} - C = (221.13/13) - C =$ **34.00**

5. Cálculo de la suma de cuadrados de tratamientos (SCT)

Para el cálculo de la SCT se utiliza la tabla 11.2, elaborada en función a las dosis y fertilizantes.

Se debe tener cuidado en el cálculo de la SCT, debido a que tenemos uno debido al experimento factorial y otro debido al experimento completo. Es decir, en otras palabras, debemos calcular dos sumas de cuadrado para tratamiento, uno factorial y otro completo ya que con el valor factorial determinaremos la interacción

Cálculo de la suma de cuadrados de tratamiento factorial

SCT = $\{[(69)^2 + (53)^2 + \ldots + (44)^2]/r\}$ - C = $\{[(69)^2 + (53)^2 + \ldots + (44)^2]/3\}$ – **17.04** = **929.6**

Cálculo de la suma de cuadrado de tratamiento completo

SCT = $\{[(69)^2 + (53)^2 + \ldots + (44)^2]/r\}$ - C = $\{[(69)^2 + (53)^2 + \ldots + (44)^2]/3\}$ – **16.98** = **1321.0**

6. Cálculo de la suma de cuadrados de Fertilizantes (SCF). Para este cálculo se utiliza los totales de fertilizantes obtenidos en la Tabla 11.2

SCF = $\{[(171)^2 + \ldots + [(145)^2]/rxd\}$ – C = $\{[(171)^2 + \ldots + [(145)^2]/3x3\}$ – C = **686.1**

7. Cálculo de la suma de cuadrados de dosis (SCD). Para este cálculo se utiliza los totales de dosis obtenidos en la Tabla 11.2

SCD = $\{[(265)^2 + (283)^2 + (236)^2]/rxf\}$ – C = $\{[(265)^2 + (283)^2 + (236)^2]/3x4\}$ – C = **92.7**

8. Cálculo de la suma de cuadrados de la interacción fertilizantes x dosis (SCFxD). Para esto se utiliza la SCT debido al factorial

SCF x D = SCT – SCF – SCD = **929.6** – 686.1 – 92.7 = **150.8**

9. Cálculo de la suma de cuadrados del testigo versus el resto. Este valor se calcula restando la suma de cuadrados de tratamiento completo menos la suma de cuadrados del tratamiento factorial

SCTvsR = 1321.0 – 929.6 = **391.4**

10. Cálculo de la suma de cuadrados del error experimental (SCE)

SCE = SCTotal – SCT – SCR = 1715.5 – 1321.0 – 34.0 **= 360.4**

11. Construcción de la tabla de análisis de varianza. En este caso particular se debe tener dos tablas, uno debido al diseño factorial y otro debido al diseño completo.

La Tabla de análisis de varianza factorial quedaría como sigue, misma que fue utilizada para el cálculo construir la Tabla 11.1.3:

Factorial			ADEVA Factorial
r	3	FC	17.074
t	12	SC Total	1.314
rt	36	SCT	929,6
F	4	SCR	44,5
D	3	SCF	686,1
		SCD	92,7
		SC F VS D	150,8
		SCE	340,1
		SC Testigo vs Resto	

Entonces la Tabla de análisis de varianza completo quedaría como sigue (Tabla 11.1.3).

Tabla 11.1.3. Análisis de varianza completo

Fuente de variación	Grados de libertad	Suma de Cuadrados	Cuadrados Medios	F calculada	$F_{0.05}$	$F_{0.01}$
Repeticiones	2	34.0	17.01	1.09	3.42	5.66
Tratamientos	12	1321.0	110.09	7.03	2.20	3.07
Fertilizantes	3	686.1	228.71	14.59**	3.03	4.76
Dosis	2	92.7	46.37	2.96**	3.42	5.66
Interacción F x D	6	150.8	25.13	1.60	2.53	3.71
Testigo vs Resto	1	391.4	391.42	24.98**	4.28	7.88
Error Experimental	23	360.4	15.67			
Total	38	1715.5				

**: Altamente significativo al p<0.01 de probabilidad.

12. Análisis de comparación de medias mediante la prueba múltiple de Tukey

a. Calculamos las medias de tratamiento

FACTOR A	FACTOR B	Medias	n	E.E.
humus de lombriz	dosis 2	30.34	3	2.31
yeso agricola	dosis 2	28.00	3	2.31
yeso agricola	dosis 3	26.89	3	2.31
yeso agricola	dosis 1	26.00	3	2.31
humus de lombriz	dosis 1	23.56	3	2.31
micorriza	dosis 1	23.11	3	2.31
humus de lombriz	dosis 3	21.00	3	2.31
micro esencial	dosis 2	18.11	3	2.31
micorriza	dosis 2	17.78	3	2.31
micorriza	dosis 3	16.22	3	2.31
micro esencial	dosis 1	15.78	3	2.31
micro esencial	dosis 3	14.55	3	2.31

b. Determinamos la Diferencia Significativa Honesta (DSH)

$$DSH = q_{\alpha,t,n}\sqrt{\frac{S^2}{r}}$$

DSH = $q_{0.05, 6,23}$ $\sqrt{15,67/3} = 4.39\sqrt{15,67/3} =$ **10.03**

FACTOR A	FACTOR B	Medias	
Humus de lombriz	dosis 2	30.34	a
Yeso agrícola	dosis 2	28.00	ab
Yeso agrícola	dosis 3	26.89	abc
Yeso agrícola	dosis 1	26.00	abcd
Humus de lombriz	dosis 1	23.56	abcd
Micorriza	dosis 1	23.11	abcd
Humus de lombriz	dosis 3	21.00	abcd
Micro esencial	dosis 2	18.11	bcd
Micorriza	dosis 2	17.78	bcd
Micorriza	dosis 3	16.22	bcd
Micro esencial	dosis 1	15.78	cd
Micro esencial	dosis 3	14.55	d

Interpretación: Valores con las mismas letras están indicando que no son significativamente diferentes al $p<0.05$ de probabilidad.

Lo que interesa es ver a que se debe esa diferencia significativa de la interacción. Para lo que se calcula la DSH.

Las diferencias de las medias que son menores al DSH calculado no son significativos al $p<0.05$ de probabilidad.

Se observa en el ejemplo realizado, que el tratamiento Humus de lombris con dósis 2 es el mejor tratamiento, respecto al tratamiento Micro esencial con dosis 3.

Ejemplo 11.2. Diseño factorial 4 x 2

En un experimento realizado en la finca Andil de la UNESUM por la Ing. Karen Quijije Quiroz (Quijije Quiróz 2018), se criaron 432 pollos, distribuidos en un diseño completamente aleatorio y analizados en arreglo factorial 4 x 2, donde los tratamientos fueron la combinación de dos factores. Factor D: tres dosis de ácidos orgánicos y el testigo que utiliza promotores de crecimiento farmacológico (oxitetraciclina) (D1, D2, D3, D4). El Factor S: Sexo (M: Machos, H: Hembras). Cada tratamiento fue aplicado a 36 pollos machos y 36 pollos hembras. Los pesos de los pollos a las seis semanas se observan en la Tabla 11.2.1.

Tabla 11.2.1 Pesos en kg de pollos hembras y machos despues de seis semanas.

Sexo	**Tratamientos**	**Dosis**			
		D1	**D2**	**D3**	**D4**
M	1,0 Cc/L Enrofloxacina	1,81	1,82	1,88	1,84
M	1,0 Cc/L Acido Orgánico	1,76	1,78	1,86	1,87
M	2,0 Cc/L Acido Orgánico	1,80	1,81	1,79	1,89
M	3,0 Cc/L Acido Orgánico	1,79	1,82	1,90	1,83
H	1,0 Cc/L Enrofloxacina	2,07	1,99	2,06	2,03
H	1,0 Cc/L Acido Orgánico	2,13	2,01	2,08	2,10
H	2,0 Cc/L Acido Orgánico	2,03	2,07	2,02	2,02
H	3,0 Cc/LAcido Orgánico	2,08	2,02	2,07	2,09

Ejemplo 11.1.

Se implementò un experimento para evaluar siete raciones de Isaño, aplicados en tres etapas, como alternativa para reemplazar las dietas en base de maíz. Los tratamientos fueron aplicados en diseño de bloques completamente aleatorizados con tres repeticiones y 30 pollos (unidades experimentales) por repetición. Las variables de respuesta fueron: Peso corporal (PC), Ganancia de peso (GP), Alimento consumido (AC) y Conversión alimenticia (CA). Estas variables fueron evaluadas durante tres semanas (etapas) consecutivas. Los análisis de varianza y de comparación de medias fueron realizados en un experimento factorial 7 x 3.

```
Data pollo2;
Input racion etapa  rep gp ca acon pp;
Cards;
1    1    1    0.190    1.454    0.244    0.190
1    2    1    0.930    0.610    0.347    0.930
1    3    1    1.884    2.502    0.891    1.884
1    1    2    0.188    1.168    0.284    0.188
1    2    2    0.908    0.532    0.383    0.908
1    3    2    1.858    2.397    0.869    1.858
1    1    3    0.196    1.510    0.239    0.196
1    2    3    0.927    0.535    0.373    0.927
1    3    3    1.761    2.487    0.853    1.761
2    1    1    0.222    1.608    0.249    0.222
2    2    1    0.913    0.484    0.348    0.913
2    3    1    1.836    2.464    0.891    1.836
2    1    2    0.198    0.829    0.259    0.198
2    2    2    0.828    0.365    0.415    0.828
2    3    2    1.897    1.781    0.914    1.897
2    1    3    0.196    1.384    0.238    0.196
2    2    3    0.868    0.454    0.375    0.868
2    3    3    1.875    2.348    0.889    1.875
3    1    1    0.168    1.456    0.208    0.168
3    2    1    0.927    0.692    0.357    0.927
3    3    1    1.838    2.479    0.897    1.838
3    1    2    0.196    1.554    0.239    0.196
3    2    2    0.886    0.527    0.365    0.886
3    3    2    1.739    2.512    0.878    1.739
3    1    3    0.185    1.360    0.246    0.185
3    2    3    0.877    0.574    0.380    0.877
3    3    3    1.635    2.373    0.877    1.635
4    1    1    0.176    1.348    0.239    0.176
4    2    1    0.931    0.579    0.321    0.931
4    3    1    1.752    2.450    0.883    1.752
4    1    2    0.191    1.308    0.249    0.191
4    2    2    0.905    0.517    0.394    0.905
4    3    2    1.801    2.323    0.900    1.801
4    1    3    0.201    1.566    0.241    0.201
4    2    3    0.876    0.514    0.393    0.876
4    3    3    1.795    2.526    0.847    1.795
5    1    1    0.208    1.472    0.252    0.208
5    2    1    0.879    0.551    0.383    0.879
5    3    1    1.760    2.391    0.888    1.760
5    1    2    0.220    1.611    0.247    0.220
5    2    2    0.822    0.416    0.344    0.822
5    3    2    1.728    2.448    0.889    1.728
5    1    3    0.216    1.474    0.242    0.216
5    2    3    0.750    0.450    0.351    0.750
5    3    3    1.677    2.280    0.877    1.677
6    1    1    0.201    1.501    0.247    0.201
6    2    1    0.765    0.379    0.378    0.765
6    3    1    1.814    2.464    0.895    1.814
6    1    2    0.181    1.363    0.241    0.181
6    2    2    0.832    0.604    0.325    0.832
```

```
6    3    2    1.655    2.389    0.886    1.655
6    1    3    0.193    1.322    0.249    0.193
6    2    3    0.795    0.385    0.413    0.795
6    3    3    1.599    2.260    0.885    1.599
7    1    1    0.164    0.886    0.255    0.164
7    2    1    0.852    0.505    0.414    0.852
7    3    1    1.797    2.095    0.903    1.797
7    1    2    0.188    1.609    0.205    0.188
7    2    2    0.892    0.456    0.396    0.892
7    3    2    1.872    2.498    0.888    1.872
7    1    3    0.175    1.332    0.227    0.175
7    2    3    0.827    0.483    0.359    0.827
7    3    3    1.647    2.339    0.869    1.647
;
Proc glm;
Classes racion etapa;
Model gp ca  acon pp= racion etapa racion*etapa;
Means racion etapa racion*etapa/Duncan tukey;
Run;
```

UNIDAD 12
SERIES DE EXPERIMENTOS

Julio Gabriel Ortega

Características

En la práctica de los diversos campos de investigación, es muy frecuente proyectar series de experimentos similares con el objeto de muestrear con mayor eficiencia el material experimental. Así, es muy común encontrar series de experimentos distribuidos en el tiempo y en el espacio, de las cuales presumiblemente se derivarán conclusiones más o me- nos definitivas. En unidad se examina algunas situaciones particulares del análisis de series de experimentos frecuentes.

Serie de experimentos similares sobre varias localidades

Considérese un grupo de experimentos en bloques completos al azar, que ensayan un conjunto de t tratamientos, sobre cada una de q localidades. El problema del análisis estadístico de la serie de experimentos puede examinarse desde dos puntos de vista diferentes: i) cuando las varianzas de los errores son homogéneas entre localidades, y ü) cuando las varianzas de los errores son heterogéneas de localidad a localidad. Para el primer caso, se tiene un método exacto de análisis, en tanto que para el segundo hay que recurrir a métodos aproximados.

En esta unidad solo se discutirá el análisis de una serie de experimentos bajo homogeneidad de varianzas. Se debe considerar dos aspectos del problema: primero, cuando sobre cada localidad los t tratamientos se ensayan en experimentos de r bloques completos al azar, y segundo, cuando sobre la localidad i se emplea un experimento de r_i bloques completos al azar, pudiendo ser diferentes los números r_i de una localidad a otra.

Serie de experimentos con igual número de bloques

Suponga que un conjunto de t tratamientos se ensayan en cada una de q localidades en diseños de r bloques completos al azar.

El modelo lineal apropiado para interpretar los resultados de la serie de experimentos viene dado por la relación:

$$Y_{ijk} = \mu + \pi_i + \beta_{ij} + \tau_k + (\pi\tau)_{ik} + e_{ijk}$$

Donde μ es un efecto común a todas las observaciones, π es el efecto de la localidad i, β_{ij} es el efecto del bloque j dentro de la localidad i, τ_k es el efecto del tratamiento k, $(\pi\tau)$ es el efecto de la interacción entre el tratamiento k y la localidad i, en el bloque j y con el tratamiento k y, finalmente, e_{ijk} es el error de observación sobre la unidad experimental (ijk). Los e_{ijk} se consideran como variables aleatorias normales no correlacionadas. con media cero y varianza constante sobre todas las unidades experimentales, hipótesis que representa la particularidad de la homogeneidad de varianzas de 1os términos de error. Por otra parte si las localidades se pueden considerar como una muestra aleatoria extraída de una población infinita de localidades, como frecuentemente ocurre en las investigaciones agrícolas, los efectos de localidad r_i y los términos de interacción $(\pi\tau)_{ik}$ puede tratarse como variables aleatorias normales, no correlacionadas no dentro ni entre ellas, con media cero y varianzas δ^2_p y δ^2_{tp}, respectivamente.

Análisis de varianza combinado

Fuentes de variación	Grados de libertad	Sumas de cuadrado	Cuadrados medios
Localidades (Loc)	q-1	SCL	CML
Bloques dentro localidades	q (r-1)	SCBDL	CMBDL
Tratamientos	t-1	SCT	CMT
Tratamientos dentro localidades	(t-1)(r-1)	SCT x L	CMT x L
Error	q (r-1)(t-1)	SCE	$CME = S^2$
Total	rtq-1	$\sum Y^2_{ijk}-Y^2\ldots/rtq$	

Series de experimentos con números desiguales de bloques

Cuando se ensayan t tratamientos en cada una de las localidades en diseño de bloques completos al azar, con r_i bloques en la localidad i, el método de análisis estadístico es completamente análogo al descrito al caso anterior, pero con algunas modificaciones.

Fuentes de variación	Grados de libertad	Sumas de cuadrado	Cuadrados medios
Localidades (Loc)	q-1	SCL	CML/s^2
Bloques dentro localidades	$\sum r_i - q$	SCBDL	
Tratamientos	t-1	SCT	
Tratamientos x localidades	(t-1)(q-1)	SCT x L	$CMT\ x\ L/s^2$
Error	$(\sum r_i-q)(t-1)$	SCE	
Total	Tsumari-1		

Programa SAS para el análisis de una serie de experimentos en bloques completos al azar

```
Data Uno;
Input exp rep trat Y1 Y2 …YP;
Cards;
…
Datos
…
Proc anova;
Classes exp rep trat;
Model  Y1-YP= exp rep(exp) trat trat x exp;
Means trat exp trat x exp;
Run;
```

UNIDAD 13

DISEÑO DE BLOQUES INCOMPLETOS CON UNA REPETICIÓN

Julio Gabriel Ortega

Introducción

Consideramos que el disño experimental que presentamos, es muy importante cuando se evaluan muchas tecnologìas como cultivares, lineas, clones, etc. El Dr. Melicio Siles Cano, un profesional boliviano de reconocida trayectoria internacional, planteó este tipo de experimentos, que nos parecio fundamental compartirlo con nuestros estudiantes en este libro. Mayores detalles del mismo lo pueden encontrar en Siles Cano (2005).

Sabemos que un programa de mejora genética de cultivos autógamos (arveja, cebada, frijol, soya, trigo, etc.) y de reproducción asexual (papa, caña de azúcar, etc) se desarrollan una gran cantidad de líneas y clones, respectivamente, con la finalidad de identificar los más sobresalientes. En especies alógamas, como el maíz, también se desarrollan un número grande de líneas, inicialmente con el objetivo de identificar las más apropiadas por su aptitud combinatoria general (ACG) y las combinaciones híbridas de estas con mejor aptitud combinatoria específica (ACE) (Fehr 1993).

Las líneas desarrolladas en autógamas (F4, F5) y los clones en especies de reproducción asexual cuentan inicialmente con poca cantidad de semilla y material vegetal, respectivamente, que en la mayoría de los casos alcanza solo para un surco de uno a tres metros de largo, por ejemplo, en arveja solo es suficiente para un surco de un metro de largo. Bajo estas condiciones, las selecciones se basan simplemente en apreciaciones subjetivas, con el gran peligro de descartar algunos genotipos potenciales o para evitar este riesgo algunas veces primeramente se incrementa el material, aumentando más el tiempo en la obtención de una variedad.

En otras situaciones, aún cuando se cuenta con suficiente cantidad de semilla, por ejemplo, algunas autógamas como cola de zorro (*Setaria italica*, (L.) Beauv) que de una sola planta se puede cosechar más de 6000000 semillas (Malm y Rachie 1971), o en maíz, el híbrido que resulta de cada línea cruzada a un probador para evaluar por su ACG o los híbridos entre distintas líneas (simples, de tres vías o dobles) puede disponer de más de 1000 semillas (Fehr, 1993), debido a la gran cantidad de material resultante (por ejemplo, 4950 híbridos simples en base solo a 100 líneas), los ensayos se establecen con una sola repetición y las selecciones también son muy subjetivas y el peligro de descartar buenos genotipos persiste, aunque este riesgo podría reducirse de acuerdo a la experiencia del mejorador.

Las evaluaciones del material genético desarrollado se realizan mediante el uso de diferentes diseños experimentales, completamente aleatorio, bloques completos aleatorizados, o filas y columnas. Con estos diseños se pueden evaluar una reducida cantidad de genotipos con al menos 2 repeticiones. Para el caso de un gran número de genotipos se desarrollaron los diseños de látices cuadrados y rectangulares. Sin embargo, al igual que los diseños elementales, requieren que los genotipos sean establecidos por lo menos con 2 repeticiones. Consecuentemente, estos diseños mencionados no serian apropiados para la evaluación del material genético desarrollado con una sola repetición. Lo cual sugiere la necesidad de buscar nuevas herramientas estadísticas

que permitan evaluar apropiadamente la gran cantidad de material genético desarrollado, cuya semilla o material vegetal no es suficiente para más que una repetición o para aprovechar mejor aquellos ensayos que se establecen con una repetición.

Análisis estadístico

Los datos de cada una de las variables de respuesta que se pueden considerar, de acuerdo a los objetivos de una investigación, previa verificación o aproximación mediante transformaciones a los supuestos, principalmente distribución normal y homogeneidad de varianzas, se analizan de acuerdo al siguiente modelo estadístico:

$$Yij = \mu + ßi + \tau j + \xi ij$$

donde:

Yij = valor observado de una variable de respuesta en una unidad experimental del i-ésimo

μ = media general

ßi = efecto aleatorio del i-ésimo bloque

ßi ≈ NIID (0,σ2b)

τj = efecto fijo del j-ésimo genotipo

ξij = efecto aleatorio de los residuales

$\xi ij \approx$ NIID (0,σ2e)

En base al modelo estadístico indicado, se realizan análisis de varianzas para probar hipótesis acerca de los efectos fijos y estimar componentes de varianza para los efectos aleatorios considerados en el modelo, de acuerdo a la teoría de los modelos mixtos (Searle *et al.,* 1992) utilizando el PROC MIXED de SAS (SAS 2004) ó algún programa estadístico equivalente. Las estimaciones de las varianzas de bloques y los residuales se logran únicamente sobre los efectos de los terceros genotipos; sin embargo, estos resultados son aplicables de igual manera a los nuevos genotipos, de acuerdo al supuesto de homogeneidad de varianzas definido en el modelo estadístico. Esto permite que las medias y diferencias entre medias de los nuevos genotipos sean estimadas y ajustadas con sus respectivos errores estándar, haciendo posible las pruebas de hipótesis acerca de estos parámetros.

