ANOVA (Análisis de varianza)

Las pruebas de hipótesis son una herramienta útil cuando se trata de comparar dos tratamientos. La experimentación usualmente requiere comparación de más de dos tratamientos simultáneamente, es allí donde se introduce Anova (teniendo en cuenta que es un procedimiento para análisis de factores cualitativos).

El análisis de varianza se deriva de la partición de la variabilidad total en las partes que la componen. ANOVA establece que la variabilidad total en los datos, medida por la suma de cuadrados total, puede ser dividida en una suma de cuadrados de la diferencia entre los promedios de los tratamientos y el gran promedio total más una suma de cuadrados de la diferencia de las observaciones entre tratamientos del promedio del tratamiento. Anova, nos da la herramienta para distinguir si un factor afecta la respuesta en promedio.

Presunciones de anova:

1. Los errores o residuales son independientes y distribuidos de manera normal o gaussiana, con promedio equivalente a 0 y varianza constante. Si su promedio no fuese 0, el modelo estaría subestimando o sobreestimando.
2. Anova presume que todas las varianzas de los niveles del factor son iguales y toma un solo cálculo de varianza llamado S_{pooled} o varianza conjunta.

Anova mira los promedios de cada nivel contra el promedio general y lo llama entre tratamientos. Anova queda con dos estimados de varianza, dentro y entre los niveles; con estos saca un cociente, si las 2 varianzas se parecen, es decir, el cociente es aproximadamente 1, el factor no tiene ningún impacto en la respuesta, pero si este cociente resulta ser grande, entonces el factor tiene mucho impacto en la respuesta.

Para ilustrar se presenta a continuación un ejemplo teniendo en cuenta un solo factor aleatorio:

	Observaciones (n replicas)						
Niveles del factor	1	2	…	n	Totales $Y_{i.}$	Promedios $\overline{Y}_{i.}$	
1	Y_{11}	Y_{21}	…	Y_{n1}	$Y_{11}+ Y_{21}+\dots Y_{n1}$	$\overline{Y}_{1.}$	
2	Y_{12}	Y_{22}	…	Y_{n2}	$Y_{12}+ Y_{22}+\dots Y_{n2}$	$\overline{Y}_{2.}$	
. .	. .	. .	…	. .	. .	…	
a	Y_{1a}	Y_{2a}	…	Y_{na}	$Y_{1a}+ Y_{2a}+\dots Y_{an}$	$\overline{Y}_{a.}$	
Totales					Y..	$\overline{Y}_{..}$	

A partir de la anterior tabla, se presenta la forma manual de hacer Anova con el fin de entender el concepto que maneja el análisis de varianza. Inicialmente se debe calcular la suma de cuadrados de los tratamientos:

$$SS_{Tratamientos} = (\frac{1}{n}\sum_{i=1}^{a} Y_{i.}^2) - \frac{Y_{..}^2}{N} \longrightarrow$$

Fuente de variación entre tratamientos

Donde:

n = Numero de tratamientos por cada nivel

N = Numero de tratamientos en total

i = 1, 2, 3… a

Luego se debe calcular la suma de cuadrados total:

$$SS_{Total} = (\sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij}^2) - \frac{Y_{..}^2}{N}$$

Donde:

N = Numero de tratamientos en total

i = 1, 2, 3… a

j = 1, 2, 3…n

Para estimar la suma de cuadrados de los errores se hace la diferencia de la suma de cuadrados total y la suma de cuadrados de los tratamientos:

$$SS_E = SS_{Total} - SS_{Tratamientos} \longrightarrow$$ Fuente de variación dentro de los tratamientos

La tabla de Anova quedaría así:

ANOVA				
Fuente de variación	**Suma de cuadrados (SS)**	**Grados de libertad**	**Promedio de los cuadrados (MS)**	**Estadístico de prueba Fo**
Tratamientos	SS tratamientos	a-1	$\frac{SS_{tratamientos}}{a-1}$	$\frac{MS_{tratamientos}}{MS_{error}}$
Error	SS error	N-a	$\frac{SS_{error}}{N-a}$	
Total	SS total	N-1		

Experimento de un solo factor aleatorio.

Este tipo de experimento es el más sencillo y consiste en analizar un solo factor evaluado en diferentes niveles, de manera que se compara las medias de la respuesta en cada uno de esos niveles y se establece si hay diferencia entre ellas.

El modelo correspondiente a este experimento esta dado por la ecuación IV.

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

Donde μ es un parámetro común para todos los tratamientos llamado la media general, τ representa el efecto del tratamiento i y ε_{ij} corresponde al error que incorpora todas las fuentes de variabilidad en el experimento.

Las hipótesis evaluadas son:

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

$$H_1 : \tau_1 \neq \tau_2 \neq ...\tau_a$$

Lo que se desea investigar es si existe diferencia o no entre los niveles del factor en consideración.

Ejemplo 1 (Tomado del libro Design and analysis of Experiments, 6 edicion, pagina 70)

En muchos procesos de manufactura de circuitos integrados, los “wafers” son revestidos con una capa de material como dióxido de silicona o un metal. Luego, el material que no se necesita es removido haciendo los grabados necesarios para crear los patrones de los circuitos, interconexiones eléctricas y áreas donde se hacen los depósitos de metal. Un proceso de grabado tipo plasma es ampliamente usado para esta operación. La energía para el proceso es suplida por un generador de radio frecuencia RF que hace que el plasma sea generado en el intervalo entre electrodos. El ingeniero del proceso esta interesado en determinar si diferentes niveles de poder de la RF afecta la tasa de grabado. Debido a que se tiene un solo factor, el ingeniero ha decidido

hacer un experimento de un solo factor aleatorio con 5 replicas. Al correr el experimento se obtuvo las siguientes respuestas:

