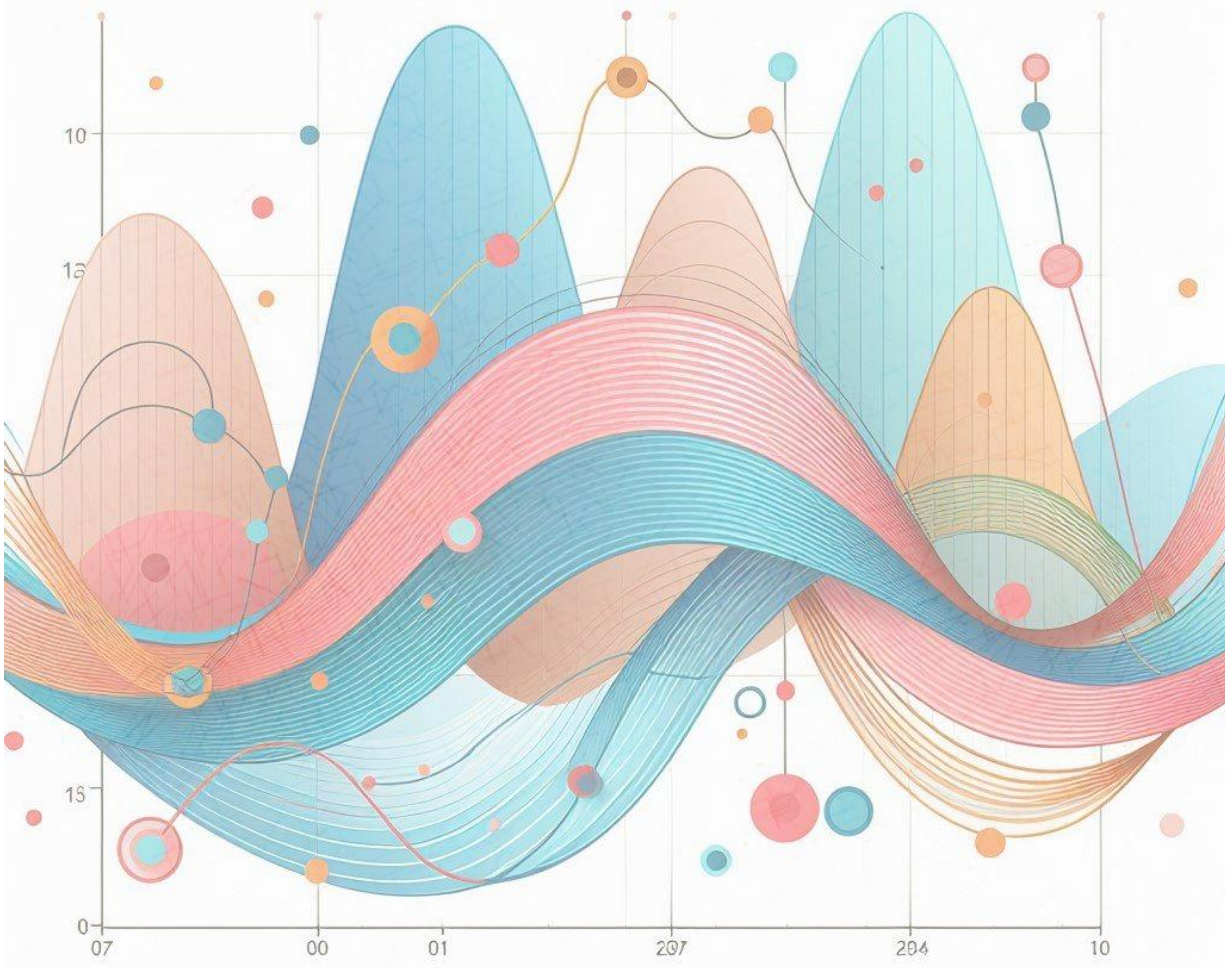
ANOVA - Análisis de varianza

Ignacio Díaz Oreiro

CI0131 Diseño de Experimentos



Agradecimiento a la profesora Dra. Kryscia Ramírez, por facilitar material usado en esta presentación.

Motivación

- Previamente se analizaron métodos para comparar dos condiciones o tratamientos.
- Por ejemplo, el experimento de la tasa de compresión incluyó dos implementaciones diferentes de un algoritmo.
- Otra forma es describir este experimento como un experimento con un solo factor con dos niveles, donde el factor es el algoritmo de compresión y los dos niveles son los dos algoritmos diferentes para hacer la compresión.

Motivación

- Sin embargo, muchos experimentos involucran más de dos niveles del factor.
- Por ejemplo, se desea investigar el rendimiento de un nuevo algoritmo de compresión que se usará en sistemas de almacenamiento.
- Se sabe por experiencia que el rendimiento del algoritmo se ve afectado por el tamaño de los bloques de datos utilizados durante la compresión.