Ejemplo 13.1.

Para demostrar la aplicación del diseño experimental desarrollado y evaluar su eficiencia, se considera la evaluación de dos conjuntos de líneas de arveja desarrolladas en Centro de Investigaciones Fitoecogenéticas de Pairumani (CIFP), 174 derivadas de la cruza entre Pea51-98-43 y Pea52-98-43 y 132 desarrolladas de la cruza entre Pea52-98-43 y Pea8-98-4 (Siles 2005). Cada grupo de 12 distintas líneas fueron distribuidas aleatoriamente en bloques de 15 unidades experimentales. Al mismo tiempo, dos cultivares, Pairumani-1 (P-1) y Pairumani-3 (P-3) de usos comerciales y considerados como testigos y la línea M-2 que disponen de suficiente cantidad de semilla fueron ubicadas al azar entre las líneas en cada uno de los bloques. La unidad experimental fue un surco de un metro de largo y espaciado a 45 cm de los otros. Esta investigación se llevó a cabo en las propiedades del CIFP durante el periodo agrícola 2002/2003 (Siles 2005). Una parte del experimento se esquematiza de la siguiente manera:

Bloque 1													
12	M-2	2	10	5	7	P-1	1	6	3	4	P-3	11	9	8
Bloque 2													
P-3	13	16	21	M-2	23	19	18	P-1	24	20	22	17	14	15
Bloque n													
93	91	P-3	89	86	95	88	90	P-1	96	87	M-2	94	92	85

En cada unidad experimental, se evaluaron la incidencia de la oidiosis en una escala de 0-5, rendimiento de grano (kg/ha), longitud de vaina (cm), peso de 100 semillas (g) y vigor (capacidad productiva y apariencia). Los datos obtenidos de cada grupo de líneas separadamente, previa verificación de la distribución normal y homogeneidad de varianzas, fueron analizados de acuerdo al siguiente modelo estadístico,

$Yij = \mu + ßi + \tau j + \xi ij$

donde:

i = 1,2,..., 15 (11) bloques,

j = 1,2,...,174 (132) líneas

Los efectos de bloques y residuales se consideraron aleatorios y NIID(0,σ2b) y NIID (0,σ2e), respectivamente. En base al modelo estadístico, se realizaron estimaciones de componentes de varianza, errores estándar para las diferencias entre medias de líneas que ocurren en el mismo o en diferentes bloques y entre líneas con los testigos. Todos estos análisis fueron llevados a cabo utilizando el PROC MIXED de SAS (SAS 2004).

Los resultados mostraron que el uso de cultivares Pairumani-1 y Pairumani-3 y la línea M-2 como terceros genotipos (testigos) permitió estimar las varianzas entre bloques y de los residuales para cada una de las cinco características y en las dos cruzas (Tabla 13.1), las mismas, bajo los supuesto de que cada observación se distribuye con los mismos parámetros, son aplicables a las líneas que han sido evaluadas con una sola repetición. Adicionalmente, las estimaciones de los componentes de varianza para cada una de las características no difieren significativamente entre cruzas, consecuentemente, el diseño experimental propuesto hace posible no solamente la evaluación de líneas de arveja con una sola repetición, sino también permite realizar comparaciones entre ellas.

Tabla 13.1. Componente de varianza estimadas para diferentes características de líneas de arveja derivadas de dos cruzas (Siles Cano 2005).

Cruza	Parámetro	Rendimiento (t/ha)	Longitud Vaina (cm)	Peso 100 semillas (g)	Vigor	Oidiosis
2 x 1	σ^2_b	0.11	0.08	0.66	0.00	0.03
2 x 4	σ^2_e	0.20	0.11	5.42	0.73	0.74
	σ^2_b	0.11	0.00	0.00	0.00	0.00
	σ^2_e	0.14	0.18	2.71	0.22	1.61

Los errores estándar estimados para las cinco características entre líneas de las dos cruzas distribuidas en el mismo y en diferentes bloques y entre líneas con un testigo (Tabla 13.2) demuestran que el diseño experimental desarrollado no solamente permite comparar genotipos distribuidos en un mismo bloque, sino también, entre genotipos ubicados en distintos bloques.

Sin embargo, estas comparaciones no tienen el mismo nivel de precisión. Los errores estándar entre líneas distribuidas en diferentes bloques son los más altos, pero no difieren significativamente entre líneas que ocurren en el mismo bloque (entre 2 a 11 %). Estos resultados demuestran que las comparaciones entre los nuevos genotipos que se han distribuido en el mismo o en diferentes bloques se logran con aproximadamente el mismo nivel de precisión.

Tabla 13.2. Errores estándar estimadas para la diferencia entre medias correspondientes a distintas características entre líneas de arveja derivadas de dos cruzas (Siles Cano 2005).

Cruza	Parámetro	Rendimiento (t/ha)	Longitud Vaina (cm)	Peso 100 semillas (g)	Vigor	Oidiosis
2 x 1	Mismo bloque	0.63	0.48	3.29	1.21	1.22
	Diferentes bloques	0.69	0.54	3.43	1.21	1.24
	Con un testigo	0.50	0.39	2.50	0.88	0.90
2 x 4	Mismo bloque	0.52	0.60	2.33	0.66	1.79
	Diferentes bloques	0.58	0.60	2.33	0.66	1.79
	Con un testigo	0.42	0.45	1.72	0.49	1.32

Para las comparaciones entre líneas con los testigos, los errores estándar son más bajos y ocurre de la misma manera con las líneas que se ubican en el mismo o en diferentes bloques, lo cual era de esperar ya que cada una de las líneas se distribuyeron junto con cada uno de los testigos en el mismo número de veces y los testigos se repiten en cada uno de los bloques. Estos resultados sugieren que las comparaciones entre los nuevos genotipos con los testigos, que es lo que comúnmente práctica el mejorador, son más precisas que entre genotipos ubicados en el mismo o en distintos bloques. Consecuentemente, las selecciones realizadas de los nuevos genotipos en relación a los testigos (terceros genotipos) se lograrán con la mejor confiabilidad.

Los resultados similares alcanzados sobre la base de las cinco características entre los dos conjuntos de líneas de arveja derivadas de dos cruzas diferentes demuestran que el diseño experimental propuesto es muy efectivo en comparar no solamente genotipos con los testigos (terceros genotipos), sino también entre genotipos distribuidos en el mismo o distintos bloques. Además, esta eficiencia podría todavía ser mejorado sí se adicionan mayor número de terceros genotipos o estableciendo bloques de unidades experimentales más homogéneas. Encontrar un tamaño ideal de la unidad experimental podría también mejorar la precisión de las estimaciones, aunque en muchos casos esto esta limitado por la disponibilidad de recursos como semilla o terreno. La adición de más terceros genotipos debe ser considerada en cada cultivo, en cada condición ambiental y dependerá principalmente de la precisión ganada en relación al consto que podría significar esta adición.

El programa general de SAS para analizar los datos de los experimentos de acuerdo al diseño experimental desarrollado se desarrolla únicamente sobre la base del PROC MIXED y consiste en lo siguiente:

```
Options is = 76 ps = 56 nodate;
Data Bl;
Input blq genot y;
Datalines;
:
```

```
Datos
:
;
Proc mixed data = Bl;
Class blq genot;
Model y = genot/ddfm = satterth;
Random blq;
Lsmeans genot/pdiff;
Run;
```

donde:
blq = variable que identifica a los distintos bloques.
genot = variable que identifica a cada uno de los genotipos.
y = variable que identifica a una variable de respuesta.
ddfm = satterth = opción que permite estimar los grados de libertad ajustados de acuerdo al método de Satterthwaite.
pdiff = opción para poder realizar todas las posibles comparaciones entre pares de medias de los distintos genotipos.

UNIDAD 14

ANÁLISIS DE VARIANZA DE MEDIDAS REPETIDAS EN EL TIEMPO

Alfredo Valverde Lucio

Julio Gabriel Ortega

Introducciòn

El ANOVA para medidas repetidas podemos considerarlo como una generalización del contraste de medias para datos relacionados (dependientes o apareados). Aquí aplicamos dos o más tratamientos a un mismo grupo de sujetos. Es una prueba, bastante más compleja que los contrastes de medias, donde podemos comprobar no solamente el efecto de varias variables intrasujetos sino también varias variables intersujetos. Podemos incluso temporalizar las diferentes medidas y tratar una nueva variable tiempo, como cuantitativa. Se consideran para su análisis: En primer lugar, el modelo de medidas repetidas para un factor intra, y en segundo lugar, el modelo para dos factores (uno intra y otro inter).

Modelo de medidas repetidas para un factor o intra sujetos. Los datos que permite analizar este modelo son los procedentes de un diseño con un solo grupo de sujetos y un único factor cuyos niveles se aplican a todos los sujetos.

Modelo de dos factores, ambos con medidas repetidas. - En un diseño de dos factores, ambos con medidas repetidas, los sujetos que participan en el experimento pasan por todas las condiciones experimentales, es decir, por todas las condiciones definidas por las posibles combinaciones entre los niveles de ambos factores.

Los investigadores del área agropecuaria frecuentemente conducen experimentos que involucran datos en cada una de varias unidades experimentales (Plantas, animales, tubos de ensayo), bajo diferentes estímulos o tratamientos (fertilizantes, dietas, drogas) y cuyo efecto se mide a través del tiempo (días, semanas, meses). En otros casos, la información se genera cuando cada uno de varios tratamientos se aplican secuencialmente a la misma unidad experimental. En general, el arreglo de medidas repetidas difiere del de parcelas divididas en que los niveles de uno o más factores no pueden ser asignados aleatoriamente a las unidades experimentales, tal como podría ser el tiempo de lectura, o en otros casos, los niveles de riego. Bajo estas circunstancias, los errores correspondientes a las unidades experimentales pueden tener o no una matriz de varianza - covarianza homogénea.

Estudio realizado por Littell, sobre análisis de "medidas repetidas" como respuesta a la toma de datos en el tiempo, misma que permite establecer curvas de crecimiento. Lo típico de experimentos de medidas repetidas en investigación es su utilización común en animales, plantas y humanos. Indica que las medidas repetidas corresponden a un experimento de tipo de factorial, con tratamiento y tiempo como los dos factores. Menciona, además que esta técnica ha sido utilizada hace mucho tiempo, pero han sido las Metodologías estadísticas e informáticas las que han permitido realizar un análisis de manera efectiva y eficiente.

Los objetivos del análisis de datos de medidas repetidas son para examinar y comparar las tendencias de respuesta sobre hora. Esto puede implicar comparaciones de tratamientos en

tiempos específicos, o promediados en el tiempo. También puede implican comparaciones de tiempos dentro de un tratamiento. Estos son objetivos comunes a cualquier experiencia factorial.

Análisis separado en cada punto de tiempo. - El análisis de datos en cada punto de tiempo examina efectos del tratamiento por separado en la observación individual veces y no hace comparaciones estadísticas entre veces. No se hace inferencia sobre las tendencias a lo largo del tiempo, así que este método no es realmente una medida repetida análisis. El uso del análisis en cada punto de tiempo suele ser en una etapa preliminar de análisis de datos y no es un método preferido para la publicación final porque no abordar los efectos del tiempo

Análisis univariado de varianza, Análisis univariado de varianza (ANOVA). - Es en términos generales, el método más comúnmente aplicado a datos de medidas repetidas que hacen comparaciones entre tiempos, trata los datos como si fueran de un diseño de parcela dividida con las unidades de animales como parcela completa y los animales en momentos particulares como subparcela unidades. Este enfoque también se conoce como un diagrama dividido en análisis de tiempo. Si las mediciones tienen la misma varianza en todo momento, y si pares de medidas en el mismo animal son igualmente correlacionado, independientemente del intervalo de tiempo entre las medidas, entonces el ANOVA univariante es válido desde un punto de vista estadístico y, de hecho, produce un método óptimo de análisis. La condición requerida para la validez de las pruebas ANOVA univariados es la llamada condición de Huynh-Feldt (HF) que es matemáticamente menos estricto que las varianzas y covarianzas iguales

análisis univariados y multivariados de variables de contraste de tiempo. - Las variables de contraste en los datos de medidas repetidas son combinaciones lineales de las respuestas a lo largo del tiempo para animales individuales. Un ejemplo básico de la metodología estadística está dada por el ortogonal polinomios, que representar tendencias lineales, cuadráticas, cúbicas, etc., sobre hora. Otro ejemplo es el conjunto de diferencias entre respuestas en puntos de tiempo consecutivos, es decir, cambia del tiempo 1 al tiempo 2, del tiempo 2 al tiempo 3, y así adelante. Se puede usar un conjunto de variables de contraste para analizar tendencias a lo largo del tiempo y hacer comparaciones entre tiempos en datos de medidas repetidas.

Se pueden aplicar análisis a estas nuevas variables. Esto proporciona un método para analizar mediciones repetidas, datos que evaden parte de la estructura de covarianza problemas que invalidan los análisis ANOVA univariados, el GLM proporciona cálculo automático y análisis para varias opciones comunes de contraste variables Los datos deben estar en modo multivariante para su uso de la declaración GLM REPEATED (Aplicado en SAS).

Análisis de modelo mixto usando el procedimiento MIXED. - Como se señaló anteriormente, el análisis de datos de medidas repetidas requiere atención especial a la estructura de covarianza debido a la naturaleza secuencial de los datos en cada animal. Procedimientos discutidos previamente o evite el problema (análisis de variables de contraste) o ignórelo (análisis univariado de varianza). Evitar los problemas puede resultar en análisis ineficientes, lo que equivale a desperdicio de los datos. El modelo mixto lineal general permite capacidad para abordar el problema directamente modelando La estructura de covarianza. Esta capacidad es incumplida en el procedimiento MIXED del SAS

Sistema. Hay dos pasos básicos para realizar una repetición Análisis de medidas utilizando metodología de modelo mixto. El primer paso es modelar la estructura de covarianza. El segundo paso es analizar las tendencias temporales de los tratamientos por estimación y comparación de medios.

Los autores citados expusieron la utilización de software como herramientas estadísticas para asegurar la confiabilidad de los datos, en este sentido y luego de este análisis, plateamos razones por las cuales se recomienda la utilización de medidas repetidas en el tiempo:

- Porque generalmente nos interesa el efecto acumulativo de los tratamientos
- Haciendo varios análisis aumentaría el número de pruebas F
- Cada prueba F adicional aumenta el riesgo de cometer error de Tipo I (concluir que existen diferencias Significativas entre tratamientos cuando, en realidad, los tratamientos son iguales)

Cuando no se debe aplicar medidas repetidas en el tiempo:

- El interés es el patrón de los TRT a través del tiempo
- Puede haber efectos escondidos al analizar solamente las medias
- Hay que analizar la interacción Tratamiento*Tiempo, si esta interacción es significativa ($P < 0.05$), es mejor analizar cada fecha aparte
- Una interacción puede esconder diferencias verdaderas entre Tratamiento.

Los autores antes citados, ratificaron la importancia de aplicar para datos en el tiempo, ANOVAS de medidas repetidas, y emplean software como el SPSS y SAS, en este capítulo se presenta un ejemplo en el software Infostat.

Ejemplo 14.1.

Se evaluó el comportamiento morfológico del café arábigo en la etapa de vivero a la aplicación de los bioestimulantes (Variedades): Starlite, Humega, Micorriza y Evergreen, en comparación con la Urea. Se empleó para el análisis de la variable altura, un ANOVA de medidas repetidas en el tiempo, considerando que cada mes y por el lapso de 4 meses se tomaría datos.
Para la aplicación del ejercicio, primero se tabulo y ordeno los datos en una tabla en Excel (Tabla 14.1), posteriormente se agregó la tabla al software estadístico, donde se comprobó inicialmente la normalidad de los datos (Tabla 14.2).

Tabla 14.1. Tabla de datos.

Repetición	Tratamiento	Variedad	Tiempo	Altura	Diámetro Raíz
1	1	Urea	30 minuto	23.4	0.816
1	2	Urea	60 minutos	26.5	0.816
1	3	Urea	90 minutos	27	0.947
1	4	Urea	120 minutos	27.5	1.048
1	1	Humega	30 minuto	23	0.816
1	2	Humega	60 minutos	24.5	0.816
1	3	Humega	90 minutos	25	0.947
1	4	Humega	120 minutos	25.6	0.947
1	1	Evergreen	30 minuto	20.2	0.816
1	2	Evergreen	60 minutos	21.5	0.816
1	3	Evergreen	90 minutos	23	0.816
1	4	Evergreen	120 minutos	23.6	0.816
1	1	Starlie	30 minuto	22.5	0.816
1	2	Starlie	60 minutos	24.7	0.816
1	3	Starlie	90 minutos	25.4	0.947
1	4	Starlie	120 minutos	25.7	0.947
1	1	Micorriza	30 minuto	18.7	0.816

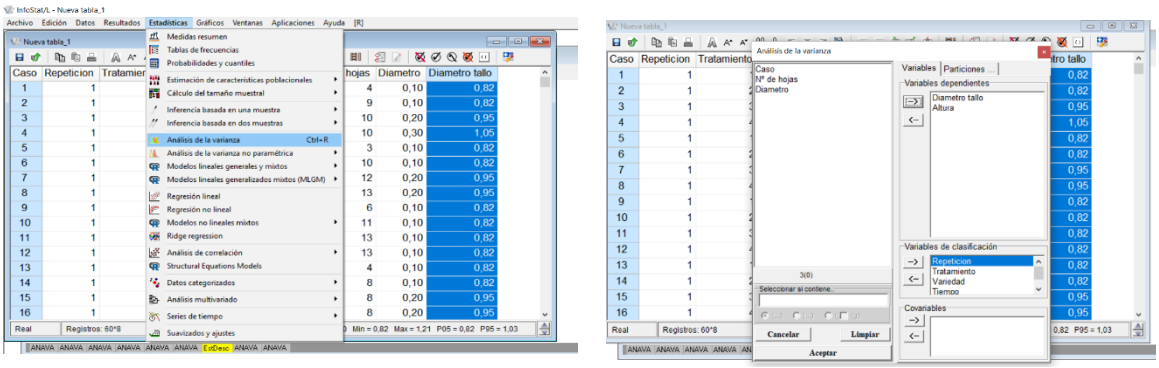
1	2	Micorriza	60 minutos	22.5	0.816
1	3	Micorriza	90 minutos	24	0.816
1	4	Micorriza	120 minutos	24.5	0.816
2	1	Urea	30 minuto	23.8	0.816
2	2	Urea	60 minutos	25	0.816
2	3	Urea	90 minutos	27	1.048
2	4	Urea	120 minutos	27.7	1.048
2	1	Humega	30 minuto	24	1.048
2	2	Humega	60 minutos	27	0.947
2	3	Humega	90 minutos	28.3	0.947
2	4	Humega	120 minutos	28.6	0.947
2	1	Evergreen	30 minuto	22.9	0.816
2	2	Evergreen	60 minutos	24.5	0.816
2	3	Evergreen	90 minutos	24.8	0.947
2	4	Evergreen	120 minutos	25	0.947
2	1	Starlie	30 minuto	20.5	0.816
2	2	Starlie	60 minutos	22.5	0.816
2	3	Starlie	90 minutos	23	0.947
2	4	Starlie	120 minutos	23.5	1.048
2	1	Micorriza	30 minuto	19.1	0.816
2	2	Micorriza	60 minutos	23	0.816
2	3	Micorriza	90 minutos	24	0.947
2	4	Micorriza	120 minutos	24.3	0.947
3	1	Urea	30 minuto	22.2	0.816
3	2	Urea	60 minutos	24	0.816
3	3	Urea	90 minutos	26.7	0.947
3	4	Urea	120 minutos	27	1.048
3	1	Humega	30 minuto	20	0.816
3	2	Humega	60 minutos	22	0.816
3	3	Humega	90 minutos	23.3	0.816
3	4	Humega	120 minutos	23.7	0.816
3	1	Evergreen	30 minuto	21.8	0.816
3	2	Evergreen	60 minutos	24.5	0.816
3	3	Evergreen	90 minutos	24.9	1.048
3	4	Evergreen	120 minutos	24.3	1.048
3	1	Starlie	30 minuto	23.8	0.816
3	2	Starlie	60 minutos	25.5	0.816
3	3	Starlie	90 minutos	26	0.816
3	4	Starlie	120 minutos	26.8	0.816
3	1	Micorriza	30 minuto	19	0.816
3	2	Micorriza	60 minutos	21	1.207
3	3	Micorriza	90 minutos	22	1.048
3	4	Micorriza	120 minutos	22.6	1.048

Tabla 14.2. Análisis de normalidad de los datos.