Poder RF (W)	Tasa de grabado observada (replicas)					Totales	Promedios
	1	2	3	4	5	$Y_{i.}$	$\overline{Y}_{i.}$
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0
						Y.. = 12,355	$\overline{Y}_{..}$ = 617.75

Ahora, las hipótesis que el investigador desea probar son:

Ho: Las medias de los niveles son iguales $\mu_{160} = \mu_{180} = \mu_{200} = \mu_{220}$

H_1: Algunas medias son diferentes

Teniendo claras las hipótesis y habiendo corrido el experimento, se procede a realizar los cálculos matemáticos que permitan llegar al estadístico de prueba Fo para tomar una decisión.

$$SS_{Total} = (\sum_{i=1}^{a}\sum_{j=1}^{n} Y_{ij}^2) - \frac{Y_{..}^2}{N} = (575^2 + 542^2 + ... + 710^2) - \frac{12{,}355}{20} = 72{,}209.75$$

$$SS_{Trat} = (\frac{1}{n}\sum_{i=1}^{a} Y_{i.}^2) - \frac{Y_{..}^2}{N} = \frac{1}{5}[2756^2 + ... + 3535^2] - \frac{12{,}355}{20} = 66{,}870.55$$

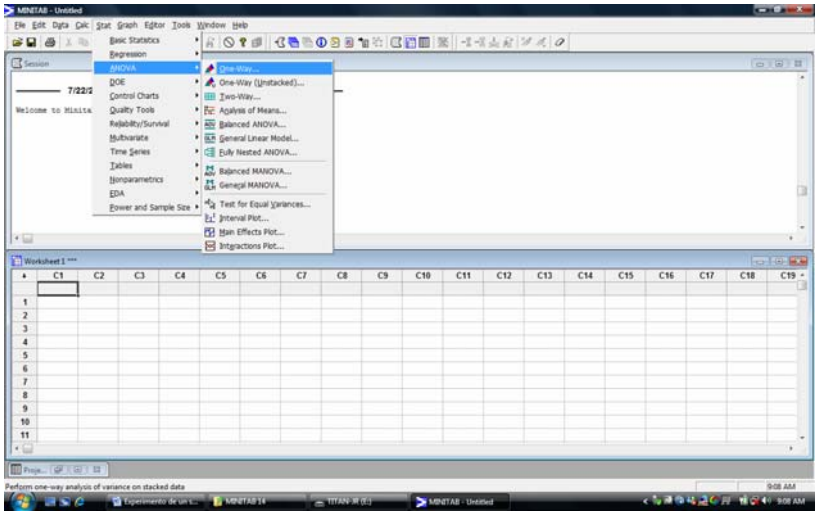
$$SS_E = SS_{Total} - SS_{Tratamientos} = 72{,}209.75 - 66{,}870.55 = 5339.20$$

ANOVA				
Fuente de variación	**Suma de cuadrados (SS)**	**Grados de libertad**	**Promedio de los cuadrados (MS)**	**Estadístico de prueba Fo**
Poder RF	66,870.55	3	$\frac{66{,}870.55}{3} = 22{,}290.18$	$\frac{22{,}290.18}{333.70} = 66.80$
Error	5339.20	16	$\frac{5339.20}{16} = 333.70$	
Total	72,209.75	19		

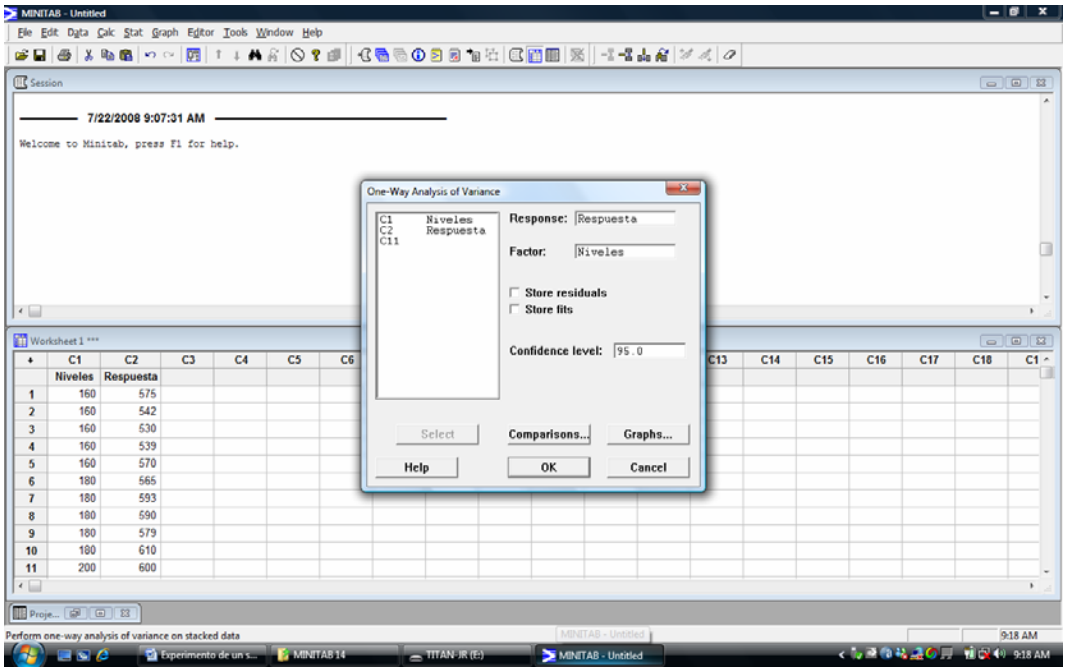
El experimentador obtiene un valor de Fo = 66.80. Tomando un nivel de significancia de 0.05, teniendo 3 grados de libertad del factor y 16 del error, se procede a buscar en la tabla de la distribución F y se obtiene un valor de 3.24. Como $66.80 > 3.24$ entonces se concluye que las medias de los niveles del factor difieren y por tanto se procede a rechazar Ho.