Análisis de Varianza

Motivación

- Se sospecha también que al aumentar el tamaño de los bloques (en KB), se incrementará la eficiencia de la compresión, pero hasta cierto punto, después del cual el rendimiento puede disminuir debido a limitaciones de memoria.
- Además, se sabe que el tamaño de los bloques debe variar entre 10 KB y 40 KB para asegurar que el algoritmo funcione adecuadamente en sistemas con limitaciones de recursos (la memoria RAM).
- Se decide probar el algoritmo en cinco niveles de tamaño de bloques: 15 KB, 20 KB, 25 KB, 30 KB y 35 KB.

Análisis de Varianza

Motivación

- Se trata entonces de un experimento con un solo factor con 5 niveles del factor.
- Si quieren revisar las diferencias entre las tasas de compresión promedio con todos los 5 niveles, el interés será probar la igualdad de las cinco medias ($\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$).
- Pudiera parecer que este problema se resolvería realizando una prueba t para todos los pares de medias posibles.
- Sin embargo, no es ésta la mejor solución, porque llevaría a una distorsión considerable en el error tipo I.

Análisis de Varianza

Motivación

- Para probar la igualdad de las cinco medias usando comparaciones por pares (pruebas *t*) habría 10 pares posibles.
- Si la probabilidad de aceptar correctamente H_0 la hipótesis nula en cada prueba individual es de $1 - \alpha = 0.95$, la probabilidad de aceptar correctamente H_0 en las 10 pruebas es de $(0.95)^{10} = 0.60$ si las pruebas son independientes.
- Por lo tanto, ha ocurrido un incremento sustancial en el error tipo I.

Motivación

- El procedimiento correcto para probar la igualdad de varias medias es el análisis de varianza.
- En el caso anterior, el análisis de varianzas permitirá mantener el error de tipo I en 0.05.
- Además, el análisis de varianza tiene un rango de aplicaciones más amplio que el problema anterior.
- Probablemente sea la técnica más útil en el campo de la inferencia estadística.

Análisis de Varianza - ANOVA

Idea intuitiva

- La hipótesis nula de la que parten los diferentes tipos de ANOVA es que la media de la variable estudiada es la misma en los diferentes grupos ($\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$)
- En contraposición, la hipótesis alternativa establece que al menos dos medias difieren de forma significativa ($\mu_i \neq \mu_j$).
- ANOVA permite comparar múltiples medias, y lo hace mediante el estudio de las varianzas.

Análisis de Varianza - ANOVA

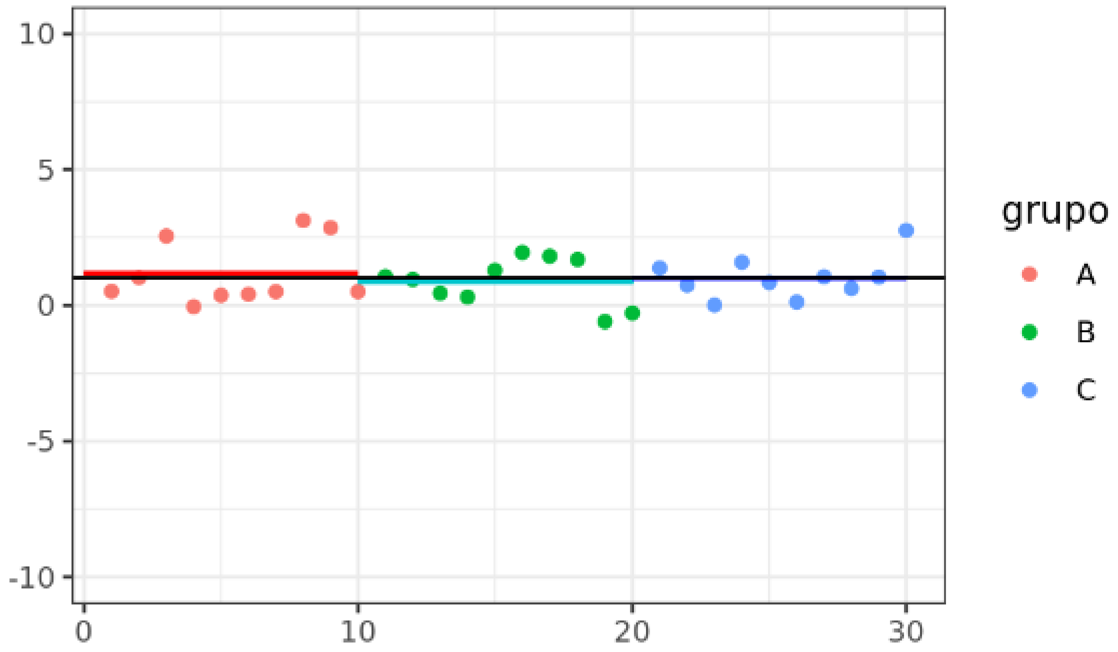
Idea intuitiva

- El funcionamiento básico de un ANOVA consiste en calcular la media de cada uno de los grupos para luego comparar la varianza de estas medias (varianza explicada por la variable grupo, inter-varianza) frente a la varianza promedio dentro de los grupos (la no explicada por la variable grupo, intra-varianza).