Variedad	Variable	n	Media	D.E.	CV	Mín	Máx	Mediana	Asimetría	Kurtosis
Evergreen	Diametro Raíz	12	0,88	0,09	10,77	0,82	1,05	0,82	1,17	-0,64
Humega	Diametro Raíz	12	0,89	0,08	9,22	0,82	1,05	0,88	0,49	-1,13
Micorriza	Diametro Raíz	12	0,91	0,13	14,49	0,82	1,21	0,82	1,25	0,01
Starlie	Diametro Raíz	12	0,87	0,08	9,37	0,82	1,05	0,82	1,25	-0,23
Urea	Diametro Raíz	12	0,92	0,11	11,95	0,82	1,05	0,88	0,30	-1,73

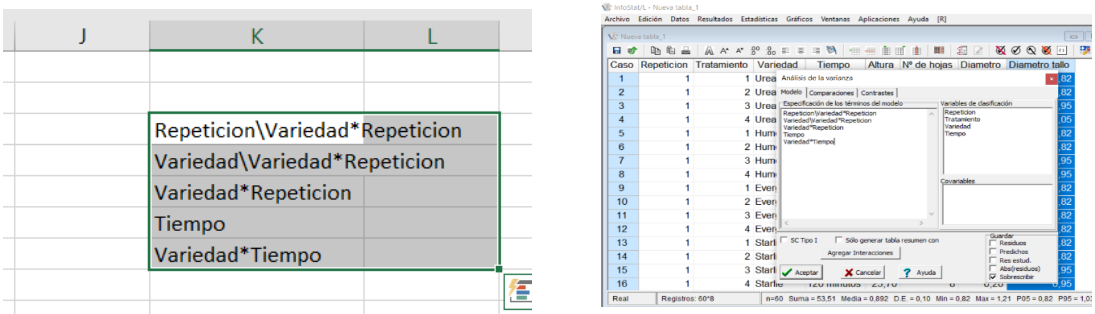
Una vez incorporados los datos en el software, se procede a realizar el análisis estadístico, tal como se demuestra en las Tabla 14.3, se va a análisis de varianza y en la siguiente ventana, se ubican las variables dependientes (diámetro y altura), y variables de estudio según corresponden.

Tabla 14.3. Guía de aplicación del ejercicio de ANOVA con medidas repetidas.



Estos pasos son iguales en los diseños simples como en los factoriales, lo cambiante es la aplicación de un modelo previamente estructurado en el excel y pegado posteriomente en el software, tal como se aprecia en la Tabla 14.4.

Tabla 14.4. aplicación del ejercicio de ANOVA con medidas repetidas.



Realizado este procedimiento se obtienen los resultados de ambas variables (Tabla 14. 5) observandose que en la variable altura, existe diferencia significativa p valor < 0.05 y en la variable diametro de tallo no existe diferencia estadistica p valor > 0.05.

Tabla 14.5. Resultados de aplicación ANOVA de medidas repetidas en ambas variables.

F.V.	SC	gl	CM	F	p-valor
Modelo	301,48	29	10,40	40,50	<0,0001
Repeticion	7,60	2	3,80	0,40	0,6861
Variedad	86,34	4	21,58	2,24	0,1537
Variedad*Repeticion	76,96	8	9,62	37,48	<0,0001
Tiempo	123,77	3	41,26	160,74	<0,0001
Variedad*Tiempo	6,80	12	0,57	2,21	0,0388
Error	7,70	30	0,26		
Total	309,18	59			

F.V.	SC	gl	CM	F	p-valor
Modelo	0,45	29	0,02	3,49	0,0005
Repeticion	0,03	2	0,02	0,79	0,4844
Variedad	0,02	4	4,9E-03	0,25	0,9036
Variedad*Repeticion	0,16	8	0,02	4,47	0,0012
Tiempo	0,16	3	0,05	11,89	<0,0001
Variedad*Tiempo	0,08	12	0,01	1,53	0,1680
Error	0,13	30	4,5E-03		
Total	0,59	59			

Para la toma de decisión es oportuno considerar las recomendaciones de analisis para el ANOVA de medidas repetidas, ante lo cual, diremos que en la variable altura al determinar interacción del tiempo durante el ciclo de crecimeinto de la planta, estableciendose que existe diferencia estadística entre tratamientos, se recomienda sin embargo realizar un análisis independiente de los factores debido a que uno es cualitativo y el otro es continuo, pudiendo realizar un polinomio ortogonal para cada factor con sus respectivos niveles, observar el comportamiento a fin de definir el modelo de analisis de regresión.

En el ANOVA de medidas repetidas de la variable diámetro de tallo p> 0.05, el resultados obtenidos, dio lugar a la aplicación de la prueba de significación de Tukey al p< 0.05 (Tabla 14.6), el ejercicio determinó diferencias estadísticas entre variedades (bioestimulantes). Apreciando que el testigo (urea) expresó mejor respuesta morfológica, y a nivel de bioestimulantes el Humega y la Micorriza, todos entre los 90 y 120 días, lo que da la pauta para comprender que, a mayor tiempo, mejor se expresa la respuesta del fertilizante y bioestimulantes en la planta de café en la etapa de vivero.

Tabla 14. 6. Prueba de Tukey al 0.05

Test:Tukey Alfa=0,05 DMS=0,21139

Error: 0,0045 gl: 30

Variedad	Tiempo	Medias	n	E.E.		
Urea	120 minutos	1,05	3	0,04	A	
Urea	90 minutos	0,98	3	0,04	A	B
Micorriza	60 minutos	0,95	3	0,04	A	B
Evergreen	90 minutos	0,94	3	0,04	A	B
Micorriza	120 minutos	0,94	3	0,04	A	B
Micorriza	90 minutos	0,94	3	0,04	A	B
Evergreen	120 minutos	0,94	3	0,04	A	B
Starlie	120 minutos	0,94	3	0,04	A	B
Starlie	90 minutos	0,90	3	0,04	A	B
Humega	90 minutos	0,90	3	0,04	A	B
Humega	120 minutos	0,90	3	0,04	A	B
Humega	30 minuto	0,89	3	0,04	A	B
Humega	60 minutos	0,86	3	0,04	A	B
Urea	60 minutos	0,82	3	0,04		B
Urea	30 minuto	0,82	3	0,04		B
Evergreen	30 minuto	0,82	3	0,04		B
Evergreen	60 minutos	0,82	3	0,04		B
Micorriza	30 minuto	0,82	3	0,04		B
Starlie	30 minuto	0,82	3	0,04		B
Starlie	60 minutos	0,82	3	0,04		B

Medias con una letra común no son significativamente

UNIDAD 15

ANALISIS T-TEST PARA SELECCION ASISTIDA POR MARCADORES MOLECULARES

Julio Gabriel Ortega

Introducciòn

El desarrollo acelerado de la biología molecular también demanda un desarrollo acelerado de la bioinformática, que hoy en día se volvió en una herramienta imprescindible para el anáisis de datos moleculares. La bioinformática es un área emergente interdisciplinaria que se ocupa de la aplicación de la informática a la recopilación, almacenamiento, organización, análisis, manipulación, presentación y distribución de información relativa a los datos biológicos, tales como macromoléculas (por ejemplo DNA o proteínas). Evolucionó para servir de puente entre las observaciones (datos) y el conocimiento que se deriva (información) sobre, por ejemplo, la función de los procesos y, posteriormente, la aplicación (conocimiento).

La bioinformática tiene un papel central en muchas áreas de la investigación en biología, como en genómica, específicamente secuenciación de genomas, mapeo, anotación y comparación de genomas. Es esencial para proteómica, permitiendo el análisis de secuencias de proteínas con el fin de determinar motivos funcionales, para la determinación de estructura de proteínas, interacciones proteína-proteína, entre otras. Asimismo, permite el descubrimiento de marcadores moleculares, como polimorfismos de un solo nucleótido (SNP), así como forma parte de los estudios de evolución y filogenia (Goodman 2002).

Esta versatilidad de la bioinformática ha permitido que hoy en día sea usada para el diseño de nuevos medicamentos y análisis forenses. En el caso del diseño de nuevos medicamentos, los estudios de interacciones proteína-ligando proveen las bases para la identificación de nuevos sitios de acción para medicamentos sintéticos, asimismo, conocer las estructuras tridimensionales de proteínas permite el diseño de moléculas que puedan unirse a un receptor de una proteína blanco con alta especificidad y afinidad (Xiong 2006).

Por otra parte, la bioinformática es de vital importancia en la secuenciación de ADN ayudando a identificar la información de importancia biológica, de manera de tener un mejor entendimiento de los organismos. Por ejemplo, la bioinformática en el campo de la biotecnología de microorganismos se emplea de diferentes formas: analizando computacionalmente la data proveniente de experimentos, secuenciación de genomas, determinación de la función de genes, construcción de árboles filogenéticos, identificación de segmentos que codifican a proteínas, entre otras (Bansal 2005).

Uso de marcadores moleculares

Es por ello que no solo es necesaria la data proveniente de los experimentos de genómica o proteómica, sino también personas formadas en esta área, capaces de interpretar dicha información. Es en este contexto que queremos contribuir en esta obra, para lo cual nos basaremos en la experiencia de Gabriel (2008) y Ritter *et al*. (2009). En estos trabajos, los autores mencionados realizaron el estudio de la resistencia en cuatro familias de cruzamientos de

especies silvestres de papa al tizón, causado por el oomycete *Phytophthora infestans* y la aplicación de marcadores moleculares para detectar genes QTLs (quantitative trait loci) o genes de caracteres cuantitativos. Para esto se aplicó marcadores polimórficos indirectos como los microsatelites o SSR - secuencias simples repetidas (Simple Sequence Repeats) y marcadores directos o de genes candidato como los cDNAs (ácido desoxirribonucleico complementario). Actualmente existen otros marcadores de genes candidato (directos) como los SNP o polimorfismo de un solo nucleótido (Single Nucleotide Polymorphism) , TDFs o fragmento de DNA objetivo (Target DNA Fragment), COS o grupo de marcadores ortólogos conservados (Conserved Ortholog Set), EST o marcador de secuencia expresada (Expressed Sequence Tag) y otros que son más modernos. Sin embargo, todos estos marcadores deben ser analizados estadísticamente para determinar el acercameinto y la detección de los genes de interés agronómico como la resistencia a enfermedades y el rendimiento.

Análisis estadístico de los alelos

Las huellas genéticas generadas por hibridación o por PCR, son heredadas a la descendencia de acuerdo las leyes mendelianas. Por lo tanto, las bandas que son polimórficas entre los progenitores se pudieron seguir en las poblaciones segregantes de las cruzas y se analizó desde el punto de vista de ligamiento en cada uno de los individuos. Es decir, los marcadores segregantes de cada progenie fueron tratados como un sistema de marcador dominante con un patrón de segregación 1:1 ó 3:1 según el modelo de segregación mendeliana (ab x aa), (aa x ab) ó (ab x ab), donde el alelo "a" representa la ausencia de la banda (0), y el alelo "b" representa la presencia de fragmentos amplificados (1). De esta forma, se dedujo fragmentos específicos de sólo un parental ab x aa; aa x ab respectivamente y factores comunes a ambos parentales (ab x ab). Los geles fueron analizados de forma visual, utilizando una matriz de presencia-ausencia del fragmento amplificado para el QTA-genotyping.

La presencia o ausencia de cada marcador segregante, tanto de la progenie como de los parentales, se anotó según la codificación mencionada. Se almacenaron en archivos con formato EXCEL para su posterior análisis estadístico.

Para cada marcador ensayado se aplicó un t-TEST del software SAS para comparar las medias de niveles de resistencia en los genotipos que pertenecen a cada clase de marcadores (ausencia/presencia) en cada caso utilizando.

Con propositos ilustrativos solo describiremos el analisis de la familia E (buk x phu), que mostrò mayor polimorfismo para los marcadores utilizados.

Programa SAS

```
TITLE ' DATA E SR.60 x MCT t-Tests ';
Data x;
input GT f1-f29 @@;
cards;
1  0 1 1 1 1 1 1 0 0 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 1
2  0 1 1 0 1 1 1 0 0 1 1 0 0 1 0 0 1 1 0 . . 0 1 0 1 0 1 1 0
3  1 0 1 0 1 0 0 1 1 0 1 0 1 1 1 0 1 1 0 0 1 0 0 1 1 0 1 0 1
4  0 1 1 0 1 0 1 1 1 1 1 0 1 1 1 1 1 0 1 0 1 1 0 0 0 1 0 0 1
5  0 1 0 1 1 1 1 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 1 0 0 0 1 0 1
6  1 1 1 0 1 1 1 0 0 1 1 0 1 0 0 1 1 0 1 1 0 1 0 1 0 1 0 1 0
7  1 1 1 0 1 0 1 0 0 1 0 1 0 1 0 0 1 1 0 1 0 1 0 1 0 0 1 1 0
8  0 1 0 1 1 1 1 0 1 0 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 0 0 0 1 1 0
9  1 1 0 1 1 1 1 0 1 1 0 1 0 0 1 0 0 1 1 0 0 1 1 0 1 0 0 1 0 1
10 0 1 1 0 1 0 1 0 0 1 1 0 0 1 1 1 1 1 0 1 0 1 1 1 0 1 0 0 1 0 1
11 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 1 0 1 1 0 1 0 1 0 1 0 1 0 1 0
12 0 1 1 0 1 1 1 1 0 1 1 1 0 0 1 1 0 1 1 0 0 1 1 0 1 0 0 1 1 0
```

```
13 1 0 1 0 1 1 1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 0 1 0 1 0 0 1
14 1 0 1 0 1 0 0 1 1 0 0 1 0 1 1 0 1 1 0 0 1 1 0 0 0 0 1 0 1
15 1 1 1 0 1 0 1 0 1 0 1 0 1 1 1 0 1 1 0 1 0 0 0 1 1 0 1 0 1
16 0 1 1 0 1 0 1 0 1 0 0 0 1 1 0 1 1 0 1 . . 0 1 0 1 1 0 1 0
17 0 1 0 1 1 1 0 1 0 1 0 1 0 .0 1 1 0 0 0 1 1 1 0 0 0 1 1 0
18 1 1 0 1 1 0 0 1 0 1 . . . . 0 0 1 1 0 1 0 1 0 1 0 0 1 1 0
19 0 1 1 1 1 0 1 1 0 1 . . .  . 1 1 1 0 0 0 1 1 0 1 0 1 0 1 0
20 1 0 1 0 1 0 1 0 1 0 0 1 1 1 1 1 1 0 0 1 1 1 0 0 0 1 0 1 0
21 0 1 1 0 1 1 1 0 1 0 1 0 0 1 1 1 1 0 0 0 1 0 0 1 1 1 0 0 1
22 1 0 1 1 1 1 0 1 1 0 1 0 0 1 1 0 1 1 0 0 1 0 1 1 0 0 1 0 1
23 1 1 0 1 1 0 0 1 1 0 0 1 0 0 1 0 1 1 0 1 0 0 1 1 0 1 0 1 0
24 0 1 0 1 1 1 0 1 1 0 0 1 0 1 0 1 1 1 1 1 0 0 1 1 0 1 0 1 0
25 1 0 1 1 1 0 1 1 1 1 1 0 0 1 1 0 1 1 0 1 0 0 1 1 0 0 1 0 1
26 1 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 1 0 0 1 1 0
27 1 0 0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0 0 1 0 1 0 0 0 0 1 1 0
28 1 1 1 0 1 1 0 1 1 0 0 1 0 1 0 0 1 1 0 1 0 1 0 1 0 0 1 1 0
29 0 1 1 0 1 0 1 1 1 1 1 0 0 1 1 1 1 0 0 0 1 1 0 0 0 0 1 1 0
30 1 1 1 0 1 0 1 0 0 1 1 0 0 1 1 0 1 1 0 1 0 1 0 0 0 0 1 1 0
31 0 1 1 0 1 1 1 1 0 1 0 1 0 1 1 0 1 1 0 1 0 1 0 0 0 0 1 1 0
33 1 0 0 1 1 0 . . . . . 0 1 0 1 1 0 0 0 1 0 0 1 1 0 1 0 1
38 0 1 1 1 1 1 0 1 1 0 1 0 1 1 0 0 1 1 0 . . 1 0 1 0 1 0 0 1
40 1 0 1 1 1 1 1 0 0 1 1 0 0 1 1 0 1 1 0 1 0 1 1 0 0 1 0 0 1
41 1 0 . . 1 1 1 0 0 1 . . . 1 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1
42 1 0 . . 1 1 1 0 1 0 0 1 0 1 1 0 1 1 0 . . 1 1 0 0 1 0 0 1
43 1 1 . . 1 1 1 0 1 0 1 0 1 1 0 1 1 0 0 1 0 0 0 1 1 1 0 0 1
44 0 1 1 0 1 1 1 0 1 0 0 1 0 1 1 1 1 0 0 0 1 1 1 0 0 0 1 1 0
45 0 1 1 0 1 0 1 1 1 1 0 1 1 0 1 1 1 0 0 0 1 . . . . 1 0 1 0
46 0 1 1 0 1 0 1 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 1 0 0 0 1 0 1
48 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 0 1 0 1
50 1 1 . . 1 1 1 0 1 0 0 1 1 . 1 1 1 0 0 0 1 1 1 0 0 1 0 1 0
51 1 1 0 1 1 0 0 1 1 0 0 1 1 . 1 1 1 0 0 1 0 1 1 0 0 1 0 1 0
53 1 1 . . 1 0 1 1 1 1 1 0 0 . 0 1 1 0 0 0 1 1 0 0 0 1 0 0 1
;
Run;
Data xq;
Input GT y IU11 @@;
Cards;
1      88.63 65.94
2      30.00 65.94
3      45.63 65.94
4       2.50 65.94
5      27.00 65.94
6       0.88 65.94
7      49.13 65.94
8     101.25 65.94
9      72.50 65.94
10     98.75 65.94
11     41.25 65.94
12     23.13 65.94
13     56.25 65.94
14     23.75 65.94
15      9.38 65.94
16      3.75 65.94
17     15.63 65.94
18     20.75 65.94
19     35.63 65.94
20    135.25 65.94
21     91.13 65.94
22    148.75 65.94
23    173.25 65.94
24    150.63 65.94
25    125.75 65.94
26    106.50 65.94
27     17.50 65.94
28     26.25 65.94
29      9.38 65.94
30     28.13 65.94
31     39.38 65.94
33    148.13 65.94
37    122.88 65.94
```

```
40  150.00 65.94
41  144.38 65.94
42  149.50 65.94
43  151.63 65.94
44  142.25 65.94
45  121.88 65.94
46  132.88 65.94
48  128.13 65.94
50  154.50 65.94
51  114.38 65.94
53  155.75 65.94
;
Data x;
merge x xq;
by  GT;
run;
options pagesize=66 linesize=130 formdlim=' ';
%macro intval(name);
%do n= 1 %to 29;
data h;
set x;
fa=&name&n;
run;
proc ttest;
Title ' TTEST for fragment: '  &name&n;
class fa;
var y;
run;
run;
%end;
%mend intval;
%intval(f);
```

Interpretación de los resultados

En la progenie E (buk 210042.5 x phu) se detectó cuatro QTLs que se encuentra en los cromosomas III, V, VI y X. Sólo un QTL en el cromosoma III viene de la fuente de resistencia P1 (*buk*) (Tabla 15.1).

Tabla 15. 1. Genotipado de QTA en la progenie E (buk x phu)

Crom	*SSR*	*QTL potencialmente asociado*	*DE*	*FR*	*V1*	*V0*	*Dif*	*Prob*
I	STM1029	Pi-1	***No segrega***					
I	STM2020a	Pi-1	***No segrega***					
I	STM2030		***No segrega***					
II	STM0038	Pi2b	**C**	1	72.42	104.23	-31.81	10.8
			P1	2	93.13	63.09	30.04	8.5
II	STM1064	Pi2c	***No segrega***					
III	STM1054	Pi3b	***No segrega***					
III	STM0040	FB-1, Pi-3a	**C**	1	67.89	88,24	-20,35	32,9
			P1	2	97.19	51,70	45,49	**1,0**
IV	STM1050	R2,Pi-4a2	***No segrega***					
V	STM1041	PiFTve-5a, Pi5a,R1, PiFTve-5c,	**P1**	1	87.80	72,23	15,56	37,4
V	STM0013	PiFTve-5b	**P1**	1	74.44	79,14	-4,70	79,2
			C	4	76.66	78,36	-1,70	93,9
			C	2	88.86	58,67	30,19	8,9
			P2	3	61.47	97,02	-35,55	**4,0**
V	STM1020	Pi-5b	***No segrega***					
VI	STM0019	Pi-6a	**P2**	1	103.09	62,72	40,37	**1,8**
			P2	2	62.72	103,09	-40,37	**1,8**
			P1	3	66.63	96,57	-29,94	8,3
			P1	4	96.57	66,63	29,94	8,3
VI	STM1100	FB-6	**P1**	1	66.75	83,53	-16,78	36,6
			P1	2	88.85	63,75	25,10	18,0
			P2	3	66.67	81,38	-14,71	44,6
VI	STM1056a	PI-6b	**No segrega**					
VII	STM1003	Pi7b	**C**	1	81.02	88,19	-7,17	77,9
			P2	2	82.79	81,52	1,28	94,5
VII	STM0052	Pi7b	**P2**	1	83.68	76,62	7,05	68,9
VIII	STM1056b	E-5	**No segrega**					
VIII	STM1024	Rblc, E-6	**C**	1	83.28	76,77	6,51	71,4
			P2	2	80.66	80,48	0,19	99,2
VIII	STM1005	Pi-8	**P2**	1	84.46	76,09	8,37	64,5
			P2	2	80.22	80,96	-0,74	96,8
IX	STM1102		**No segrega**					
IX	STM1051		**No segrega**					
IX	STM3012	Pi-9	**No segrega**					
X	STM0051	Rber	**C**	1	69.58	114,99	-45,41	**3,1**
			P2	3	107.49	63,66	43,83	**1,4**
			P1	2	83.84	74,33	9,52	59,2
XI	STM1009	Pi-11	**No segrega**					
XI	STM0037	Pi-11	**No segrega**					
XI	STM0025	RP-11,Pi-19	**No segrega**					
XII	STM0007		**No segrega**					
XII	STM0030		**No segrega**					
XII	STM2028	Pi-12	**No segrega**					

Crom = cromosoma; SSR: marcador microsatélite; .DE = Fragmento de un descendiente [P1 = Fragmento específico para el parental 1, P2 = Fragmento específico para el parental 2, C = Fragmento común de ambos parentales]; Fr = Numero de fragmentos; V1, V0 = Valor promedio de AUDPC para cada genotipo donde un marcador molecular particular está presente o ausente respectivamente; Dif = Diferencia entre V1 - V0; Prob = probabilidad para la existencia del QTL [%]. Los marcadores significativos de los QTLs están remarcados en negrilla.