Es importante notar que el procedimiento descrito anteriormente es hecho a mano. Para esto existen programas como Minitab quienes realizan los cálculos a partir de los datos ingresados. A continuación se ilustra el procedimiento en Minitab:

1. En el menú de stat se busca la opción anova, allí se hace doble click en la opción one way anova como muestra la figura



2. Aparece entonces una ventana que permite ingresar las columnas de valores para el análisis. En la primera casilla que dice response, se ingresa la columna que contiene los valores de la respuesta, en la siguiente casilla de factor, se ingresa la columna que tiene los niveles del factor, se dejo una confianza del 95% que equivale al nivel de significancia de 0.05 utilizado en los cálculos manuales:



3. Al dar clik en OK se obtiene la siguiente respuesta:

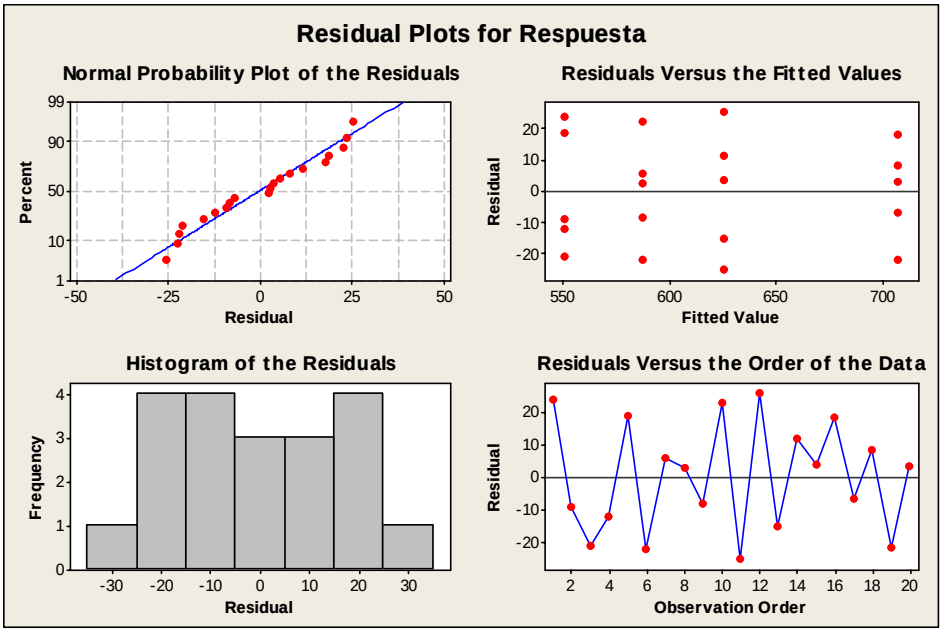
One-way ANOVA: Respuesta versus Niveles

Source	DF	SS	MS	F	P
Niveles	3	66871	22290	66.80	0.000
Error	16	5339	334		
Total	19	72210			

S = 18.27 R-Sq = 92.61% R-Sq(adj) = 91.22%

Se obtienen los mismos valores que se obtuvieron con los cálculos manuales. En este caso se ve que el P-value es de 0, esto implica un valor menor al del nivel de significancia (0.005). al ser $0 < 0.005$ se rechaza Ho y el investigador puede concluir entonces que los niveles del poder afectan la tasa de grabado.

4. Al dar clik en OK se obtiene también una grafica con 4 metodos de análisis graficos para los residuales, esto con el fin de cotejar la idoneidad del modelo:



- Normal probability plot of the residuals (trazo de probabilidad normal): Este grafico muestra que los residuals se encuentran al rededor de la línea del medio, lo cual quiere decir que no hay ninguna desviación significativa de la presunción de normalidad para los residuals.
- Residuals versus the fitted values (trazo de residuales contra los valores estimados): este grafico muestra que no hay un patrón definido.
- Histogram of the residuals (histograma de los residuales): la forma del mismo muestra un comportamiento aproximadamente normal o gaussiano.
- Residuals versus the order of the data (trazo de residuales vs orden de la experimentación): Este grafico muestra que los datos no siguen ningún patrón.

Ejemplo 2

La compañía Mush, productora de setas, ha elaborado un proceso de deshidratación de las mismas. Para el proceso se estableció una caja de cartón equipada con una entrada de aire, una

chimenea, una parrilla para poner las setas a deshidratar y un foco debajo de la misma, el cual provee el calor necesario para deshidratar las setas. El ingeniero encargado del proceso sabe que 150 gramos de setas tardan de 9 a 18 horas en deshidratarse pero no sabe el tiempo exacto. Se sabe también que las setas deben llegar a reducir su peso en un 87% aproximadamente para considerarse deshidratadas. Debido a esto se estableció un experimento tomando un solo factor en consideración (tiempo). El experimentador determino 4 niveles de tiempo entre 9 y 18 horas con intervalos de 3 horas entre cada nivel.