En la progenie E se detectaron cuatro QTLs de resistencia a *P. infestans* en los cromosomas V (PiTFve-5b), VI (Pi-6a) y X (Rber), que provienen de *S. phureja* y uno en el cromosoma III (FB-1, Pi-3a), del parental *S. bukasovii,* lo que muestra que ambos parentales son fuentes valiosas de resistencia a *P. infestans*. Estos marcadores mostraron significancia al nivel de $P<0.05$ de probabilidad en la prueba de t-test.

PARTE V
ANALISIS DE REGRESION, CORRELACION Y COVARIANZA

UNIDAD 16
ANÁLISIS DE REGRESIÓN Y CORRELACIÓN

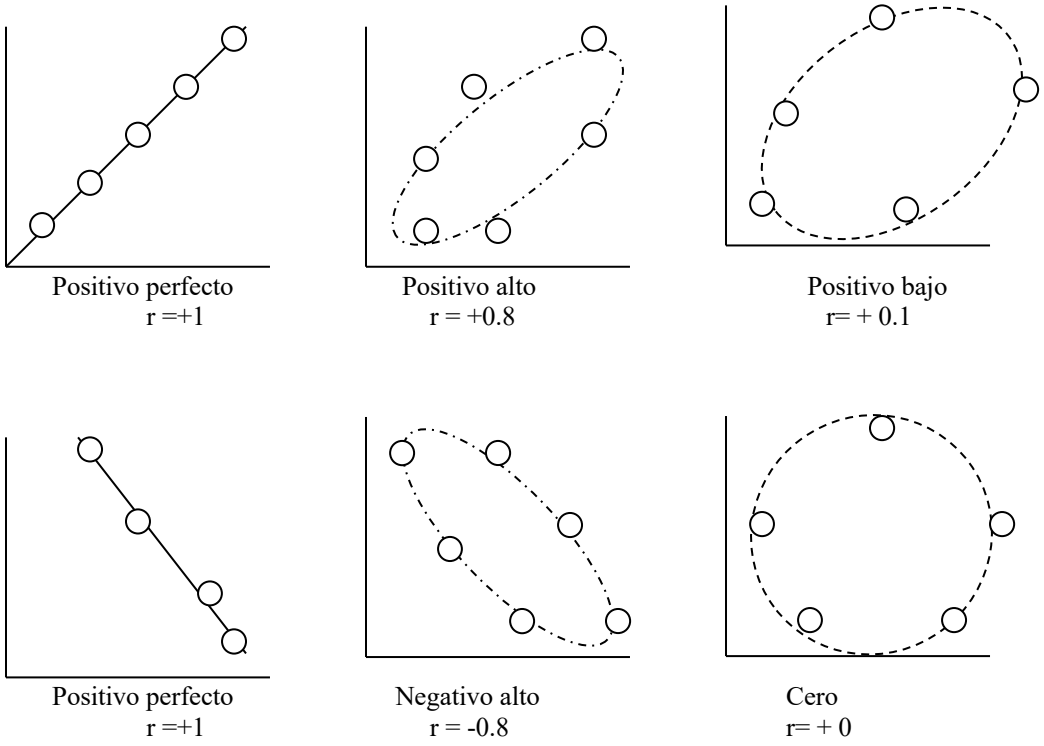
Julio Gabriel Ortega

Raquel Vera Velázquez

Alfredo Valverde Lucio

Características

El científico a veces esta interesado en la relación que existe entre dos o mas variables. Por ejemplo, interesante saber la interdependencia que existe entre la edad de un animal y su peso, en este caso hablamos de correlación positiva si al aumentar la edad del animal aumenta su peso, y da correlación negativa si al aumentar la edad del animal disminuye su peso. La correlación (R) puede tomar valores de -1 a 1. Un valor de R igual a 1 denota una correlación positiva perfecta; contrariamente, un valor de R igual a -1 denota una correlación negativa perfecta.



Por ejemplo, tenemos los datos de 12 corderos de diferentes años y pesos, asi como se observa en la siguiente tabla:

Edad (años)	1	3	2	4	2	3	2	1	2	3	4	5
Peso (Kgs)	20	40	32	48	21	39	34	17	31	37	51	55

En el programa SAS seria:

```
Data pesos;
Input edad peso;
Cards;
1 20
3 40
2 32
4 48
2 21
3 39
2 34
1 17
2 31
3 37
4 51
5 55
proc corr;
var edad peso;
run;
```

Si el conjunto de datos tiene más variables pueden incluirse todas en la lista de variables. La salida de este programa es:

Variable	N	Mean	STV DEV	SUM	MINIMUN	MAXIMUM
Edad	12	2.667	1.2309	32.00	1.000	5.000
Peso	12	35.417	12.2062	425.00	17.000	55.000

Pearson Correlation coeficients / Prob |R| Under H0: RH0=0 / M= 12

	Edad	Peso
Edad	**1.000**	0.96 (1)
	0.000	0.001(2)
Peso	0.9600	**1.000**
	0.001	0.000

Como se puede ver, SAS produce una matriz de correlación con el coeficiente de correlación (1) (por ejemplo, el coeficiente de correlación entre peso y edad es 0.96, ó sea que hay una relación positiva muy alta entre edad y peso). El valor que se encuentra debajo del coeficiente de correlación (R) es la probabilidad (2) de que el coeficiente sea cero (o sea que no haya correlación). En el ejemplo vemos que la probalidad de que la correlación sea casi cero, es decir la probabilidad de que haya correlación, es estadísticamente altamente significativa (1 – 0.001) al 99.9%.

Regresión

La regresión es la cantidad de cambio de una variable asociada a un cambio único de otra variable. Esta definición es susceptible de critica en aquellas áreas en que resulta insuficientemente precisa o general desde el punto de vista matemático; sin embargo , para nuestros propósitos debe servir con el fin de puntualizar la principal distinción entre regresiòn y correlación. Nótese que la **correlación** se refiere al hecho de que **dos variables se encuentran relacionadas** y a la estrechez de dicha relación. La **regresión** a su vez se refiere a la **naturaleza de la relación**.

Volvamos a considerar algunos adagios familiares y veamos como el concepto de regresión aflora en nuestro pensamiento diario.

“Un centavo ahorrado es un centavo ganado”

“Mas vale pájaro en mano que siento volando”

“Un espacio de tiempo ahorra nueve”

“Un cuadro vale mas que mil palabras”

Variable independiente (X)	Variable dependiente (Y)	Ecuación de regresion	Coeficiente de regresion
centavos ahorrados	centavos ganados	Y=X	1
Pajaros en mano	Pajaro volando	Y=2X	2
Un espacio de tiempo	Espacios ahorrados	Y=9X	9
Cuadros	Palabras	Y=1000X	1000

La ecuación de regresión denota una recta que es: Y = a + bX. El símbolo **a** recibe el nombre de **intercepto**, puesto que cuando X es igual a cero, Y = a; de donde la recta corta al eje de las X en **a** unidades a partir del origen. Cuando **a** es igual a cero, la recta pasa a través del origen, pues cuando X es igual a cero, Y es también igual a cero. El símbolo **b** se denomina pendiente, puesto que determina la inclinación de la recta.

Resulta fácil apreciar que **b** es la cantidad de cambio de **Y**, asociada a un cambio de **X**; y esta es exactamente, la forma en que hemos definido la regresión, por tanto, resulta lógico denominar a **b** como **coeficiente de regresión**.

Nótese que todos estos refranes implican la correlación de dos variables, pero van más allá y nos dicen en términos numéricos como están relacionadas ambas variables.

Ejemplo 16.1.

Un técnico veterinario decide hacer un seguimiento a un tipo de forraje para el engorde da las ovejas, para esto hace mediciones de crecimiento del cultivo de forraje, tuvo la idea de que podia juzgar lo bien que estaba creciendo el forraje a partir del volumen de la copa. Muy simple: copa grande - crecimiento bueno; copa pequeña - crecimiento deficiente. Pero no pudo precisar el grado de frondosidad o pujanza ni el de escasez o deficiencia. Lo que le hizo falta fue un analisis de regresion; este lo habria capacitado para expresar una relacion entre el crecimiento del forraje y el volumen de la copa por medio de una ecuación. Dado cierto volumen de copa, el podria usar la ecuación para predecir cual seria el crecimiento del cultivo de forraje. A continuación mostramos los datos obtenidos por el tecnico.

Volumen de la copa (X)	Crecimiento (Y)
22	0.36
6	0.09
93	0.67
62	0.44
84	0.72
14	0.24
52	0.33
69	0.61
99	0.64
98	0.65
41	0.47
85	0.60
90	0.51

Si la relación entre las dos variables fuese lineal, el modelo que predeciría el crecimiento seria:

Crecimiento = a + b (volumen)

El programa SAS para ajustar este modelo de regresión seria:

```
Data regre;
Input crecim volumen;
Cards;
22      0.36
 6      0.09
93      0.67
62      0.44
84      0.72
14      0.24
52      0.33
69      0.61
99      0.64
98      0.65
41      0.47
85      0.60
90      0.51
;
Proc glm;
Model crecim = volumen;
Run;
```

Se usa el comado GLM para hacer regresiones. Notese que es necesario escribir la palabra model y luego Y = X. No deben incluirse las constantes.

La salida de SAS incluye un análisis de varianza, que nos dice si la regresión es significativa. También el SAS calcula los estimados **a** (intercepto) y **b** (coeficiente de regresión). La salida producida por el SAS es el siguiente:

GENERAL LINEAR MODELS PROCEDURE

Depedent variable: crecim

F. variacion	**GL**	**Suma de cuadrados**	**Cuadrado medio**	**F value**	**Pr>F**
Model	1	0.34951099	0.34951099	48.93	0.0001 **(1)**
Error	11	0.0785693	0.00714236		
Total	12	0.42807692			

R-Square	**C.V**	**Root MSE**	**Rend mean**
0.81647 **(2)**	17.36	0.0845235	0.48692

Source	**DF**	**Type I SS**	**F value**	**Pr>F**
Volumen	1	0.34951099	48.93	0.0001
Source	DF	Type III SS	F value	Pr>F
Volumen	1	0.34951099	48.93	0.0001

Parameter	Estimate	T FOR HO: Parameter=0	PR>|T|	STD error of estimate
Intercept	0.16244 **(3)**	3.13	0.0097	0.05119753
Volumen	0.00518 **(4)**	7.00	0.0001 **(5)**	0.00073969

Lo que nos interesa de esta salida es lo siguiente:

Ver si la regresion es significativa, es decir si hay regresion entre crecimiento y volumne. El analisis de varianza en este caso prueba la hipotesis **Ho**: no hay regresion (b=0) contra **H1**: si hay regresion (b#0). Si observamos el valor debajo de Prob > F **(1)**, el valor obtenido fue 0.0001, lo que significa que la regresion es significativa a niveles menores de 0.0001, suficiente para rechazar **Ho**.

El **R cuadrado o R^2** (R Square), nos dice "que tanto de la variacion en **Y** se encontró asociada con **X**". En este caso el **R^2 (2)** fue de 0.81. Esto quiere decir que el modelo lineal explica el 81% de la variacion. En otras palabras el 81% de los datos están explicando el modelo determinado.

Los valores estimados para **"a"** (intercepto) y para **"b"** (coeficiente de regresion). En este caso **"a"** (3.13) es el valor que aparece a la par de intercept (0.00518) en la última linea. Entonces la ecuación para predecir el crecimiento seria:

Crecimiento = 0.16244 + 0.00518 (volumen)

El nivel de significancia para el cual rechazariamos las hipotesis que el parametro **b (4)** sea cero (es decir que no haya regresion). En este caso rechazamos la hipotesis con nivel 0.0001 (0.01%). Notese que la prueba **(5)** coincide con la **(1)** en este caso particular.

Si desea hacer un gráfico de los datos superponiendo la linea de regresion, simplemente añada las siguientes lineas al programa:

```
Output predicted = py;

Proc plot;

Plot crecim*volumen =´*´ py*volumen ´+´ / overlay;

Run;
```

En el gráfico los datos seran denotados con ´*´y la linea de regresion (datos predichos) con ´+´.

Ajuste de modelos cuadraticos y cubicos

Algunas veces la relacion entre dos variables no es lineal, pero si cuadrática o cúbica una relación típica es por ejemplo cuando la variable independiente **(x)** es un fertilizante aplicado con varias dósis. Conforme aumentamos la dósis, aumenta el rendimiento, pero llega al momento en que el rendimiento empieza a bajar debido a que dósis muy altas del fertilizante más bien produce efectos negativos (dañinos).

Para ajustar un modelo cuadrático (usando los datos del ajemplo anterior) simplemente escribimos las siguientes lineas:

```
Proc glm;

Model crecim = volumen volumen*volumen
```

Estas dos lineas de programa ajustan el siguiente modelo:

Crecimiento = a + b*volumen + C+volumen^2 + E

La salida de SAS es similar a la mostrada en el ejemplo anterior excepto que ahora tendriamos un parametro mas **(C)**, con su probabilidad de que sea significativo. Para saber si el modelo cuadrático se

ajusta mejor que el lineal, debemos fijarnos en la reducción producida en el error. Si la reducción es grande entonces quiere decir que el modelo cuadráqtico es mejor. También podemos ver si el parametro C es estadísticamente significativo, lo mismo que si el $\mathbf{R^2}$ aumento considerablemente.

El $\mathbf{R^2}$del modelo cuadrático siempre sera mas alto que el del modelo lineal (lo mismo aplica para el modelo cubico), pero debemos considerar y pensar si este aumento en el $\mathbf{R^2}$ realmente vale la pena como para complicar el modelo con un factor mas.

Si deseamos ajustar un modelo cúbico, es decir:

Crecimiento = a + b volumen + C volumen^2 + d volumen^3 + E

Simplemente escribimos en nuestro programa lo siguiente:

Proc glm;

Model crecim = volumen volumen*volumen volumen*volumen*volumen;

La salida incluira el parámetro adicional (d), asi como la probabilidad de que sea significativa. Las mimas observaciones observadas anteriromente aplican aquí para saber si el modelo cúbico es mejor que el cuadrático

Ejemplo 16.2.

Se implemento un experimento para evaluar siete raciones de Isaño como alternativa para reemplazar alimentos a base de maíz.. Los tratamientos fueron aplicados en diseño de bloques completos al azar con tres repeticiones y 30 pollos (unidades experimentales) por repetición. Las variables de respeuesta fueron Peso corporal (PC), Ganancia de peso (GP), Alimento Consumido (AC) y Conversión alimenticia (CA). Se obtuvo las ecuaciones de regresión cuadrática de peso corporal y ganancia de peso.

Análisis en SAS

```
Data Reg1;
Input racion etapa  rep gp ca acon pp;
Etapa1=etapa;
Etapa2=etapa**2;
Cards;
1 1 1 0.190 1.454 0.244 0.190
1 2 1 0.930 0.610 0.347 0.930
1 3 1 1.884 2.502 0.891 1.884
1 1 2 0.188 1.168 0.284 0.188
1 2 2 0.908 0.532 0.383 0.908
1 3 2 1.858 2.397 0.869 1.858
1 1 3 0.196 1.510 0.239 0.196
1 2 3 0.927 0.535 0.373 0.927
1 3 3 1.761 2.487 0.853 1.761
2 1 1 0.222 1.608 0.249 0.222
2 2 1 0.913 0.484 0.348 0.913
2 3 1 1.836 2.464 0.891 1.836
2 1 2 0.198 0.829 0.259 0.198
2 2 2 0.828 0.365 0.415 0.828
2 3 2 1.897 1.781 0.914 1.897
2 1 3 0.196 1.384 0.238 0.196
2 2 3 0.868 0.454 0.375 0.868
2 3 3 1.875 2.348 0.889 1.875
3 1 1 0.168 1.456 0.208 0.168
3 2 1 0.927 0.692 0.357 0.927
3 3 1 1.838 2.479 0.897 1.838
3 1 2 0.196 1.554 0.239 0.196
3 2 2 0.886 0.527 0.365 0.886
3 3 2 1.739 2.512 0.878 1.739
3 1 3 0.185 1.360 0.246 0.185
3 2 3 0.877 0.574 0.380 0.877
3 3 3 1.635 2.373 0.877 1.635
4 1 1 0.176 1.348 0.239 0.176
4 2 1 0.931 0.579 0.321 0.931
4 3 1 1.752 2.450 0.883 1.752
4 1 2 0.191 1.308 0.249 0.191
4 2 2 0.905 0.517 0.394 0.905
4 3 2 1.801 2.323 0.900 1.801
4 1 3 0.201 1.566 0.241 0.201
4 2 3 0.876 0.514 0.393 0.876
4 3 3 1.795 2.526 0.847 1.795
5 1 1 0.208 1.472 0.252 0.208
5 2 1 0.879 0.551 0.383 0.879
5 3 1 1.760 2.391 0.888 1.760
5 1 2 0.220 1.611 0.247 0.220
5 2 2 0.822 0.416 0.344 0.822
5 3 2 1.728 2.448 0.889 1.728
5 1 3 0.216 1.474 0.242 0.216
5 2 3 0.750 0.450 0.351 0.750
5 3 3 1.677 2.280 0.877 1.677
6 1 1 0.201 1.501 0.247 0.201
6 2 1 0.765 0.379 0.378 0.765
6 3 1 1.814 2.464 0.895 1.814
6 1 2 0.181 1.363 0.241 0.181
6 2 2 0.832 0.604 0.325 0.832
6 3 2 1.655 2.389 0.886 1.655
6 1 3 0.193 1.322 0.249 0.193
6 2 3 0.795 0.385 0.413 0.795
6 3 3 1.599 2.260 0.885 1.599
7 1 1 0.164 0.886 0.255 0.164
7 2 1 0.852 0.505 0.414 0.852
7 3 1 1.797 2.095 0.903 1.797
7 1 2 0.188 1.609 0.205 0.188
7 2 2 0.892 0.456 0.396 0.892
7 3 2 1.872 2.498 0.888 1.872
7 1 3 0.175 1.332 0.227 0.175
7 2 3 0.827 0.483 0.359 0.827
```

```
7    3    3    1.647    2.339    0.869    1.647
;
Proc reg;
model gp pp = etapa1 etapa2;
Run;
```

UNIDAD 17

ANÁLISIS DE COVARIANZA

Julio Gabriel Ortega

Raquel Vera Velázquez

Carlos Castro Piguave

Características

En el análisis de covariancia se combinan los conceptos del análisis de variancia para un diseño experimental y para regresión. El análisis de covariancia es utilizado en casos en los que la variable respuesta de un diseño experimental esté relacionada con una o más variables concomitantes. En este capítulo se tratará el caso de la covariancia lineal con una sola variable concomitante y se presentará el análisis para el Diseño de Bloques Completos al Azar. El estudiante sin embargo, no tendrá ningún problema en llevar esta técnica a un Diseño Completamente al Azar.

Covarianza simple

Suponganos que sobre la unidad experimental (i,j), donde i=1,2,.....,r, Y j=1,2,....t, (la parcela con el tratamiento j ennel bloque i), de un experiemento de bloques completos al azar, se observan, ademas de los valores de la caracteristica de interés, Y_{ij}, los valores de otra variable, X_{ij}, a la cual denominaremos variable compañera o covariable. Si el investigador sospecha que esta ultima ejerce alguna influencia sobre el valor, de la priemra es posible que el modelo lineal:

$$Y_{ij} = \mu + \beta_i + \tau_j + \gamma X_{ij} + e_{ij}, i = 1,.....,r, j = 1,.....,t,$$

$$E(e_{ij}) = 0, E(e^2{}_{ij}) = \sigma^2, E(e_{ij}e_{i'j'}) = 0$$

Yij= valor observado de la carateristica en estudio

Xij=Covariable

У=El coeficiente de varianza

Para explicar el valor de la respuesta observada sobre las unidades experiemntales, el investigador estara interesado en estimar contrastes entre efectos de tratamientos , en probar la significancia de los mismos y, además en en probar la significancia de la covariable. Así además de las probpiedades supestas de los errores, estos pueden considerarase normalmente distribuidos.

Ejemplo 17.1.

Un experimento de fertilizante con el diseño San cristobal (doce tratamientos en cuatro bloques completos al azar), realizado por el IMPA, en la zona de abastecimiento del ingenio Motzorongo, en el estado de Vera Cruz, cosechado en la plantilla en la zafra 1977-1978, produjpo los resulatados de la tabla, donde Y es el rendimiento de la caña, en toneladas por haectárea, y X es el número onservado de tallos molederos por parcela experiemntal. Se propone examinra el efecto de los nutrientes sobre el rendimiento de la caña, elimenado a través de la técnica de covarianza, el efecto del número de tallos molederos. Si el número de tallos molederos observados fuese independieentemente

del del tratamiento, la aplicación de la técnica de covarianza seria correcto, pero si ocurre que el número de tallos molederos por parcela es inducido por el tratamiento aplicado, entonces el uso del modelo de covarianza seria inapropiado para interpretar los resultados.

Como regla general para decidri sobre el empleo de covarianza, el investigador debiera tener la certeza de que sus covariables no estan influenciadas por lo tratamientos estudiados. El presente trabajo tal ves no sea muy claro para nuestros proópistos, pues podria pensarse que el fertilizante influye en cierta forma sobre el número de tallos moledrros, esperandose por consiguiente, que a dosis alta de nutrientes se observe un mayor número de dichos tallos.

Es común que en la práctica, para probara la significancia del efecto de los tratamientos sobre los valores de la propia covarable, se realice el análisisde varianza sobre los valores observados de la covarianza. Esta manera de proceder de acuerdo a Anderson Bancroft (1952), no es muy adecuada y recomiendan los autores que los investigadores basen su técnica del análisis en un juicio riguroso de su experiemento para bien detectar la existencia de la dependencia o no de las covariables para con los tratamientos.

Tratamientos	**I**		**II**		**III**		**IV**		**Sumas**	
	Y	**X**	**Y**	**X**	**Y**	**X**	**Y**	**X**	**Y**	**X**
1 0-0-0	107.5	319	103.6	308	85.5	319	115.6	275	412.2	1221
2 20-0-0	89.2	300	102.8	307	84.4	320	108.1	302	384.5	1229
3 0-20-0	102.2	280	110.0	280	76.9	299	87.5	268	376.6	1127
4 20-20-20	88.1	318	105.0	315	104.7	319	120.3	311	418.1	1263
5 0-0-20	121.4	308	100.3	304	111.7	315	126.1	290	459.5	1217
6 20-0-20	119.4	306	111.1	310	100.8	334	119.2	296	450.5	1246
7 0-20-20	110.6	316	113.6	303	114.7	284	122.2	295	461.1	1198
8 20-20-20	106.4	290	120.0	306	88.9	314	130.0	299	445.3	1209
9 10-10-10	114.7	315	106.9	299	114.4	310	115.8	297	451.8	1221
10 30-10-10	116.4	330	129.2	315	106.4	319	136.9	317	488.9	1281
11 10-30-10	96.1	302	107.8	353	106.6	310	122.8	294	433.3	1259
12 10-10-30	102.5	321	114.4	307	116.4	316	126.7	302	460.0	1246
Sumas	1274.5	3705	1324.7	3707	1211.4	3759	1431.2	3546	52418.8	14717

Procedimiento con nuestro ejemplo, se realizan los cálculos para contruir la tabla de suma de cuadrados y productos cruzados.

FV	**gl**	**SC de Y productos cruzados**		
		X.X	**X.Y**	**Y.Y**
Bloques	3	2129.1	-2043.29	2157.25
Tratamientos	11	4323.7	1904.43	3042.45
Error	33	4574.7	-404.26	2780.86
Total	47	11.27.5	-543.12	7980.56
E'=T+E	44	8898.4	1500.17	5823.31

Covariancia en un Diseño de Bloques Completos al Azar

Modelo Aditivo Lineal

El modelo aditivo lineal para un análisis de covariancia en un Diseño de Bloques Completos al Azar es el siguiente:

$Yij = \mu + ti + \gamma j + ß\ (Xij - X\bullet\bullet) + eij$

donde:

Yij= es el valor o rendimiento observado en el i-ésimo tratamiento, j-ésimo bloque.
μ = es el efecto de la media general.
ti =es el efecto del i-ésimo tratamiento.
γj =es el efecto del j-ésimo bloque.
ß = es el coeficiente de regresión lineal de Y sobre X.
Xij= es el valor de la variable independiente en el i-ésimo tratamiento, j-ésimo bloque.
X•• es la media de la variable independiente.
eij= es el efecto del error experimental en el i-ésimo tratamiento, j-ésimo bloque.
t es el número de tratamientos.
b es el número de bloques.

Analisis en Sas

```
Data ancova1;
Input Rep Trat  X  Y ;
Cards;
1	1	319	107.5
1	2	300	89.2
1	3	280	102.2
1	4	318	88.1
1	5	308	121.4
1	6	306	119.4
1	7	316	110.6
1	8	290	106.4
1	9	315	114.7
1	10	330	116.4
1	11	302	96.1
1	12	321	102.5
2	1	308	103.6
2	2	307	102.8
2	3	280	110.0
2	4	315	105.0
2	5	304	100.3
2	6	310	111.1
2	7	303	113.6
2	8	306	120.0
2	9	299	106.9
2	10	315	129.2
2	11	353	107.8
2	12	307	114.4
3	1	319	85.5
3	2	320	84.4
3	3	299	76.9
3	4	319	104.7
3	5	315	111.7
3	6	334	100.8
3	7	284	114.7
3	8	314	88.9
3	9	310	114.4
3	10	319	106.4
3	11	310	106.6
3	12	316	116.4
4	1	275	115.6
```

```
4       2       302     108.1
4       3       268     87.5
4       4       311     120.3
4       5       290     126.1
4       6       296     119.2
4       7       295     122.2
4       8       299     130.0
4       9       297     115.8
4       10      317     136.9
4       11      294     122.8
4       12      302     126.7
Proc glm;
Classes trat;
Model Y=X trat;
Lsmeans trat/pdiff stderr;
Means trat;
Run;
```

Ejemplo 17.2.

Se desarrolló un experimento cuyo objetivo era determinar si la exposición en agua calentada artificialmente afectaba el crecimiento de las ostras. Cinco bolsas con diez ostras cada una fueron aleatoriamente asignadas a cinco temperaturas (T1, T2, T3, T4, T5); cada bolsa constituía una unidad experimental. Se utilizaron cinco estanques, cada uno calentado a una de las cinco temperaturas. Las ostras fueron limpiadas y pesadas al comienzo y al final del experimento un mes después. El experimento se repitió cuatro veces para lo cual fueron necesarios 4 meses. Cada repetición constituye un bloque. Los pesos iniciales y finales se presentan en la siguiente tabla:

Bloq.	**T1**		**T2**		**T3**		**T4**		**T5**		**Total**	
	X	**Y**	**X**	**Y**	**X**	**Y**	**X**	**Y**	**X**	**Y**	**X**	**Y**
I	20.4	24.6	27.2	32.6	26.8	31.7	22.4	29.1	21.8	27.0	118.6	145.0
II	19.6	23.4	32.0	36.6	26.5	30.7	23.2	28.9	24.3	30.5	125.6	150.1
III	25.1	30.3	33.0	37.7	26.8	30.4	28.6	35.2	30.3	36.4	143.8	170.0
IV	18.1	21.8	26.8	31.0	28.6	33.8	24.4	30.2	29.3	35.0	127.2	151.8
Total	83.2	100.1	119.0	137.9	108.7	126.6	98.6	123.4	105.7	128.9	515.2	616.9

Analisis en Sas

```
Data ancova2;
Input Rep  trat X  Y ;
Cards;
1       1       20.4    24.6
1       2       27.2    32.6
1       3       26.8    31.7
1       4       22.4    29.1
1       5       21.8    27.0
2       1       19.6    23.4
2       2       32.0    36.6
2       3       26.5    30.7
2       4       23.2    28.9
2       5       24.3    30.5
3       1       25.1    30.3
3       2       33.0    37.7
```

```
3    3    26.8    30.4
3    4    28.6    35.2
3    5    30.3    36.4
4    1    18.1    21.8
4    2    26.8    31.0
4    3    28.6    33.8
4    4    24.4    30.2
4    5    29.3    35.0
Proa glm;
Classes trat;
Model Y=X trat;
Lsmeans trat/pdiff stderr;
Means trat;
Run;
```

Ejemplo 17.3.

```
Data ancova;
Input cerdo 1-2 trat$ 4 X 6 Y 8-9;
Cards;
1    A    4    10
2    B    1    5
3    C    2    5
4    D    3    5
5    A    2    5
6    B    3    9
7    C    6    12
8    D    1    7
9    A    4    8
10   B    2    5
11   C    4    10
12   D    4    6
13   A    6    15
14   B    1    1
15   C    6    10
16   D    2    9
17   A    4    12
18   B    3    5
19   C    2    8
20   D    1    3
Proa glm;
Classes trat;
Model Y=X trat;
Lsmeans trat/pdiff stderr;
Means trat;
Run;
```

PARTE VI

TECNICAS DE ANALISIS MULTIVARIANTE

UNIDAD 18

ANÁLISIS MULTIVARIANTE

Julio Gabriel Ortega

Alfredo Valverde Lucio

Introducción

La moderna investigación que se realiza demanda del uso de técnicas actualizadas de análisis estadísticos, que contribuyan y ayuden al investigador a llegar a interpretaciones cercanas a la realidad y en la cual se involucran muchas variables; esto no es posible hacer con las técnicas clásicas del diseño experimental, de ahí que muchos investigadores innovaron técnicas valiosas para estos análisis. Nosotros quisimos compartirles en esta obra los más útiles en base a la experiencia de nuestro primer autor, a lo largo de los 30 años de investigación. Hemos considerado diversos y valiosos documentos, pero también experiencias como de las Doctoras Carmen García y Sagrario Gómez, docentes de la Catedra de "Técnicas estadísticas multivalentes aplicadas", de la Universidad Pública de Navarra de Pamplona, España. Estas técnicas estadísticas son aplicadas a datos multidimensionales. Se pretende conseguir que estudiantes con distinto bagaje académico (ingenieros, biólogos, sociólogos, matemáticos, etc.) puedan realizar sus propios análisis empíricos con rigor.

Franco e Hidalgo (2003), mencionan que el origen del análisis multivariado se remonta a los comienzos del siglo XX, con Pearson y Sperman, época en la cual se empezaron a introducir los conceptos de la estadística moderna. Las bases definitivas de este tipo de análisis se establecieron en la década 1930-40 con Hotelling, Wilks, Fisher, Mahalanobis, y Bartlett (Bramardi, 2002). En términos generales, el análisis multivariante se refiere a todos aquellos métodos estadísticos que analizan simultáneamente medidas múltiples (más de dos variables) de cada individuo. En sentido estricto, son una extensión de los análisis univariados (análisis de distribución) y bivariados (clasificaciones cruzadas, correlación, análisis de varianza y regresiones simples) que se consideran como tal si todas las variables son aleatorias y están interrelacionadas (Hair *et al.,* 1992).

Los métodos multivariantes Franco e Hidalgo (2003), los clasifican en dos grandes grupos (Tabla 18.1): 1) los de dependencia, que son aquellos en los cuales una variable o conjunto de variables es identificado como dependiente de otro conjunto conocidas como independiente o predictor; y 2) los de interdependencia, o aquellos en que ninguna variable o grupo de variables es definido como independiente o dependiente y, más bien, el procedimiento implica el análisis simultáneo de todo el conjunto de variables (Hair *et al*., 1992).

Tabla 18.1. Métodos estadísticos de análisis multivariante.

Métodos de dependencia (tipo de análisis)	**Método de independencia (tipo de análisis)**
Discriminante múltiple	Componentes principales
Correlación canónica	Factorial
Regresión múltiple	Conglomerados (clusters)
Multivariante de la varianza	Multidimensional
Conjunto	Correspondencia

Fuente; Hair et al. (1992), Franco e Hidalgo (2003)

En esta obra nos ocuparemos básicamente de los métodos de interdependencia, o aquellos en que ninguna variable o grupo de variables es definido como independiente o dependiente y, más bien, el procedimiento implica el análisis simultáneo de todo el conjunto de variables (Franco e Hidalgo 2003). Haremos el esfuerzo de explicar de manera concisa y sencilla las técnicas más usadas y cómo analizarlas aplicando los comandos del software estadísticos SPSS.

18.1. ANÁLISIS FACTORIAL

18.1.1. Análisis de componentes principales (ACP)

Este método se basa en la transformación de un conjunto de variables cuantitativas originales en otro conjunto de variables independientes no correlacionadas, llamadas componentes principales. Los componentes deben ser interpretados independientemente unos de otros, ya que contienen una parte de la varianza que no está expresada en otro componente principal (Uriel y Aldas 2005).

En pocas palabras se trata de disminuir la dimensión de los problemas con muy poca pérdida de información.

Obtención de los componentes principales

1 - **En definitiva:** la obtención de componentes principales consiste en pasar de la matriz $\sum_x$ simétrica a la matriz $\sum_p$ diagonal cuyos elementos van en orden decreciente. El paso se hace mediante

$$\sum_p = A \sum_x A'$$

con A matriz cuyas filas son los vectores propios (ortogonales) de $\sum_x$ correspondientes a los valores propios ordenados de mayor a menor. Además A es ortogonal $(A' = A^{-1})$

2 - **¿Mantenemos la varianza? Sí.**

$\sum_1^p Var(X_i) = \sum_1^p Var(P_i)$

$\sum_1^p Var(Pi) = tr(\Sigma_p) = tr(A\Sigma_x A') = tr(A'A\Sigma_x) = tr(\Sigma_x) = \Sigma_1^p Var X_i$

(pues $AA' = I$ por ser A ortogonal)

¿Cuál es la varianza total del modelo?

$\sum Var X_i = \sum Var Pi = \sum \lambda_i$ con λ_i valores propios de Σ_x

¿Cuánta varianza retiene el primer componente?

Como $Var(P_1) = \lambda_1$; la proporción de varianza retenida por P_1 es:

$$\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_p}$$

¿Cuánta varianza retienen los dos primeros componentes?

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \cdots + \lambda_p}$$

3 - Si algún $\lambda_i = 0$ entonces las siguientes son 0 pues Σ_x es semidef. positiva. Si $\lambda_i = 0 \rightarrow VarPi = 0 \rightarrow Pi$ es constante.

X_1	X_2	$\cdots$	X_p	P_1	P_2	P_i	$\cdots$	P_p
$\vdots$	$\vdots$	$\vdots$	$\vdots$			C		
$\vdots$	$\vdots$	$\vdots$	$\vdots$			C		
$\vdots$	$\vdots$	$\vdots$	$\vdots$			C		
$\vdots$	$\vdots$	$\vdots$	$\vdots$			C		

$a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p = C \quad \forall$ todos los individuos.

Yo podría despejar cualquier variable X en función de las demás. $\Rightarrow$ una columna c.l. exacta de otras (en los datos originales)$\Rightarrow$ sobra una variable.

4 - Los componentes no son invariantes frente a cambios de escala. El cambio en la unidad de medida produce cambios en los resultados. Esto se evita con variables tipificadas.

En lugar de diagonalizar Σ_x, diagonalizamos Σ_z que es la matriz de covarianzas de las X tipificadas (o la matriz de correlaciones de la X originales).

$$Z_i = \frac{X_i - \bar{X}_i}{\sigma_i}$$

Hacemos todo exactamente igual pero sobre Σ_z en vez de sobre Σ_x con la ventaja de que

$tr(\Sigma_z) = p = Var(Z_1) + Var(Z_2) + \cdots + Var(Z_p)$

$\lambda_1/p \rightarrow$ proporción de información que retiene P_1

$\lambda_2/p \rightarrow$ proporción de información que retiene P_2

$\frac{\lambda_1+\lambda_2}{p} \Rightarrow$ proporción de información que retienen conjuntamente los ejes P_1 y P_2

5 - Si se ha decidido utilizar sólo k componentes ¿Cómo quedan representadas las variables X_i(ó Z_i si hemos tipificado)?

Si nos quedamos con algunas componentes perdemos cierta información, pero no la perdemos por igual en todas las variables.

$\mathbf{P} = \mathbf{AX}$ Cómo A es ortogonal: $A' = A^{-1}$

$\mathbf{X} = \mathbf{A}^{-1}\mathbf{P} = \mathbf{A}'\mathbf{P}$

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{j1} & a_{p1} \\ a_{12} & a_{22} & \cdots & a_{j2} & a_{p2} \\ a_{1i} & a_{2i} & \cdots & a_{ji} & a_{pi} \\ a_{1p} & a_{2p} & \cdots & a_{jp} & a_{pp} \end{pmatrix} \cdot \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{pmatrix}$$

$X_1 = a_{11}P_1 + a_{21}P_2 + \cdots + a_{p1}P_p$

$\vdots$

$X_i = a_{1i}P_1 + a_{2i}P_2 + \cdots + a_{pi}P_p$

$\vdots$

$X_p = a_{1p}P_1 + a_{2p}P_2 + \cdots + a_{pp}P_p$

$\Rightarrow VarX_1 = a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \cdots + a_{p1}^2\lambda_p$

$\Rightarrow VarX_i = a_{1i}^2\lambda_1 + a_{2i}^2\lambda_2 + \cdots + a_{pi}^2\lambda_p$

Si yo decido emplear sólo los 2 primeros componentes, del total de la varianza de X_1 : $a_{11}^2\lambda_i + a_{21}^2\lambda_2 + \cdots + a_{p1}^2\lambda_p$ me quedo con $a_{11}^2\lambda_1 + a_{21}^2\lambda_2$;

Comunalidad

$$\frac{a_{11}^2\lambda_1 + a_{21}^2\lambda_2}{a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \cdots + a_{p1}^2\lambda_p}$$

Si he decidido emplear k componentes, la comunalidad de X_1 será:

$$\frac{a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \cdots + a_{k1}^2\lambda_k}{a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \cdots + a_{p1}^2\lambda_p}$$

Así podemos conocer la comunalidad de cada variable X_i al quedarnos con $1, 2, \cdots, k$ factores.

Si hemos trabajado con datos tipificados.

$\mathbf{P} = A\mathbf{Z}$;

$\mathbf{Z} = A'\mathbf{P}$;

$Z_1 = a_{11}P_1 + a_{21}P_2 + \cdots + a_{p1}P_p$;

$Var(Z_1) = 1 = a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \cdots + a_{p1}^2\lambda_p$

Varianza retenida por las k primeras componentes $P_1, \cdots, P_k$

$a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \cdots + a_{k1}^2\lambda_k$

Comunalidad.

para$Z_1 : a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \cdots + a_{k1}^2\lambda_k$

para$Z_i : a_{1i}^2\lambda_1 + a_{2i}^2\lambda_2 + \cdots + a_{ki}^2\lambda_k$

¿Qué parte de la información de X_1 queda recogido en P_1?

$$\frac{a_{11}^2\lambda_1}{a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \cdots + a_{p1}^2\lambda_p}$$

(Si hemos trabajado con variables tipificadas el denominador es 1)

6 -¿Cuál es la correlación entre dos variables originales (tipificadas)?.

Corr Zi y Zj.

Corr(Zi,Zj)=Cov(Zi,Zj)=$E(a_{1i}P_1 + a_{2i}P_2 + \cdots + a_{pi}P_p)(a_{1j}P_1 + a_{2j}P_2 + \cdots + a_{pj}P_p) =$
$a_{1i}a_{1j}Var(P_1) + a_{2i}a_{2j}Var(P_2) + \cdots + a_{pi}a_{pj}Var(P_p) = a_{1i}a_{1j}\lambda_1 + a_{2i}a_{2j}\lambda_2 + \cdots + a_{pi}a_{pj}\lambda_p$

Si sólo consideramos k variables cortamos esa cadena de sumandos y tenemos la **correlación reproducida.**

P.ej. si sólo tomamos 2 factores P_1 y P_2.