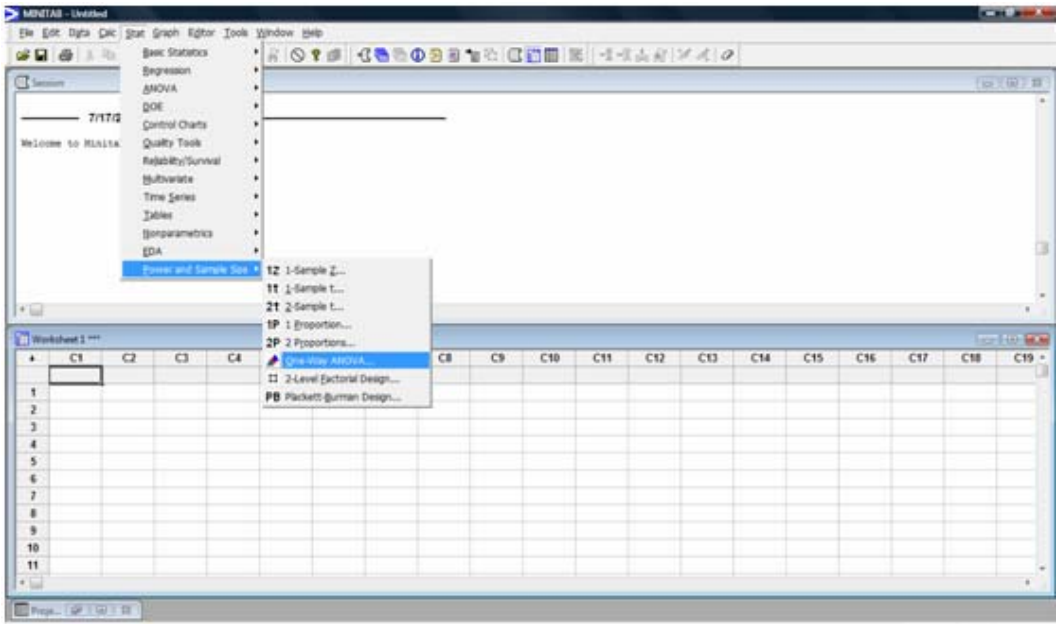
Lo anterior conlleva entonces a la siguiente configuración:

Factor: Tiempo			
Nivel 1: 9 horas	Nivel 2: 12 horas	Nivel 3: 15 horas	Nivel 4: 18 horas
X	X	X	X

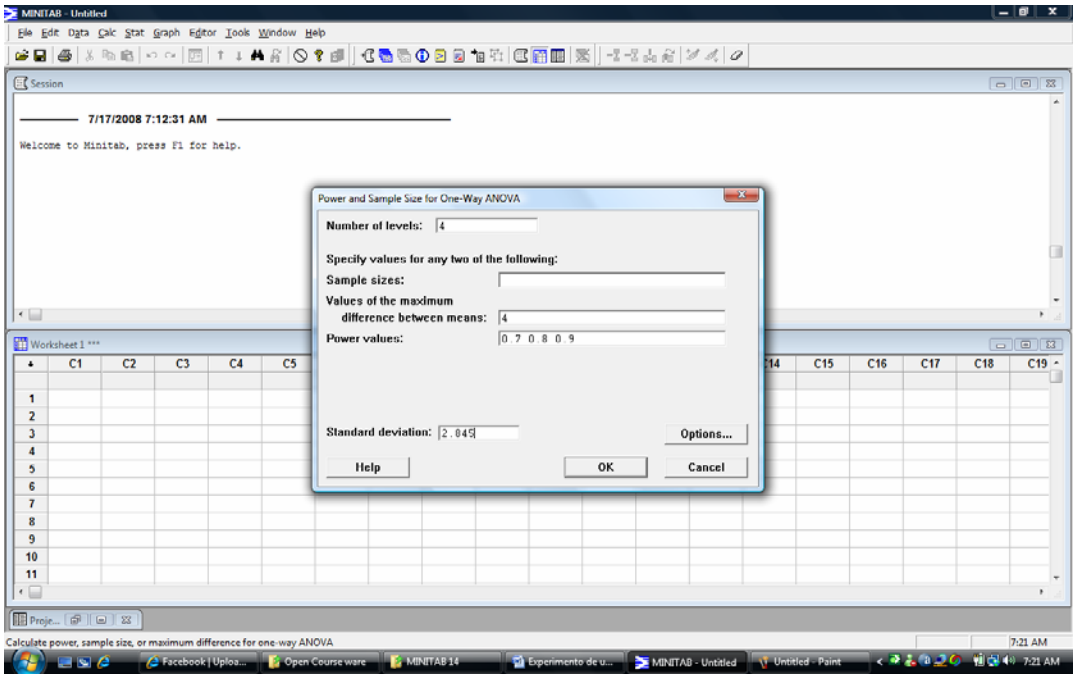
El experimentador sabe que debe realizar replicas de su experimento; para esto el realizó una prueba de poder y tamaño de muestra (power and sample size) en el programa Minitab.

El poder es la probabilidad de que la prueba rechace la hipótesis nula (en este caso es que no exista diferencia entre las medias de los pesos para los niveles de la variable tiempo o que no haya diferencia entre el efecto de los niveles de la variable) cuando la misma es falsa; se denomina como 1-β, siendo β la probabilidad de aceptar algo que debió ser rechazado. Se presumieron 3 valores para el poder (0.7, 0.8 y 0.9) para evaluar la cantidad de replicas de acuerdo a cada uno de ellos. En cuanto a la diferencia entre las medias de los factores, el experimentador hizo una presunción de 4 gramos de manera que se pueda detectar la diferencia entre los efectos de los niveles cuando las medias varíen en más de 4 gramos la una de la otra. El valor de la desviación estándar de los pesos era previamente conocido (2.845 gramos). Los valores del poder, la diferencia entre medias, la desviación estándar y un nivel de significancia de 0.05 fueron ingresados a Minitab de la siguiente manera:

1. En Minitab, en el menú de stat se encuentra la opción de power and sample size y allí la opción de one way anova como muestra la próxima figura:



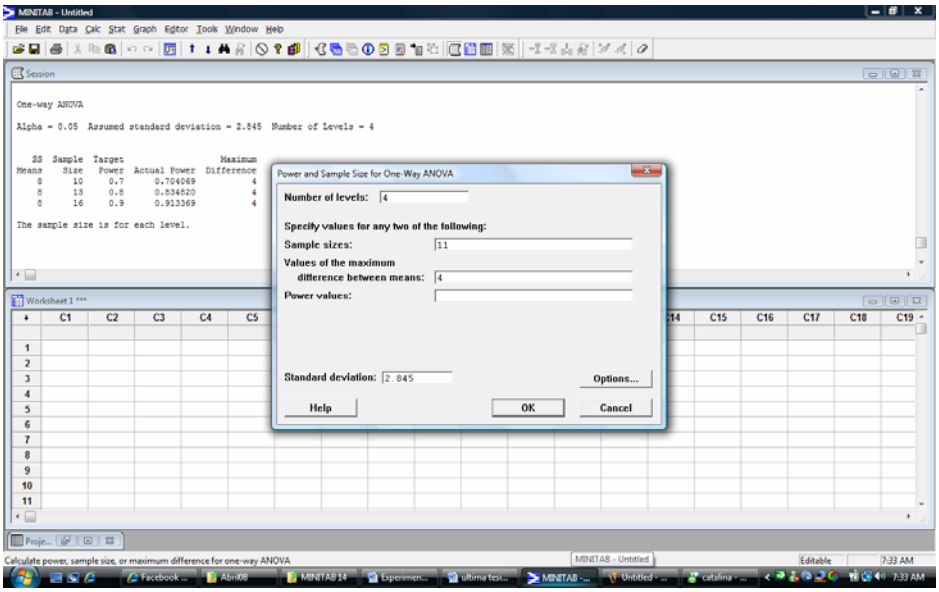
2. Al abrir la opción one way anova, se encuentra entonces la pantalla donde se ingresan los datos del experimento, es decir, el numero de niveles del factor, el valor de la diferencia máxima que se desea entre las medias de los pesos para cada uno de los niveles, los valores del poder y la desviación estándar de los pesos. La siguiente figura ilustra el procedimiento:



3. Al dar click en el botón de OK se obtiene el siguiente resultado:

```
Power and Sample Size

One-way ANOVA

Alpha = 0.05  Assumed standard deviation = 2.845  Number of Levels = 4

   SS   Sample  Target                    Maximum
Means    Size    Power  Actual Power  Difference

    8      10      0.7      0.704069           4

    8      13      0.8      0.834820           4

    8      16      0.9      0.913369           4

The sample size is for each level.
```

El experimentador entonces concluye que para obtener un poder de 0.704069 debe realizar 10 replicas del experimento, para un poder de 0.834820 debe hacer 13 replicas y para un poder de 0.913369 debe hacer 16 replicas. Debido a que el mínimo de replicas es de 10, el experimentador decide entonces buscar el poder que se conseguiría al realizar 11 replicas del experimento. Este procedimiento se hace mediante la misma herramienta de Minitab pero dejando en blanco la casilla de power y poniendo el número 11 en sample size. A continuación se ilustra el procedimiento y la respuesta obtenida:

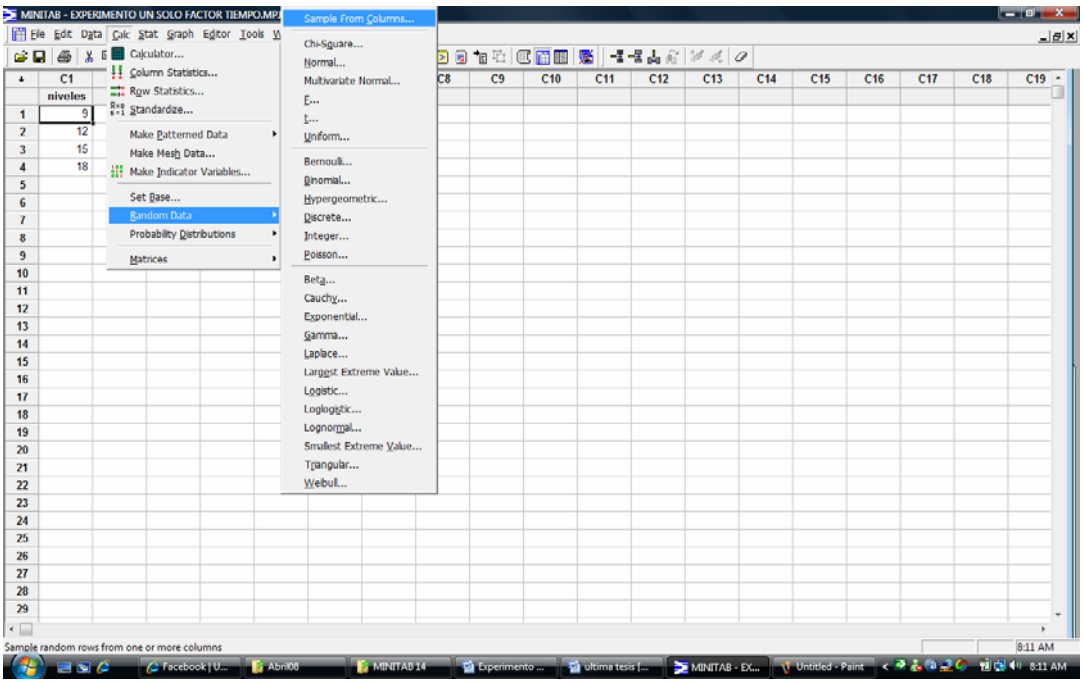


```
Power and Sample Size

One-way ANOVA

Alpha = 0.05  Assumed standard deviation = 2.845  Number of Levels = 4


   SS  Sample                Maximum
Means    Size     Power  Difference
    8      11  0.754440           4


The sample size is for each level.
```