Corr reproducida$(Z_1, Z_2) = a_{1i}a_{1j}\lambda_1 + a_{2i}a_{2j}\lambda_2$

7 - Correlación entre variables originales y factores

$$Corr(Zi, Pj) = \frac{Cov(Zi, Pj)}{\sqrt{Var(P_j)}\sqrt{VarZ_i}} = \frac{Cov(Z_i, P_j)}{\sqrt{\lambda_j}}$$

$$Cov(Z_i, P_j) = E(Z_i.P_j) = E[(a_{1i}P_1 + a_{2i}P_2 + \cdots + a_{pi}P_p)P_j] = a_{ji}Var(P_j) = a_{ji}\lambda_j$$

$$Corr(Zi, Pj) = \frac{a_{ji}\lambda_j}{\sqrt{\lambda_j}} = a_{ji}\sqrt{\lambda_j}$$

En particular:

$$Corr(Zi, P_1) = a_{1i}\sqrt{\lambda_1}$$

$$Corr(Zi, P_2) = a_{2i}\sqrt{\lambda_2}$$

	P_1	P_2
Z_1	$a_{11}\sqrt{\lambda_1}$	$a_{21}\sqrt{\lambda_2}$
Z_2	$a_{12}\sqrt{\lambda_1}$	$a_{22}\sqrt{\lambda_2}$
Z_i	$a_{1i}\sqrt{\lambda_1}$	$a_{2i}\sqrt{\lambda_2}$
Z_p	$a_{1p}\sqrt{\lambda_1}$	$a_{2p}\sqrt{\lambda_2}$

TABLA que en SPSS se llama matriz de componentes.

Conociendo el contenido de esta tabla se puede obtener todo lo demás.

La suma por columnas de los elementos al cuadrado proporciona las λ_i.

La suma por filas de los elementos al cuadrado proporciona las comunalidades.

También se pueden obtener las correlaciones reproducidas.

8 - Gráfico de variables. Representación de las variables originales en los ejes formados por los componentes.

Normalmente representamos en P_1 y P_2 ó P_1 y P_3 ó P_2 y P_3.

$d = \sqrt{(a_{1i}\sqrt{\lambda_1})^2 + (a_{2i}\sqrt{\lambda_2})^2} = \sqrt{comunalidad(Z_1)}$

Se representan las variables de acuerdo a las coordenadas obtenidas en la tabla anterior. (matriz de componentes).

Cuanto más alejada está Z_i del origen, mayor comunalidad tiene, por lo tanto está mejor representada. Las variables cercanas al origen están mal representadas. Se necesita otro eje para representarlas bien (y por tanto otros gráficos).

La representación está limitada por el círculo unidad.

9 - Gráfico para observaciones.

Z_i tipificada.

Los factores P_j son tales que:

$E(P_j) = 0$

$Var(P_j) = \lambda_j$

Ahora hacemos componentes tipificadas.

$$Z_{p_j} = \frac{P_j - 0}{\sqrt{\lambda_j}}$$

$$Z_{p_j} = \frac{1}{\sqrt{\lambda_j}}(a_{j1}Z_1 + a_{j2}Z_2 + \cdots + a_{jp}Z_p)$$

$$Z_{p_j} = \frac{a_{j1}}{\sqrt{\lambda_j}}Z_1 + \frac{a_{j2}}{\sqrt{\lambda_j}}Z_2 + \cdots + \frac{a_{jp}}{\sqrt{\lambda_j}}Z_p$$

Coeficientes para puntuaciones factoriales.

En el individuo (r).

$$Z_{p_j}^{(r)} = \frac{a_{j1}}{\sqrt{\lambda_j}}Z_1^{(r)} + \frac{a_{j2}}{\sqrt{\lambda_j}}Z_2^{(r)} + \cdots + \frac{a_{jp}}{\sqrt{\lambda_j}}Z_p^{(r)}$$

Puntuación tipificada (o coordenada) del individuo r en el eje P_j = **Puntuaciones factoriales**.

10 - Criterios para decidir el número de ejes (componentes) a retener.

- Podemos fijar un valor de varianza que queremos retener.
- Podemos fijar un valor mínimo para las comunalidades, o al menos para la comunalidad de alguna variable de especial interés.
- Podemos observar los valores propios (gráfico de sedimentación), ver cómo van disminuyendo y decidir con cuántos nos quedamos.
- Como las variables tipificadas tienen varianza 1, podemos seleccionar aquellas componentes que tengan varianza mayor que 1 (o sea que expliquen más que las variables originales). El criterio en este caso es seleccionar los componentes asociados a valores propios >1.

11 - Test de Barlett.

¿Tiene sentido aplicar CP?

¿Hay alguna estructura de correlación suficiente?

¿Hay suficiente dependencia entre las variables?

Hipótesis nula: No existe dependencia entre las Z_i.

$$(\Sigma_z \approx \begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{pmatrix} \Rightarrow |\Sigma_z| = 1)$$

Si se acepta la Hipótesis nula no tiene sentido seguir adelante.

Estadístico de contraste:

$-|n - 1 - 1/6(2_p + 5)| \ln |R| \approx \chi^2_{(1/2).p(p-1)}$

Estad. calculado $< \chi^2$ tabla $\rightarrow$ Acepto$\Rightarrow$ No aplicar CP. (sig>0.05)

Estad. calculado $> \chi^2$ tabla $\rightarrow$ Rechazo$\Rightarrow$ Sí aplicar CP. (sig<0.05)

12 - Rotación de los ejes factoriales

Supongamos que nos hemos quedado con k ejes, en ocasiones una vez que se ha realizado el análisis y se ha cumplido el principal objetivo (reducir) es posible que se observe que una rotación de los ejes podría proporcionar una mejor interpretación de los resultados.

Existen varios tipos de rotación:

Rotación Varimax:

→Mantiene las comunalidades.

→Mantiene el Total de Varianza explicada.

→Cambia el reparto entre los ejes.

Con la rotación es puede ser más fácil poner nombre a los ejes.

13 - Patrones Con frecuencia se observan como componentes principales:

P_1: Tamaño

P_2: Forma.

- Todas las variables correlacionando positivamente con P_1

  "Cuanto más de cualquier variable, mayor P_1"

- Algunas variables más específicas correlacionadas(positiva o negativamente) con P_2.

Ordenes en SPSS para Análisis de Componentes Principales (ACP)

Las clasificaremos en tres tipos:

1. *Las específicas del análisis*
2. *Las de obtención de puntuaciones factoriales y representación gráfica de las observaciones*
3. *Las de cambios (mejoras) en los gráficos*

1. Analizar ↴ Reducción de datos ↴ Análisis factorial ↴

Introducir las variables propias del análisis

En **DESCRIPTIVOS**: solución inicial, coeficientes, niveles de sig, determinante, **KMO** y prueba de esfericidad, reproducida.

En **EXTRACCIÓN**

- Método Componentes principales
- Analizar matriz de correlaciones

Mostrar: solución factorial sin rotar (solo en el caso SIN rotación) y gráfico de sedimentación

Extraer : -si se desea obtener los ejes con valor propio mayor que 1: autovalor mayor que 1 (la opción por defecto)

- si se desea un nº concreto de ejes: opción Nº de factores igual a

En **ROTACIÓN**: -ninguno (sin rotación)

- VARIMAX (cuando se desea rotar)

Mostrar gráficos de saturaciones (siempre, se trata del gráfico de variables, construido con las correlaciones de la matriz de componentes)

2. En **PUNTUACIONES**:

- Guardar como variables (nuevas columnas en e archivo de datos FACij)
- Método regresión
- En OPCIONES: ordenar coeficientes por tamaño

Gráfico de observaciones: En el menú Gráfico

Dispersión Simple Definir

- Al eje X el factor 1, al eje Y el factor 2 (u otros)
- Etiquetar por la variable que identifica las observaciones (si la hay)
- Opciones: Mostrar etiquetas.

Si hay muchas observaciones este gráfico sale confuso. En ocasiones se puede segmentar el fichero antes de hacer el gráfico y entonces aparecen varios gráficos más claros.

3. **Modificar un gráfico**: Doble click en el gráfico (aparece una nueva ventana)

Cambiar de dimensión 3 a dimensión 2: edición propiedades variables (y se elimina una de las 3)

Colocar ejes en el gráfico: opciones linea de referencia etc.

Ejemplo 18.1.

Este ejemplo corresponde a un trabajo realizado por Gabriel *et al.* (2013) y López *et al*. (2015), en el que estos investigadores evaluaron 30 híbridos de tomate **[***Solanum lycopersicumL. (Mill.)]* proporcionados por la Fundación PROINPA en Bolivia. Estos hìbridos fueron obtenidos mediante el cruzamiento entre 21 líneas parcialmente puras obtenidas de cultivares comerciales de polinización abierta de tres ciclos de autofecundación (progenitores femeninos) y dos líneas silvestres obtenidas por el Dr. Mikel Stevens de San Diego State University-California, Estados Unidos: 70 con resistencia al tospovirus TSWV o peste negra y 71susceptible a tospovirus (progenitores masculinos). Se evaluaron variables cuantitativas y cualitativas (Tabla 18.2), en base a los descriptores del IPGRI (1996). Se realizaron tres tipos de análisis: i) análisis de correlación, ii) análisis de componentes principales (ACP) y iii) análisis de correspondencia múltiple (ACM).

El **análisis de correlación** permitió conocer el grado de asociación existente entre las diferentes variables sin afectar la relación presente entre unidades (cm, mm, g y días). Para este análisis se utilizó el coeficiente de **Pearson** aplicado para datos multiestratos cuantitativos, y para los datos cualitativos se utilizó el coeficiente de correlación de **Spearman.**

El **ACP**, consistió en transformar un conjunto de variables cuantitativas originales a un nuevo conjunto de variables independientes y no correlacionadas donde los primeros componentes llevan la mayor información o variabilidad. Los componentes deben ser independientes unos de otros, ya que contienen una parte de la varianza que no está expresada en otros componentes principales y el número de componentes depende del número de variables incorporadas en el análisis y la contribución de estas a cada componente principal se expresa en valores y vectores propios. El valor propio representa la varianza asociada con el componente principal y decrece a medida que se generan estos componentes. En cambio, el vector propio contiene los coeficientes de las combinaciones lineales de las p variables originales.

El **ACM,** permitió identificar las variables cualitativas más importantes que expliquen la mayor parte de la variabilidad genética que presenta la muestra original. Este método permitió describir la relación existente entre las variables cualitativas heterogéneas conforme a los diferentes estados que presentaron los híbridos evaluados.

Tabla 18.2. Detalle de variables cualitativas y cuantitativas evaluadas en 30 híbridos de tomate (Fuente: López 2014).

	Variables cuantitativas	Código	Variables cualitativas	Código
Variables morfológicas	1. Tamaño del fruto	TmF	1. Tipo de crecimiento de la planta	TCP
	2. Longitud del fruto (mm)	LnF	2. Forma predominante del fruto	FPF
	3. Ancho del fruto (mm)	AnF	3. Color exterior del fruto maduro	CFM
	4. Rendimiento por planta (Kg)	Y	4. Color de la carne del pericarpio (interior)	CCP
			5. Forma del corte transversal del fruto	CTF
			6. Numero de lóculos	NOL
Variables agronómicas	1. Número de días a la madurez	NoDM		
Variables agroindustriales	1. Perdida en PH	PpH		
	2. Perdida en Textura	PTxT		
	3. Ganancia en Grados Brix	G°Bx		

La descripción de los resultados lo haremos por subíndice, con el propósito de ser ordenados y que el lector pueda seguir una secuencia lógica.

La Tabla 18.3. muestra una matriz con las correlaciones de Pearson para cinco variables cuantitativas.

Tabla 18.3. Coeficientes de correlación de Pearson entre cinco variables cuantitativas evaluadas en 30 híbridos de tomate (Fuente: Lòpez 2014).

	NoDM	**TmF**	**LNF**	**AnF**	**Y**
NoDM	**1**				
TmF	0.26*	**1**			
LnF	0.33*	**0.56****	**1**		
AnF	0.39*	**0.55****	**0.67****	**1**	
Y	0.15	**0.35****	0.33*	0.27*	**1**

NoM = Nùmero de días hasta la madurez, TmF = Tamaño del fruto, LnF = Longitud del fruto, AnF = Ancho del fruto, Y = Rendimiento de fruto/planta. **: Altamente significativo al $p<0.01$ de probabilidad, *: Significativo al $p<0.05$ de probabilidad.

En la tabla 18.3 se observa una correlación positiva altamente significativa entre las variables LnF con AnF ($r = 0.67$), esto indicaria que a mayor longitud de fruto, más ancho es el fruto. Por otra parte, la correlaciòn del TmF fue positivo y altamente significativo con LnF ($r = 0.56$) y AnF ($r=0.55$). Por lo tanto, los híbridos con mayor TmF exhibieron mayor AnF y LnF. Hubo correlación positiva baja significativa entre las variables TmF y Y ($r = 0.35$).

El ACP generó componentes, que respresentaron el número de componentes y su valor propio en su abscisa y el porcentaje de la varianza en la ordenada. Esto mostró el decrecimiento de los primeros componentes en relación a los demás y fueron seleccionados aquellos componentes más significativos.

Se consideró como componentes significativos aquellos valores anteriores al punto de inflexión (Figura 18.1.). Se retuvo dos componente cuyo valor propio fue ≥1 y que expresaron más del 69.88% de la varianza total. Aunque el

segundocomponente tiene un valor de 0.9 es próximo a uno; por tanto,se le tomó en cuenta como componente con valor de uno.

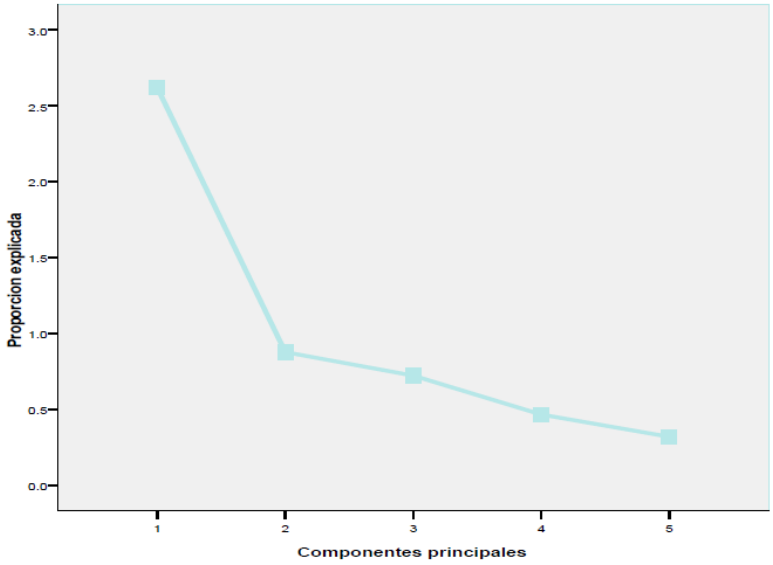


Figura 18.1. Gràfico de sedimentación para cinco variables cuantitativas agromorfologicas evaluadas en 30 híbridos de tomate (Lòpez 2014, López *et al*. 2015).

La Tabla 18.4. expresa la contribución de cada una de las variables que se asociaron al componente principal y la varianza total del componente. Mientras más altos fueron los coeficientes de la varianza sin importar el signo, más eficaces serán los componentes en la discriminación de los híbridos.

Tabla 18.4. Matriz factorial correspondiente a variables relacionadas de los híbridos de tomate evaluados en invernadero (López 2014)**.**

Componenetes	**1^0**	**2^0**
Valor propio	2.62	-0.63
Porcentaje de varianza	52.35	17.53
Variables	**Coeficientes de correlación**	
NoDM	0.55	-0.63
TmF	0,78	0.16
LnF	0.84	-0.02
AnF	0.83	-0.16
Y	0.53	0.66

NoM = Nùmero de días hasta la madurez, TmF = Tamaño del fruto, LnF = Longitud del fruto, AnF = Ancho del fruto, Y = Rendimiento de fruto/planta.

El primer componente contribuyó con el 52.35% del total de la varianza (Tabla 16.2). Aportando en forma positiva las variables TmF, LnF y AnF. Esto indicaria que este componente identifico a los híbridos con TmF, LnF y AnF. El segundo componente contribuyó con el 17.53% de la varianza total, en el cual la variable Y aporto positivamente. Por otra parte, la variable NoDM, contribuyó en forma negativa. Este componente distinguió a los híbridos con mayor rendimiento y precocidad.

18.1.2. Análisis Factorial de Correspondencias (AFC)

- Técnica de Interdependencia.
- Procedimiento factorial para variables cualitativas.
- Se basa en el análisis de las asociaciones entre las modalidades de dos o más variables cualitativas. Cuando el análisis se realiza para dos variables nos referimos a **Correspondencias simples** y si son más variables, **Correspondencias múltiples**.

Objetivo

La factorización (búsqueda o construcción de nuevos factores, nuevas variables a partir de iniciales).

Consiste en reducir un espacio de grandes dimensiones (número de modalidades o categoría) a otro de menor dimensión (generalmente dos) en el que se ubiquen o proyecten simultáneamente las modalidades de las variables. Además, podemos observar:

- Semejanzas entre modalidades de una misma variable.
- Asociaciones (y repulsiones) entre modalidades de distintas variables.
- Modalidades "raras"

Presentación de datos

Correspondencias simples

(Dos variables, tabla contingencia)

	a_1	a_2	...	a_p
b_1				
b_2				
		Frecuencias n_{ij}		
b_n				

Correspondencias múltiples

(Varias variables, tabla disyuntiva completa)

Observación	$a_1, a_2 ... a_p$	$b_1, b_2 ... b_r$	$c_1, c_2 ... c_s$
1	$0,1,0,0,0$	$1,0,\ldots,0$	$0,0,\ldots,1$
2	$1,0,0,0,0$	$0,1,\ldots,0$	$1,0,\ldots,0$
3			
$\vdots$			

Extracción de factores

En ACP el criterio para extraer factores era que se mantuviera la mayor parte de la varianza.

En AFC el criterio para extraer factores es captar la mayor parte de ASOCIACION entre modalidades. A esa asociación la llamamos inercia, es decir, extraemos los factores de modo que se produzca lo más fielmente posible las asociaciones entre modalidades.

Los datos se transforman de tal forma que las modalidades de una variable juegan el papel de variables y las de la otra de observaciones (y recíprocamente).

Obtendremos:

- Cuanta inercia retiene cada eje. (en C.P. cuanta varianza explica cada factor)
- Contribución absoluta de las modalidades a la formación de cada eje.
- Contribución relativa con que cada eje representa a cada modalidad, es decir, cómo queda representada cada modalidad (en ACP comunalidad).
- Modalidades semejantes de la misma variable.
- Asociación (+ o -) de modalidades de distintas variables.
- Se puede proyectar en el mismo grafico variables suplementarias que no intervienen en el proceso pero que enriquecen la presentación.

Correspondencias simples

2 variables A variable cualitativa con p modalidades (a1, a2, ..., aj, ..., ap)

B variable cualitativa con n modalidades (b1, b2, ..., bi, ..., bn)

N observaciones nij = número de observaciones en las modalidades aj y bi

	a_1	a_2	...	a_j	...	a_p		
b_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1p}	$N_{1.}$	$=\sum_{j=1}^{p} n_{1j}$
b_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2p}	$N_{2.}$	$=\sum_{j=1}^{p} n_{2j}$
b_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ip}	$N_{i.}$	$=\sum_{j=1}^{p} n_{ij}$
b_n	n_{n1}	n_{n2}	...	n_{nj}	...	n_{np}	$N_{n.}$	$=\sum_{j=1}^{p} n_{nj}$
	$N_{.1}$	$N_{.2}$	...	$N_{.j}$	...	$N_{.p}$	N	

$$N_{.1} = \Sigma_{i=1}^{n} n_{i1}$$

$$N_{.2} = \Sigma_{i=1}^{n} n_{i2}$$

$$N_{.j} = \Sigma_{i=1}^{n} n_{ij}$$

$$N_{.p} = \Sigma_{i=1}^{n} n_{ip}$$

$$N = \Sigma_{i=1}^{n} N_{i.} = \Sigma_{j=1}^{p} N_{.j}$$

Perfiles de las filas.

$$Fila\ 1 = b_1 = (\frac{n_{11}}{N_{1.}}, \frac{n_{12}}{N_{1.}}, \ldots, \frac{n_{1j}}{N_{1.}}, \ldots, \frac{n_{1p}}{N_{1.}})$$

$$Fila\ i = b_i = (\frac{n_{i1}}{N_{i.}}, \frac{n_{i2}}{N_{i.}}, \ldots, \frac{n_{ij}}{N_{i.}}, \ldots, \frac{n_{ip}}{N_{i.}})$$

Perfiles promedio o MASA : (N.1 N , N.2 N , . . . , N.j N , . . . , N.p N)

Los perfiles-fila muestran cómo se distribuye la variable A en las modalidades de B. El perfil promedio o MASA muestra la distribución marginal de la variable A.

Si los perfiles de dos filas son idénticos significa que A se comporta igual en ellas, es decir A se distribuye independientemente de las modalidades de B.

Por el contrario, perfiles muy distintos indican dependencia fuerte entre variables.

Análogamente se pueden calcular los perfiles-columna.