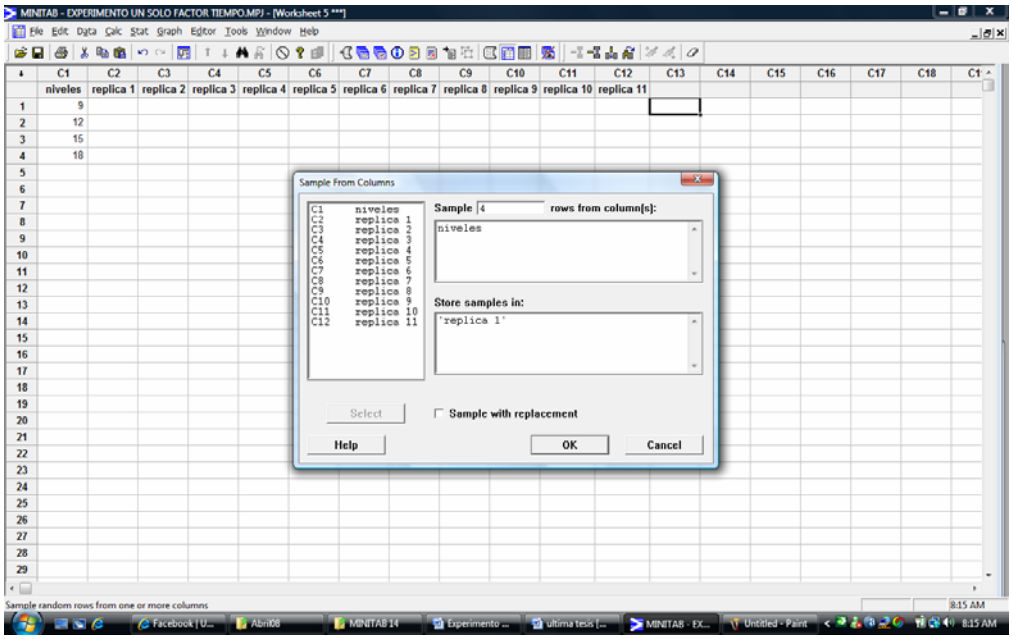
Según el anterior resultado, al realizar 11 replicas se obtiene un poder de 0.7544 que el experimentador considera razonable para los resultados que desea obtener. Por lo anterior el número de replicas que se deben realizar en el experimento de un solo factor aleatorio es de 11.

Después el experimentador hace la aleatoriedad con la que va a realizar la experimentación para cada replica, es decir, en el programa Minitab se ingresan los valores de los niveles (9,12, 15 y 18 horas) y se hace un procedimiento para obtener el orden en que se van a hacer las corridas para cada replica. La siguiente figura ilustra el procedimiento en el programa Minitab:

1. En el menú de calc, en la opción Random data, se despliega otro menú donde se escoge la opción sample from column:



2. Al hacer click en sample from column se despliega una ventana donde se ingresa el numero de filas que contienen los datos a organizar, luego una casilla donde se ingresa la columna de la cual se hace la aleatoriedad, esto haciendo doble click en los nombres de las columnas que se despliegan en la casilla de la izquierda, finalmente en la última casilla se ingresa el nombre de la columna donde se desea que se almacene el resultado (la organización aleatoria de la réplica). La siguiente figura ilustra el procedimiento:



3. Al hacer click en OK se despliega el siguiente resultado:

MINITAB - EXPERIMENTO UN SOLO FACTOR TIEMPO.MPJ - [Worksheet 5 ***]

File Edit Data Calc Stat Graph Editor Tools Window Help

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
	niveles	replica 1	replica 2	replica 3	replica 4	replica 5	replica 6	replica 7	replica 8	replica 9	replica 10	replica 11	
1	9	9											
2	12	15											
3	15	18											
4	18	12											
5													

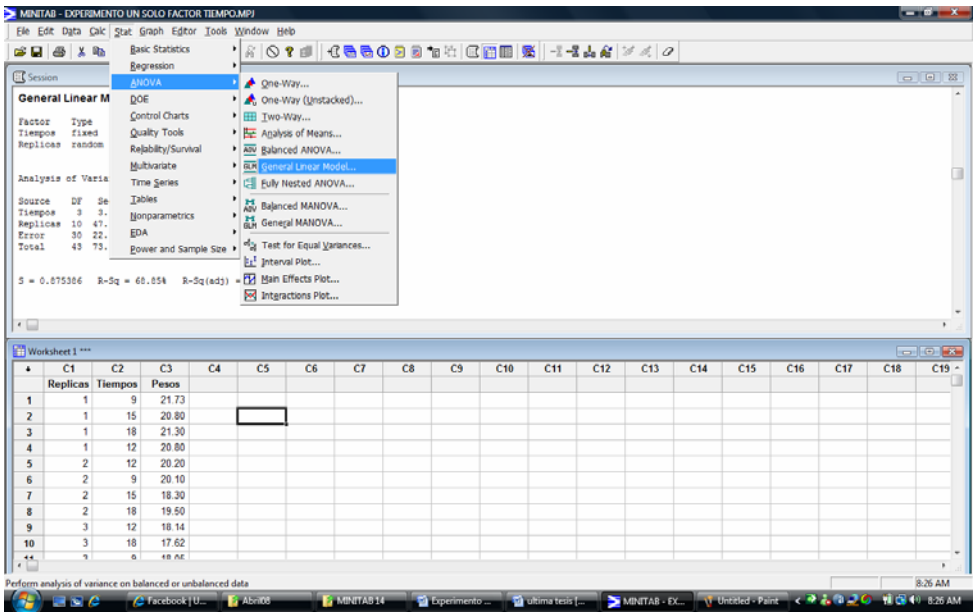
Entonces el experimentador debe correr la primera réplica poniendo las setas en la caja por 9 horas inicialmente, luego debe sacarlas, pesarlas y poner un segundo lote de setas en la caja por 15 horas y así hasta completar la réplica. Para la aleatoriedad de las demás replicas, se repite el procedimiento anteriormente mencionado