Podemos calcular distancias entre filas, distancia de una fila al promedio, distancia entre columnas, distancia de una columna al promedio:

$$d_{x^2}(b_i, b_{i'}) = \sqrt{\frac{(\frac{n_{i1}}{N_{i.}} - \frac{n_{i'1}}{N_{i'}})^2}{\frac{N_{.1}}{N}} + \frac{(\frac{n_{i2}}{N_{i.}} - \frac{n_{i'2}}{N_{i'.}})^2}{\frac{N_{.2}}{N}} + \dots}$$

$$d_{x^2}(b_i, 0) = \sqrt{\frac{(\frac{n_{i1}}{N_{i.}} - \frac{N_{.1}}{N})^2}{\frac{N_{.1}}{N}} + \frac{(\frac{n_{i2}}{N_{i.}} - \frac{N_{.2}}{N})^2}{\frac{N_{.2}}{N}} + \dots}$$

Dependencia e independencia entre variables

Cuando para todo

$$n_{ij} = \frac{N_{i.}N_{.j}}{N}$$

Los perfiles resultan idénticos, ello implica que hay independencia entre las variables.

Por el contrario, cuando

$$\frac{N_{i.}N_{.j}}{N} = n'_{ij} \neq n_{ij}$$

decimos que hay asociación.

El cuantificador clásico de la asociación entre variables es:

$$X^2 = \sum \frac{(n'_{ij} - n_{ij})^2}{n'_{ij}}$$

Ho: Hay independencia entre valores.

Bajo Ho

$$X^2 \approx X^2_{(p-1)(n-1)}$$

Calculamos

$$X^2 = \Sigma \frac{(n'_{ij} - n_{ij})^2}{n'_{ij}}$$

Si

$$X^2 > X^2_{(p-1)(n-1)}$$

Rechazamos la independencia (Tiene sentido factorizar).

Si

$$X^2 \leq X^2_{(p-1)(n-1)}$$

Aceptamos la independencia (No tiene sentido factorizar)

Llamamos inercia de la tabla de contingencia a

$$\frac{X^2}{N}$$

La inercia mide el grado de intensidad en la asociación entre variables.

Cuanto mayor es el valor X^2 calculado, mayor es la inercia que tienen los datos y más sentido tiene factorizar.

Extracción de factores

La inercia total se puede descomponer en:

i_1 (inercia en el eje 1) + i_2 (inercia en el eje 2) + i_3 (inercia en el eje 3) + etc

$i_1 > i_2 > i_3 > i_4 \ldots$

i_1/I = Porcentaje de inercia que retiene el primer eje.

i_2/I = Porcentaje de inercia que retiene el segundo eje.

i_1+i_2/I = Porcentaje que retienen conjuntamente los dos primeros ejes.

Los factores se extraen al diagonalizar la matriz Qf = A.A' con

$$A = \left(\frac{P_{ij}}{\sqrt{P_{i.}}\sqrt{P_{.j}}}\right)_{nxp}$$

y $P_{ij} = n_{ij}/N$ (frecuencias relativas)

$$P_{i.} = \frac{N_{i.}}{N}$$

$$P_{.j} = \frac{N_{.j}}{N}$$

La matriz Q siempre tiene "1" como valor propio (se omite).

Valores propios de Q = 1, λ_1, λ_2, $\lambda_{\min(p,n)-1}$

Número máximo de ejes: min(p,n)-1
Con ellos retenemos el 100% de la inercia

La suma de los valores propios 6= 1 es la inercia.
Para cada valor propio (6= 1) obtendremos un vector propio que normalizamos.

$$(Q - \lambda I)v = \bar{0}$$

Por ejemplo, para _1 obtenemos úv1 = ($ú_{11}$, $ú_{12}$, . . . , $ú_{1n}$) que es el 1er. eje.

Para el segundo valor propio, _2, obtendremos un vector propio normalizado. El segundo

eje es

$ú_2$ = ($ú_{21}$, $ú_{22}$, . . . , $ú_{2n}$)

.

($ú^2_{21} + ú^2_{22} + \ldots + ú^2_{2n} = 1$)

Si $\lambda_1 \neq \lambda_2$ entonces $\acute{u}_1$ y $\acute{u}_2$ son ortogonales.

Así sucesivamente.

Obtenemos

Coordenadas de cada modalidad fila en los nuevos ejes.

Sea λ_1 el mayor valor propio $\neq$ obtendremos $\acute{u}_1$, vector propio $(\acute{u}_{11}, \acute{u}_{12}, \ldots, \acute{u}_{1n})$.

$$B_i^{(1)} = \frac{\sqrt{\lambda_1}}{\sqrt{P_{i.}}} v_{1i}$$

Coordenadas de las filas i en el eje 1.

Sea λ_2 el siguiente valor propio $\acute{u}_1$, su vector propio $(\acute{u}_{21}, \acute{u}_{22}, \ldots, \acute{u}_{2n})$.

$$B_i^{(2)} = \frac{\sqrt{\lambda_2}}{\sqrt{P_{i.}}} v_{2i}$$

Coordenadas de las filas i en el eje 2.

$$\begin{array}{|c|cc|c|cc|} & \text{1er eje} & \text{2 eje} & & & \\ \text{Coord. de } b_1 & \frac{\sqrt{\lambda_1}}{\sqrt{P_{1.}}} v_{11} & \frac{\sqrt{\lambda_2}}{\sqrt{P_{1.}}} v_{21} & & B_1^{(1)} & B_1^{(2)} \\ \text{Coord. de } b_2 & \frac{\sqrt{\lambda_1}}{\sqrt{P_{2.}}} v_{12} & \frac{\sqrt{\lambda_2}}{\sqrt{P_{2.}}} v_{22} & & B_2^{(1)} & B_2^{(2)} \\ \vdots & \vdots & \vdots & = & B_i^{(1)} & B_i^{(2)} \\ \text{Coord. de } b_n & \frac{\sqrt{\lambda_1}}{\sqrt{P_{n.}}} v_{1n} & \frac{\sqrt{\lambda_2}}{\sqrt{P_{n.}}} v_{2n} & & B_n^{(1)} & B_n^{(2)} \end{array}$$

Estas coordenadas son las que se representan en la normalización principal.

Contribución absoluta de cada modalidad fila a la formación del eje.

$$C_{b_1}^a = \frac{P_{1.}}{\lambda_1} (B_1^{(1)})^2$$

$$C_{b_2}^a = \frac{P_{2.}}{\lambda_1} (B_2^{(1)})^2$$

$$C_{b_n}^a = \frac{P_{n.}}{\lambda_1} (B_n^{(1)})^2$$

La suma de las contribuciones absolutas a un mismo eje es 1.

Contribución de $b_1, b_2, \ldots, b_n$ a la formación del 1er. eje siendo $B^{(1)}{}_1, B^{(1)}{}_2, \ldots, B^{(1)}{}_n$ las coordenadas de $b_1, b_2, \ldots, b_n$ en ese primer eje.

Para la contribución de las filas a la formación del segundo eje, se cambiaran las coordenadas por $B^{(2)}{}_1, B^{(2)}{}_2, \ldots, B^{(2)}{}_n$ etc.

Además:

$$C_{b_1}^a = v_{11}^2$$

$$C_{b_2}^a = v_{12}^2$$

$$C_{b_n}^a = v_{1n}^2$$

$$(\sqrt{C^a_{b_1}}, \sqrt{C^a_{b_2}}, \ldots, \sqrt{C^a_{b_n}}) = (v_{11}, v_{12}, \ldots, v_{1n})$$

Contribución relativa de los factores a las modalidades fila.

$$C^i_r = \frac{B^2_i}{d^2(0,i)}$$

Con B^2_i coordenada al cuadrado de la fila considerada y $d^2(0, i)$la distancia X^2 de la fila considerad a la media (MASA) de las filas.

Esto es para cada eje.

La suma de dos contribuciones relativas nos dice como ha quedado representada esa modalidad con esos dos ejes.

Mod	Eje 1	Eje 2
b_1	$\frac{(B^{(1)}_1)^2}{d^2(0,1)}$	$\frac{(B^{(2)}_1)^2}{d^2(0,1)}$
b_2	$\frac{(B^{(1)}_2)^2}{d^2(0,2)}$	$\frac{(B^{(2)}_2)^2}{d^2(0,2)}$
b_i	$\frac{(B^{(1)}_i)^2}{d^2(0,i)}$	$\frac{(B^{(2)}_i)^2}{d^2(0,i)}$

Con $B^{(1)}_1, B^{(1)}_2, B^{(1)}_i$ coord. de b_1, b_2, b_i en el eje 1.

Con $B^{(2)}_1, B^{(2)}_2, B^{(2)}_i$ coord. de b_1, b_2, b_i en el eje 2.

Respecto de las columnas:

Coordenadas de las columnas:

$$A_j = \sum \frac{P_{ij} v_i}{\sqrt{P_{i.}} P_{.j}}$$

$$a_1 = (A^{(1)}_1, A^{(2)}_1) = (\sum \frac{P_{i1} v_1 i}{\sqrt{P_{i.}} P_{.1}}, \sum \frac{P_{i1} v_2 i}{\sqrt{P_{i.}} P_{.1}})$$

$$a_2 = (A^{(1)}_2, A^{(2)}_2) = (\sum \frac{P_{i2} v_1 i}{\sqrt{P_{i.}} P_{.2}}, \sum \frac{P_{i2} v_2 i}{\sqrt{P_{i.}} P_{.2}})$$

$$a_j = (A^{(1)}_j, A^{(2)}_j) = (\sum \frac{P_{ij} v_1 i}{\sqrt{P_{i.}} P_{.j}}, \sum \frac{P_{ip} v_2 i}{\sqrt{P_{i.}} P_{.j}})$$

$$a_p = (A^{(1)}_p, A^{(2)}_p) = (\sum \frac{P_{ij} v_1 i}{\sqrt{P_{i.}} P_{.p}}, \sum \frac{P_{ip} v_2 i}{\sqrt{P_{i.}} P_{.p}})$$

(Para una misma modalidad las coordenadas en el 1er y 2 eje solo cambian los coeficientes $ú_{ij}$ del vector propio).

(Para distintas modalidades cambia la columna de referencia y su total de $P_{.j}$). Contribución absoluta de cada modalidad columna a la formación del eje:

$$C_j^a = \frac{P_{.j}}{\lambda} A_j^2$$

	eje 1	eje 2
C_{a1}^a	$\frac{P_{.1}}{\lambda_1}(A_1^{(1)})^2$	$\frac{P_{.1}}{\lambda_2}(A_1^{(2)})^2$
C_{a2}^a	$\frac{P_{.2}}{\lambda_1}(A_2^{(1)})^2$	$\frac{P_{.2}}{\lambda_2}(A_2^{(2)})^2$
$\vdots$	$\vdots$	$\vdots$
C_{aj}^a	$\frac{P_{.j}}{\lambda_1}(A_j^{(1)})^2$	$\frac{P_{.j}}{\lambda_2}(A_j^{(2)})^2$

(La suma de todas las contribuciones absolutas a la formación de un mismo eje es 1).

Contribución relativa de los ejes a las modalidades columna:

$$C_j^r = \frac{A_j^2}{d^2(0,j)}$$

	Eje 1	Eje 2
C_{a1}^r	$\frac{(A_1^{(1)})^2}{d^2(0,j)}$	$\frac{(A_1^{(2)})^2}{d^2(0,j)}$
C_{a2}^r	$\frac{(A_2^{(1)})^2}{d^2(0,j)}$	$\frac{(A_2^{(2)})^2}{d^2(0,j)}$
$\vdots$	$\vdots$	$\vdots$
C_{ar}^r	$\frac{(A_j^{(1)})^2}{d^2(0,j)}$	$\frac{(A_j^{(2)})^2}{d^2(0,j)}$

$d^2(0, j) = (\text{distancia}X^2)^2$ entre el perfil de la columna promedio (sin raíz) y la columna considerada.

(La suma por filas me da la calidad de la representación para la modalidad Aj considerada).

Correspondencias múltiples

Más de dos variables ⟶ tabla disyuntiva completa.

Procedimiento de normalización y diagonalización diferente. Hay una inercia que solo depende del número de variables y modalidades.

$$I = \frac{N.modalidades}{N.variables} - 1$$

El grado de asociación viene dado por el reparto de esa inercia entre ejes.

Reparto equilibrado ⟶ Independencia

Reparto desequilibrado ⟶ asociación

N.max ejes = n. de modalidades - n variables.

Ordenes en SPSS para Análisis Factorial de Correspondencia (AFC)

A) Para introducir una tabla de contingencia

Abrir un archivo nuevo de datos y definir tres variables: En una se introducirán las modalidades-fila, en otra las modalidades-columna y en la tercera las frecuencias conjuntas. (Por ejemplo, si es una tabla 3x4 el archivo tendrá 12 filas y 3 columnas).

Datos Ponderar casos Ponderar mediante la variable frecuencia

(En la parte inferior derecha de la pantalla aparecerá la palabra "ponderado")

B) Análisis factorial de correspondencias

Analizar Reducción de datos Análisis de correspondencias

Introducir en fila: la variable correspondiente a las modalidades-fila

Definir rango (mínimo 1 máximo: nº de filas de la tabla) actualizar

Introducir en columna: la variable correspondiente a las modalidades-columna

Definir rango (mínimo 1 máximo: nº de columnas de la tabla) actualizar

En **MODELO**: Cambiamos el método de normalización a Principal

(el resto como está)

En **ESTADISTICOS**: añadimos a las opciones marcadas (tabla, inspección de filas, inspección de columnas) los perfiles fila y perfiles columna

SIN entrar en **GRÁFICOS**: Aceptar (Aparecerá toda la salida excepto gráficos)

Para obtener el gráfico conjunto con todas las modalidades fila y columna:

Repetimos el análisis señalando en **MODELO**: método de normalización Simétrica

(de forma automática en el apartado **GRAFICOS** queda señalada la opción diagrama de dispersión biespacial) Aceptar (obtenemos la salida completa donde únicamente habrán cambiado las coordenadas y aparece un gráfico que hemos de modificar)

Modificación del gráfico: doble click en el gráfico

Galería Dispersión Simple Reemplazar.

Diseño Línea de referencia ….. (para obtener las líneas de los ejes)

Formato Marcadores (si queremos que las modalidades-fila y las modalidades- columna tengan distinto símbolo, tamaño,…)

C) Gráfico de barras

En el menú principal ir a **GRAFICOS** Barras Agrupado Definir

Las barras representan: número de casos

Eje de categorías: introducir la variable que deseemos en abcisas

Definir grupos por: la otra variable Aceptar

Ejemplo 18. 2.

El análisis de correlaciòn de Spearman y el correspondencia múltiple (ACM), fueron realizado para el ejemplo 18.1, pero lo discutimos en esta secciòn para seguir un orden y contribuya a la compresiòn del lector.

La Tabla 18.5. muestra una matriz con las correlaciones de Spearman entre variables cualitativas.

Tabla 18.5. Coeficientes de correlación de Spearmanentre seis variables cualitativas evaluadas en 30 híbridos de tomate (Fuente: López 2014)

	TCP	**FPF**	**CFM**	**CCP**	**CTF**	**NOL**
CTP	**1**					
PFP	**-0.62**	**1**				
CFM	0.23	-0.51	**1**			
CCP	0.14	0.23	0.09	**1**		
CTF	0.23	-0.26	0.32	-0.13	**1**	
NOL	0.08	**-0.41**	0.35	-0.07	0.33	**1**

TCP = Tipo de crecimiento de la planta, FPF = Forma predominante del fruto, CFM = Color exterior del Fruto maduro, CCP = Color de la carne del pericarpio (interipr), CFT = Forma del corte trasnversal del fruto, NOL = Nùmero de lòculos- **: Altamente significativo al $p<0.01$ de probabilidad, *: Significativo al $p<0.05$ de probabilidad.

Hubo correlación negativa altamente significativa (r = -0.62) entre las variables TCP y FPF (Tabla 16.2). Esto indicaría que la FPF depende del TCP. En las plantas indeterminadas predominan los frutos redondos y redondos achatados (92%) y solamente el 5% fueron redondos alargados. En cambio, en las determinadas el porcentaje de frutos redondos alargados fue mayor (30%) que en las indeterminadas y el porcentaje de frutos redondos y redondos achatados fue inferior que en las indeterminadas (65%). Asimismo, se presentaron frutos cordiformes (tipo pera) en un 5%. La variable FPF se correlacionó negativa altamente significativa con la variable CFM (r = -0.51). Esto mostraría que en los frutos alargados predomina el color naranja y en los frutos redondos los colores rojo y rosado. Asimismo, la variable FPF se correlacionó negativa y altamente significativa con la variable NOL (r = -0.41). Lo que denota que en los frutos alargados predominan tres lóculos por fruto y que en los frutos redondos predominan cuatro lóculos.

El **ACM** mostró variabilidad genética dentro de los 30 híbridos evaluados de tomate (Figura 17.2). Aunque la mayoría de los híbridos se agruparon por características similares, hubieron casos aislados como el híbrido 20 que no se agrupo con ninguno de los híbridos. Este híbrido tuvo frutos con color exterior rosado y carne del pericarpio del mismo color (características únicas de este hibrido). Por otro lado, los híbridos 2, 17 y 23 se agruparon al presentar frutos redondos alargados con un color exterior naranja y la carne del pericarpio del mismo color. Otros híbridos que se diferenciaron fueron los híbridos 3, 5, 18, 27 y 28 al presentar forma redonda y un color exterior rojo y la carne del pericarpio también del mismo color. En general hubo una correlación entre la forma predominante del fruto, el color exterior y el color del pericarpio.

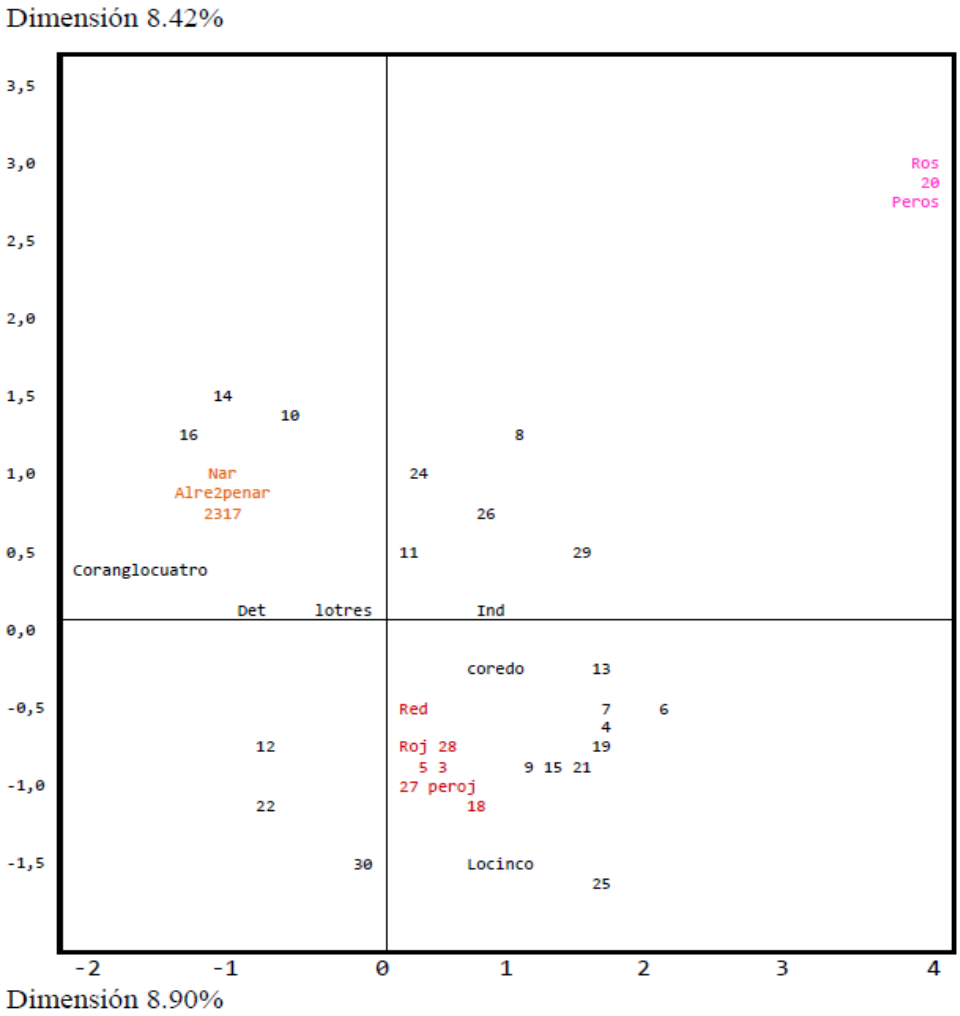


Figura 18.2. Asociación entre estados de cada una de las variables cualitativas (Fuente: López 2014, López *et al.* 2015). Tipo de crecimiento de las plantas (Ind = Indeterminadas, Det = Determinadas); Forma predominante del fruto (Alre = Redondo alargado, Red = Redondo, Lach = Ligeramente achatado);Color exterior del fruto maduro (Roj = Rojo, Nar = Naranja, Ros = Rosada);color carne del pericarpio (Peroj = Rojo, Penar = Naranja, Peros = rosado); Corte transversal del fruto (Coredo = Redondo, Corang = Angular); Numero de lóculos (Lotres = tres lóculos, Locuatro = cuatro lóculos y Locinco = cinco lóculos) y los números son los 30 híbridos de tomate.

18.2. Análisis de conglomerados (clúster)

Es un método analítico que se puede aplicar para clasificar las accesiones de un germoplasma (o variables) en grupos relativamente homogéneos con base en alguna similitud existente entre ellas (Franco e Hidalgo 2005).

El objetivo es reunir las observaciones en grupos (clúster) en función de su semejanza. Se trata de que las observaciones semejantes entre si se hallen en un mismo grupo (homogeneidad interna) y los grupos difieran unos de otros (heterogeneidad entre grupos).

El método de conglomerados o análisis de clúster se puede aplicar sobre una matriz básica de datos **n** x **p** o sobre una matriz **n** x **n**, o **p** x **p**, donde **n** es el número de accesiones que se quieren agrupar y **p** son las variables.

Es importante aclarar que el análisis de conglomerados se aplica sobre una matriz de distancias y no sobre una de similitud. Para descriptores cualitativos, esta última debe ser transformada en una de distancia. Para datos cuantitativos, los programas de estadística actualmente disponibles calculan directamente los valores de distancia según el método que se aplique.

Básicamente los métodos de agrupamiento más usados en el análisis conglomerado son: 1) jerárquico, que forma grupos a varios niveles; y 2) no jerárquico o de partición que también forma grupos a través de criterios predefinidos (Franco e Hidalgo 2005).

Relación con otras técnicas

- comparte con las técnicas factoriales la idea de reducción, solo que ahora se trata de agrupar observaciones (filas) y no variables (columnas)
- comparte con el análisis discriminante la idea de clasificar, solo que ahora los grupos no existen a priori (no hay una variable que indique la pertenencia de las observaciones a grupos), sino que los grupos se van a construir a partir de la semejanza entre observaciones respecto a las variables disponibles

Procedimientos

Existen distintos métodos de agrupación que se clasifican en: técnicas jerárquicas y técnicas de optimización.

En las técnicas jerárquicas los grupos se van formando en sucesivas etapas por unión (caso ascendente) o por división (caso descendente) de grupos ya existentes, de modo que si dos observaciones se han unido (caso ascendente) o se han separado (caso descendente), permanecerán en ese mismo estado (unidas o separadas) hasta el final del proceso.

En las técnicas de optimización se fija una función a optimizar (por ejemplo el cociente entre la variabilidad externa y la interna) y generalmente mediante procedimientos iterativos se van ubicando y reubicando las observaciones en distintos grupos. En cada etapa se mide la “mejoría” experimentada por la función a optimizar y, cuando la mejoría no supera cierto umbral prefijado, el proceso se detiene.

Tanto en unas como en otras existen distintas variantes que pueden producir resultados diferentes.

Aspectos teórico-prácticos a considerar

1) La influencia de las unidades de medida de las variables en los resultados. Se aconseja, particularmente cuando las variables se miden en unidades diferentes unas de otras, realizar el proceso con variables tipificadas.
2) La estabilidad de la solución. Valorar los cambios que se producen el resultado final (es decir, qué observaciones pertenecen a cada grupo) al introducir cambios (distintas variantes) en el proceso. Los grupos son más consistentes (podríamos decir que están geométricamente más diferenciados) cuando, pese a utilizar distintas variantes en el procedimiento de agrupación, el resultado final es similar.
3) Comparación entre diferentes soluciones (que posean el mismo número de grupos):
   a) utilizando solamente las variables del análisis: la mejor es aquella en la que el cociente entre la variabilidad externa y la interna sea máximo (L de Wilks mínima).
   b) utilizando otras variables auxiliares distintas a las del análisis: puede considerarse mejor aquella solución donde la variable auxiliar se comporte de forma más diferenciada en los grupos.
4) Número de grupos en la solución final: no hay reglas fijas. Puede depender del criterio del analista. Si nos basamos únicamente en las variables del análisis, se pueden utilizar criterios basados en la minimización de la variabilidad interna: siempre que aumentemos el número de grupos ocurre que la variabilidad interna disminuye, con lo cual merecerá la pena considerar k+1 grupos en lugar de k si la disminución que se consigue es superior a un cierto umbral. También podemos guiarnos por la evolución del proceso y detenerlo cuando se produzca un "salto" importante (se ve en la salida de ordenador)
5) Caracterización de los grupos. Tan importante como identificar los grupos es caracterizarlos, observar cómo son. Para ello se pueden calcular las medias de las variables utilizadas en el análisis en cada grupo y representarlas gráficamente para ver su perfil.
6) En ocasiones y para evitar que tengan excesiva influencia en el resultado varias variables muy correlacionadas entre sí pueden ser conveniente aplicar componentes principales y, con los factores resultantes, realizar el clúster.

Descripción del proceso en el caso jerárquico ascendente

En este caso se considera que cada una de las N observaciones es un clúster y el proceso consiste en ir uniendo dichas observaciones en grupos, hasta llegar a configurar un número de grupos fijado de antemano por el analista (por ejemplo 2, 3 o 4 grupos) o bien observar el proceso hasta el final (cuando todas las observaciones configuran un único grupo) para después decidir el número de grupos a considerar como mejor solución.

Dado que vamos buscando observaciones semejantes (en las variables consideradas) hay una primera decisión a tomar ¿qué medida de semejanza entre observaciones se va a elegir?

Cuando las variables son **cuantitativas**, las más usadas son:

- Distancia euclídea
- Distancia euclídea al cuadrado
- Distancia de bloques

Con la medida elegida, se construye la matriz de distancias entre observaciones (matriz NxN, simétrica y con ceros en la diagonal principal).

Observando esta matriz, encontraremos la pareja de observaciones "más parecidas". Son las que dan lugar a la primera unión, al primer clúster.

En ese momento, disponemos de N-2 observaciones y un clúster con dos observaciones. Para continuar el proceso, debemos tomar otra decisión ¿qué medida de distancia entre clúster (vinculación) se va a utilizar?

Dentro de las más utilizadas están:

- Vinculación vecino más próximo (simple)

D_{AB} = min d_{ij} (la menor de las distancias entre pares de observaciones)

Donde i representa cualquier observación del clúster A; j representa cualquier observación del clúster B

- Vinculación vecino más lejano (completa)

D_{AB} = max d_{ij} (la mayor de las distancias entre pares de observaciones)

Donde i representa cualquier observación del clúster A; j representa cualquier observación del clúster B

- Vinculación inter-grupos (promedio)

$D_{AB} = S\ d_{ij} / n_A\ n_B$ (la media aritmética de las distancias entre pares de observaciones)

Donde i representa cualquier observación del clúster A; j representa cualquier observación del clúster B; n_A y n_B representan los tamaños (nº de observaciones) de los clústers A y B respectivamente.

Una vez elegida la vinculación, procedemos a construir una nueva matriz de distancias, en este caso N-1 x N-1 (que será simétrica, con ceros en la diagonal principal) cuyos elementos habrán sido calculados de acuerdo a la vinculación elegida. De esta matriz se localiza el menor valor, que da lugar a la segunda unión.

Y se reitera el proceso. La vinculación elegida se ha de mantener durante todo el proceso y las d_{ij} se toman siempre de la primera matriz de distancias.

Ejemplo Se dispone de 6 observaciones a, b… de las que se han medido dos variables X_1 y X_2. Realizar el proceso de agrupamiento, utilizando la distancia euclídea al cuadrado y la vinculación promedio (en este ejemplo se realizará sin tipificar)

Datos

Observaciones	**X_1**	**X_2**
A	1	6
B	2	6.5
C	2	7
D	2.5	3
E	6	4
F	9	6

Calculo de distancias (euclídea al cuadrado):

$$d_{ab} = (1-2)^2 + (6-6.5)^2 = 1.25$$ lo mismo las demás.

Primera matriz de distancias

	a	**b**	**c**	**d**	**e**	**f**
a	0	1.25	2	11.25	29	64
b		0	0.25	12.5	22.25	49.25
c			0	16.25	25	50
d				0	13.25	51.25
e					0	13
f						0

El primer clúster lo forman b y c a la distancia de 0.25.

Segunda matriz de distancias

	a	**bc**	**d**	**e**	**f**
a	0	1.625	11.25	29	64
bc		0	14.375	23.625	49.625
d			0	13.25	51.25
e				0	13
f					0

Con la vinculación promedio inter-grupos: (solo cambiarán respecto a la primera matriz las distancias respecto al grupo bc , en cursiva)

$d_{bc,a} = (d_{ab} + d_{ac})/2 = (1.25+2)/2 = 1.625$ las demás igual Por tanto en esta 2ª etapa se une "a" al grupo bc a la distancia de 1.625.

Tercera matriz de distancias

	abc	**d**	**e**	**f**
abc	0	13.333	25.417	54.417
d		0	13.25	51.25
e			0	13
f				0

Donde 13.333 = (11.25 + 12.5 + 16.25)/3 donde se han promediado tres distancias. En esta etapa se unen e y f a la distancia 13

Cuarta matriz de distancias

	abc	**d**	**ef**
abc	0	13.333	¿?
d		0	¿?
ef			0

Continuar el ejercicio hasta unir las seis observaciones.

Referencias

Bansal AK. Bioinformatics in microbial biotechnology – a mini review. Microb Cell Fact. 2005; 4:19.

Barnett V (1981) "Interpreting Multivariate Data". Ed. Wiley.

Bayat A. Science, medicine, and the future: bioinformatics. BMJ. 2002; 324: 1018 - 1022.

Benza J (1982) Métodos Estadísticos para la Investigación. Universidad La Molina, Lima, Perú

Bernstein Ira H (1987) "Applied Multivariate Analysis". Springer-Verlag, 1st Edition. Ships from the UK.

Burgos López G (2019) araterizacion de lineas parentales de melón (*Cucumis melo* L.) para obtener semilla de hibridos F l en invernadero. Tesis para obtener la Lic. En Ing. Agropecuaria, Universidad Estatal del Sur de Manabí, Jipijapa, Ecuador. 108 p.

Caballero W (1985) Introducción a la estadística. IICCA, San José, Costa Rica. 289 p.

Ching Chun Li (1977) Introducción a la estadística experimental. Editorial Omega, Madrid, España.

Cochran GW y Cox MG (1990) Diseños experimentales. 2da. Ed., Trillas, México D. F., México. 661 p.

Dean A, Voss D (1999) Design and Analysis of Experiments. Springer, Netherlands. 735 p.

Dillon, W. y Goldstein, M. (1984). "Multivariate Analysis". Ed. Wiley

Fehr NR (1993) Principies of cultivar development. Theory and technique. Department of Agronomy, Iowa State University. Ames, Iowa, USA. pp. 315-438.

Franco TL, Hidalgo R (eds.). (2003) Análisis Estadístico de Datos de Caracterización Morfológica de Recursos Fitogenéticos. Boletín técnico no. 8, Instituto Internacional de Recursos Fitogenéticos (IPGRI), Cali, Colombia. 89 p.

Gabriel J (2008) Aplicación de marcadores moleculares para cribado de QTLs en diferentes fuentes de resistencia a tizón tardío (*Phytophthora infestans*) en papa. Tesis para obtener el PhD. en Producción agraria y Aplicaciones biotecnológicas, Unversidad Pública de Navarra, Pamplona, España. 108 p.

Gabriel J, López E, Magne J, Angulo A, Luján R, La Torre J, Crespo M (2013) Genetic basis of inheritance for morphological, agronomic and agro-industries characteristics in hybrid tomato Solanum lycopersicum L. (Mill) (en línea). J Selva Andina Biosph. 1(1):45-54. Diponible en http://www.scielo.org.bo/pdf/jsab/v1n1/v1n1_a05.pdf

Gómez S (2017) Uso de software para análisis de datos cuantitativos. Disponible en http://softwareanalisisdedatoscuantitativos.blogspot.com/2014/04/tipos-de-paquetes-estadisticos.html.

Goodman N (2002) Biological data becomes computer literate: new advances. Curr Opin Biotechnol. 13: 68 – 71.

Hair JF, Anderson RE, Tatham RL, Black WC. (1999): “Análisis Multivariante”. Prentice Hall Iberia, Madrid.

Holguin Flores G (2019) Comportamiento morfológico del café (*Coffea arábica* L.) sarchimor 4260 en etapa de crecimiento con fertilizantes químicos y orgánicos. Tesis para obtener la Lic. En Ing. Agropecuaria, Universidad Estatal del Sur de Manabí, Jipijapa, Ecuador. 80 p.

Infante GS, Zárate de Lara GP (1991) Métodos estadísticos. Trillas, México D.F., México. 643 p.

Kendall M (1980) "Multivariate Analysis". 2.ed. London: Charles Griffin, 1980. 210p.

Little TM, Hills FJ (1991) Métodos estadísticos para la investigación en la agricultura. Trad. Del inglés por Anatolio de Paula Crespo. Trillas, México D.F., México. 270 p.

López E (2014) Estimación de la herencia y varianzas genéticas para caracteres morfológicos, agronómicos y agroindustriales en híbridos de tomate **[***Solanum lycopersicumL. (Mill.)]*. Tesis Ing. Agrònomo, Fac. de Ciencias Agrìcolas y Pecuarias "Martín Cárdenas", Universidad Mayor de San Simón, Cochabamba, Bolivia. 69 p.

López E, Gabriel J, Angulo A, Magne J, La Torre J, Crespo M (2015) Herencia y relación genética asociados al rendimiento, madurez en híbridos de tomate [*Solanum lycopersicum* L. (MILL.)] (en linea). Agronomía Costarricense 39(1): 107-119. Disponible en https://www.scielo.sa.cr/pdf/ac/v39n1/a08v39n1.pdf

Malm NR, Rachie KO (1971) The setaria millets; a review of the world literature. S. B. 513. University of Nebraska College of Agriculture. Lincoln, Nebraska, USA. pp. 7-12.

Marascuilo R, Calvin JR (1982). "Multivariate Statistics in the Social Sciences". Ed. Books Cole

Mardia KV, Kent JT, Bibby JM (1979) "Multivariate Analysis". Ed. Academic Press.

Martínez GA (1988) Diseños experimentales: Métodos y elementos de teoría. Trillas, México D. F., México. 756 p.

Muñoz A (2007) Aprendizaje del Software Estadístico R: un entorno para simulación y computación estadística. Departamento de Estadística, Universidad Carlos III de Madrid, Madrid, España. http://ocw.uc3m.es/estadistica/aprendizaje-del-software-estadistico-r-un-entorno-para-simulacion-y-computacion-estadistica.

Ostle B (1994) Estadística aplicada. Limusa, México D. F., México. 629 p.

Pardo A y Ruiz MA (2002) SPSS 11: Guía para el análisis de datos. Madrid: McGraw-Hill. ISBN 9788448137502.

Payne R (2008) A Guide to REML in GenStat. 15th Edition. VSN International, 5 The Waterhouse, Waterhouse Street, Hemel Hempstead, Hertfordshire HP1 1ES, UK. 88 p.

Portilla M. Eraso S, Gale C, García I, Moler JA, Palacios MB (2002) "Manual práctico del paquete estadístico SPSS para windows". Edita Universidad Pública de Navarra.

Quijije Quiróz K (2018) Uso de ácidos orgánicos para mejorar los parámetros zootécnicos y la calidad de la carcasa de pollos de engorde. Tesis para obtener la Lic. en Ing. Agropecuaria, Universidad Estatal del Sur de Manabí, Jipijapa, Ecuador. 46 p.

Quinn GP, Keough MJ (2002) Experimental Design and Data Analysis for Biologists. Cambridge University Press, United Kingdom. 527 p.

Ritter E, Ruiz de Galarreta JI, Hernandez M, Plata G, Barandalla L, Lopez R, Sanchez I, Gabriel J (2009) Utilization of SSR and cDNA markers for screening known QTLs for late blight (Phytophthora infestans) resistance in potato. Euphytica: 1-10. DOI 10.1007/s10681-009-9986-4

Ruíz-Ramirez J (2010) Eficiencia relativa y calidad de los experimentos de Fertilización en el cultivo de caña de azúcar. Terra Latinoamericana 28(2): 149 – 154.

SAS Institute Inc. SAS/STAT (2004) Users Guide, Version 9.2, Fourth Edition, Vol. 2, SAS Institute Inc., Cary, N.C. 2004.

Searle SR, Casella G, McCulloch CE (1992) Variance components. John Willey & Sons, Inc. lthaca, New York, USA. pp. 258-289.

Siles - Cano M (2005) Diseño de bloques incompletos con una sola repetición en la evaluación preliminar de material avanzado en programas de mejoramiento genético de cultivos. Revista de Agricultura 34: 1 – 5.

Snedecor GW, Cochran WG (1989) Statistical Methods. Hardcover, Iowa State University Press. ISBN 10: 0813815614 / ISBN 13: 9780813815619

Steel DDR, Torrie HJ (1990) Bioestadística: Principios y procedimientos. Trad. del Inglés por Ricardo Martínez. 2da. Ed., Mc Graw Hill, México D. F., México. 622 p.

Uriel E, Aldás J (2005) Análisis multivariante aplicado; aplicaciones al marketing, investigación de mercados, economía, dirección de empresas y turismo. Paraninfo Cengage Learning, Madrid, España. 531 p.

Xiong J (2006) Essential bioinformatics. 1ra ed. Cambridge: Cambridge University Press.

Descubre tu próxima lectura

Si quieres formar parte de nuestra comunidad,
regístrate en https://www.grupocompas.org/suscribirse
y recibirás recomendaciones y capacitación

compAs

Grupo de capacitación e investigación pedagógica

@grupocompas.ec
compasacademico@icloud.com

Autores

Blanca Indacochea Ganchozo
Ingeniero Forestal. Maestra en Ciencias en Agroecología y Agricultura Sostenible. Doctora en Ciencias Forestales. Facultad de Ciencias Naturales y de la Agricultura, Universidad Estatal del Sur de Manabí, Jipijapa, Manabí, Ecuador.

Carlos Castro Piguave
Ingeniero Agropecuario. Maestro en Ciencias en Administración Ambiental. Facultad de Ciencias Naturales y de la Agricultura, Universidad Estatal del Sur de Manabí, Jipijapa, Manabí, Ecuador.

Máximo Vera Tumbaco
Ingeniero Agropecuario. Maestro en Ciencias en Administración Ambiental. Facultad de Ciencias Naturales y de la Agricultura, Universidad Estatal del Sur de Manabí, Jipijapa, Manabí, Ecuador.

José Alcívar Cobeña
Ingeniero Zootecnista. Maestro en Ciencias en Gestión Ambiental. Universidad Estatal del Sur de Manabí, Jipijapa, Manabí, Ecuador.

Raquel Vera Velásquez
Licenciada en Educación (matemáticas). Máster en Ciencias de la Educación. Facultad de Ciencias Naturales y de la Agricultura, Universidad Estatal del Sur de Manabí, Jipijapa, Manabí, Ecuador.

Editores

Julio Gabriel Ortega
De nacionalidad boliviana, Ing. Agrónomo de la Facultad de Ciencias Agrícolas, Pecuaria, Veterinarias y Forestales de la Universidad Mayor de San Simón, Cochabamba, Bolivia. Maestro en Ciencias (MSc) en Genética del Colegio de Posgraduados, Montecillo, México. Diplomado en Educación Superior de la UNITEPC, Bolivia, Diplomado en Formación Gerencial de la Universidad Católica Boliviana, Bolivia. Diplomado en Estudios Avanzados de la Universidad Pública de Navarra, España. Doctor (PhD) en Producción Agraria y Aplicaciones Biotecnológicas de la Universidad Pública de Navarra, España. Trabajo 27 años en la Fundación PROINPA en Bolivia como investigador, Editor Principal de la Revista Latinoamericana de la Papa de la ALAP, parte del Comité Editor de la Revista UNESUM Ciencia en Ecuador, co-editor de la Revista de Agricultura en Bolivia y par evaluador en varias revistas científicas. Publicó más de 50 artículos científicos en revistas indexadas y varios libros. Actualmente Docente – investigador en la Universidad Estatal del Sur de Manabí, Ecuador.

Alfredo Valverde Lucio
De nacionalidad ecuatoriana, realizo sus estudios de pre-grado en la Universidad Laica "Eloy Alfaro de Manabí", donde obtuvo el título de Ing. Agropecuario, los estudios de cuarto nivel los realizó en la Universidad Tecnológica Indoamérica de Ambato, Ecuador, obteniendo el título de Magister (MSc) en Gestión de Proyectos socio productivos. Con 15 años de experiencia en proyectos de desarrollo social y productivos en el campo agropecuario, laborando para diferentes programas de desarrollo rural y urbano marginal. Actualmente trabaja como profesor Titular en la Universidad Estatal del Sur de Manabí (UNESUM), Carrera de Ingeniería Agropecuaria, dictando las asignaturas tales como Diseño experimental, Proyectos Agropecuarios, Estadística, Bioquímica, Pasto y Forrajes, Extensión Rural.

compAs
Grupo de capacitación e investigación pedagógica

@grupocompas.ec
compasacademico@icloud.com

ISBN: 978-9942-33-381-0

@grupocompas.ec
compasacademico@icloud.com

compAs
Grupo de capacitación e investigación pedagógica