Los resultados de los pesos en gramos para las 11 replicas son:

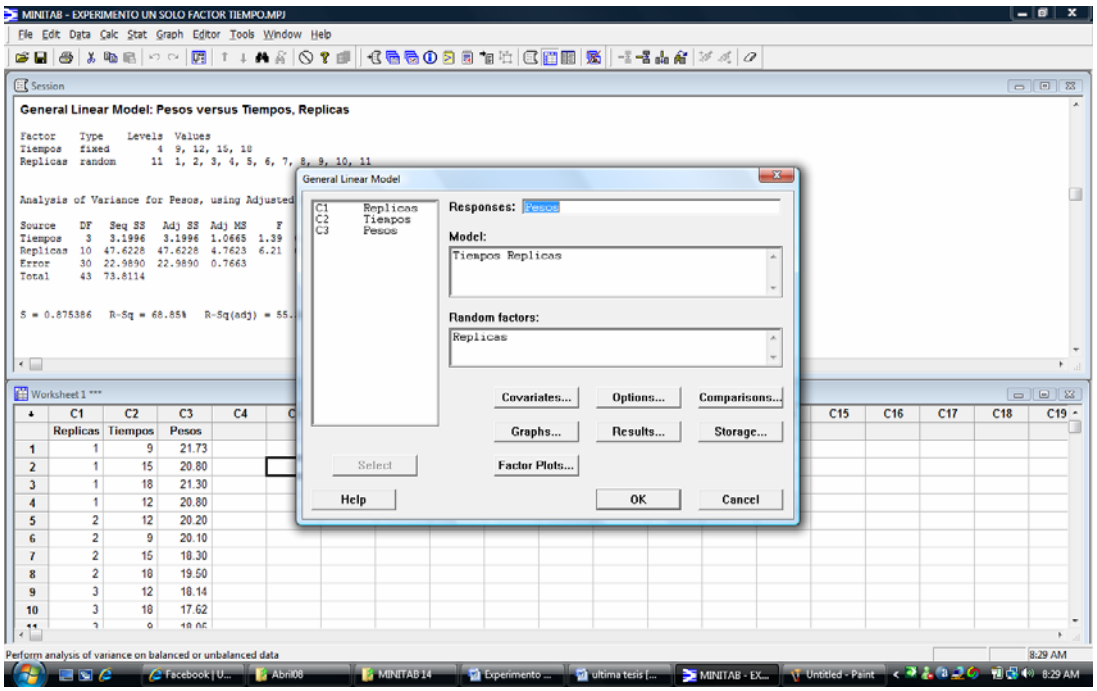
	Factor: Tiempo			
Replica	Nivel 1: 9 horas	Nivel 2: 12 horas	Nivel 3: 15 horas	Nivel 4: 18 horas
1	21.73	20.80	20.80	21.30
2	20.10	20.20	18.30	19.50
3	18.05	18.14	18.40	17.62
4	20.05	19.30	18.85	19.30
5	19.01	19.42	20.27	18.75
6	21.64	21.81	20.06	21.88
7	23.21	20.22	19.04	22.02
8	20.34	18.20	18.74	18.85
9	18.50	18.02	18.30	19.30
10	19.34	20.05	19.53	18.70
11	19.39	18.90	21.43	20.54

El experimentador ingreso los datos a Minitab y realizo el análisis de los mismos de la siguiente manera:

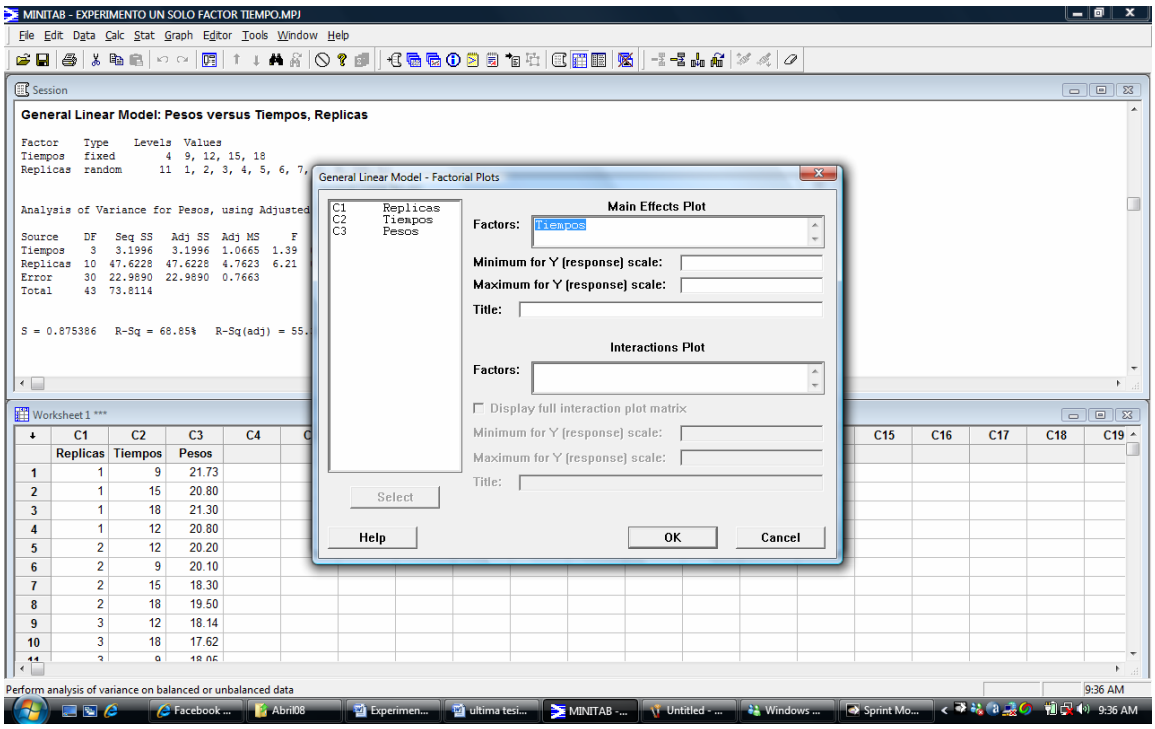
1. En el menú de stat, se despliegan diferentes opciones, debido a que se desea realizar un análisis de varianza, se despliega entonces el menú de ANOVA, donde se escoge la opción de General linear model como muestra la figura:



2. Al dar click en General linear model se obtiene una ventana donde se ingresa en la primera casilla la columna de respuestas denominada como pesos, en la casilla de Model se ingresa el modelo, en este caso el factor tiempo y las replicas, siendo el factor tiempo un factor fijo y las replicas un factor aleatorio. En la última casilla (random factors) se especifica que el factor replica es aleatorio



3. La ventana muestra 7 botones que permiten especificar o adquirir información adicional en el análisis. Para este caso, se oprime el botón factor plots y se obtiene la siguiente ventana:



4. La anterior opción permite realiza un grafico de los efectos principales de los niveles del factor. En la casilla Factors se ingresa entonces el factor tiempo, se oprime OK y regresa a la ventana principal donde se oprime OK de nuevo y se obtiene el siguiente resultado:

General Linear Model: Pesos versus Tiempos, Replicas

Factor	Type	Levels	Values
Tiempos	fixed	4	9, 12, 15, 18
Replicas	random	11	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Analysis of Variance for Pesos, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Tiempos	3	3.1996	3.1996	1.0665	1.39	0.264
Replicas	10	47.6228	47.6228	4.7623	6.21	0.000
Error	30	22.9890	22.9890	0.7663		
Total	43	73.8114				

S = 0.875386 R-Sq = 68.85% R-Sq(adj) = 55.36%

Unusual Observations for Pesos

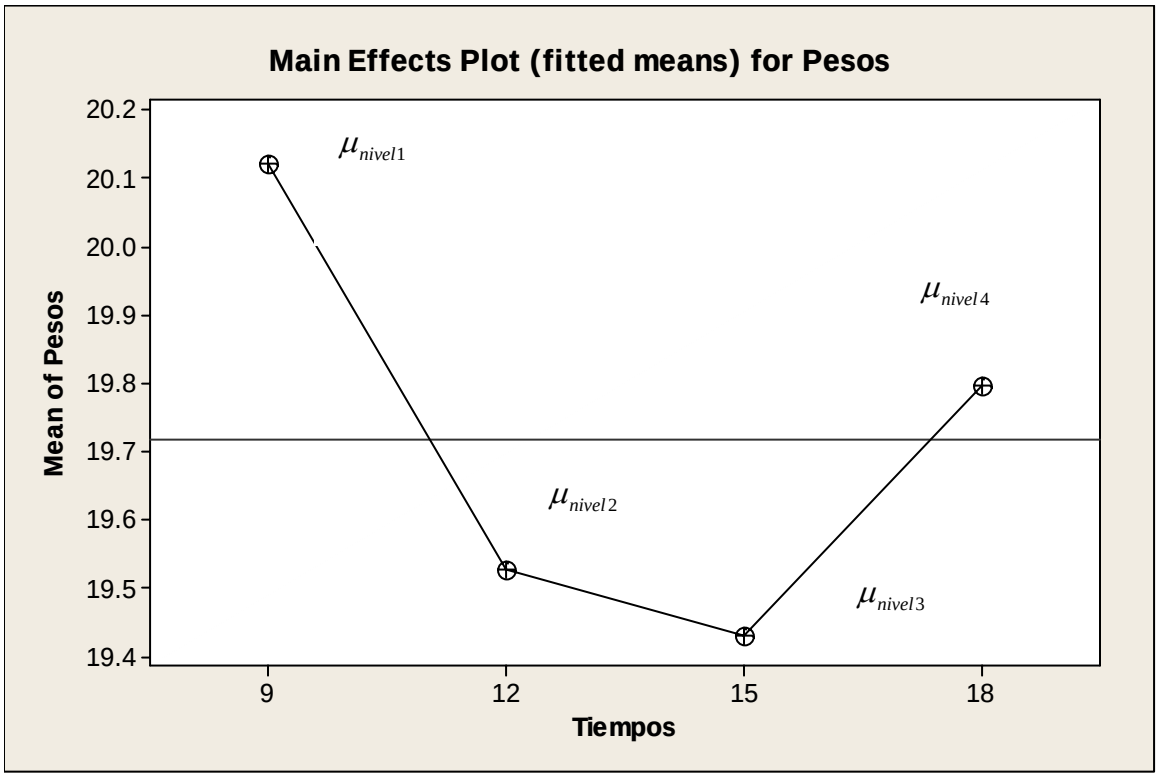
El valor P es mayor al valor de alfa de 0.05 por lo tanto no se puede rechazar Ho y se determina que no hay diferencia entre los niveles del factor.

Obs	Pesos	Fit	SE Fit	Residual	St Resid
25	19.0400	20.8327	0.4938	-1.7927	-2.48 R
26	23.2100	21.5273	0.4938	1.6827	2.33 R
44	21.4300	19.7752	0.4938	1.6548	2.29 R

R denotes an observation with a large standardized residual.

Residual Plots for Pesos

Main Effects Plot (fitted means) for Pesos



El experimentador deduce que no hay diferencia entre los niveles del factor tiempo debido a su valor P. Al observar la grafica se encuentra que la diferencia entre las medias de los niveles no sobrepasan los 4 gramos de diferencia entre las medias que el experimentador quería detectar, por lo tanto, el tiempo que debe durar el proceso de deshidratación es de 9 horas.