Análisis de Varianza - ANOVA

Idea intuitiva

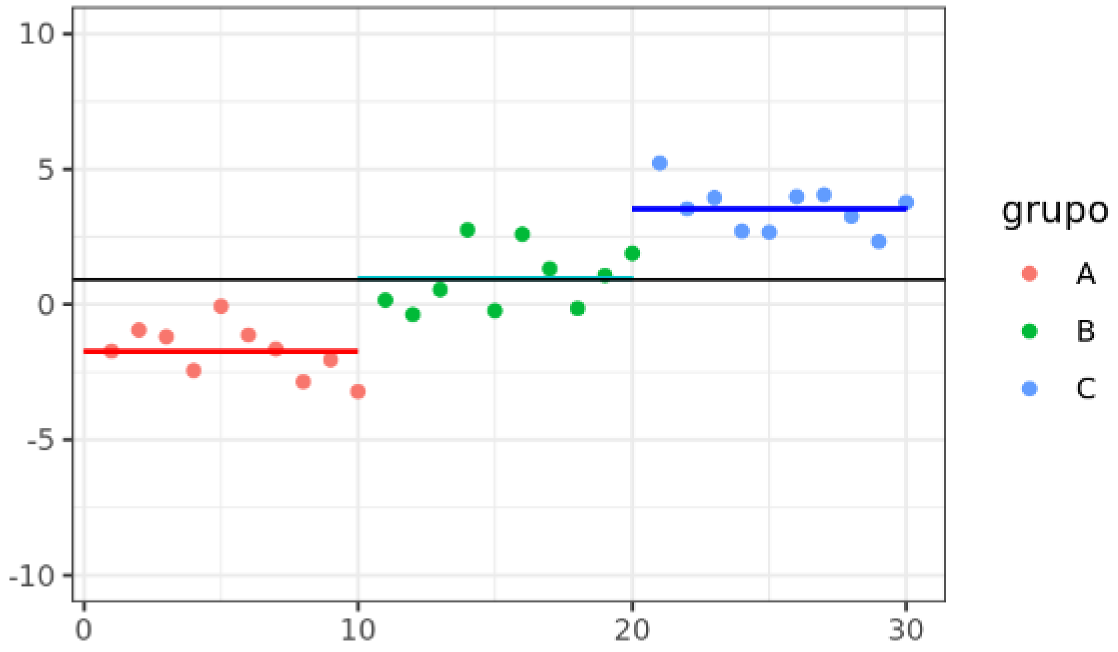
- Bajo la hipótesis nula de que las observaciones de los distintos grupos proceden todas de la misma población (tienen la misma media y varianza), la varianza ponderada entre grupos será la misma que la varianza promedio dentro de los grupos.



Análisis de Varianza - ANOVA

Idea intuitiva

- Conforme las medias de los grupos estén más alejadas las unas de las otras, la varianza entre medias se incrementará y dejará de ser igual a la varianza promedio dentro de los grupos.



Medias muestrales y la gran media

- En el cálculo de ANOVA hay dos tipos de medias:
- Medias muestrales separadas de cada grupo.
- La gran media, que es la media de las medias muestrales o la media de todas las observaciones combinadas, independientemente de la muestra.

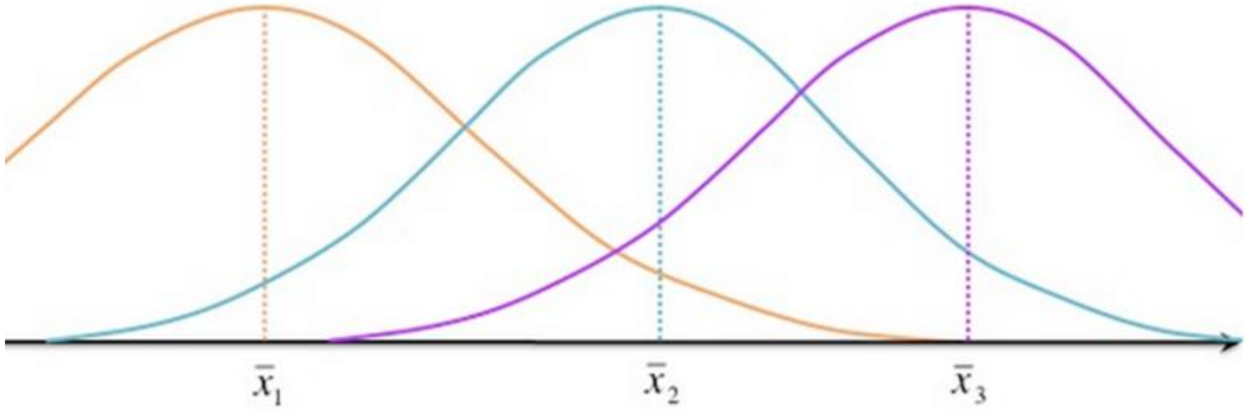
Análisis de Varianza - ANOVA

Variabilidad entre grupos

- Considere las siguientes distribuciones de muestras:

- Como las muestras se superponen, sus medias individuales no difieren demasiado.

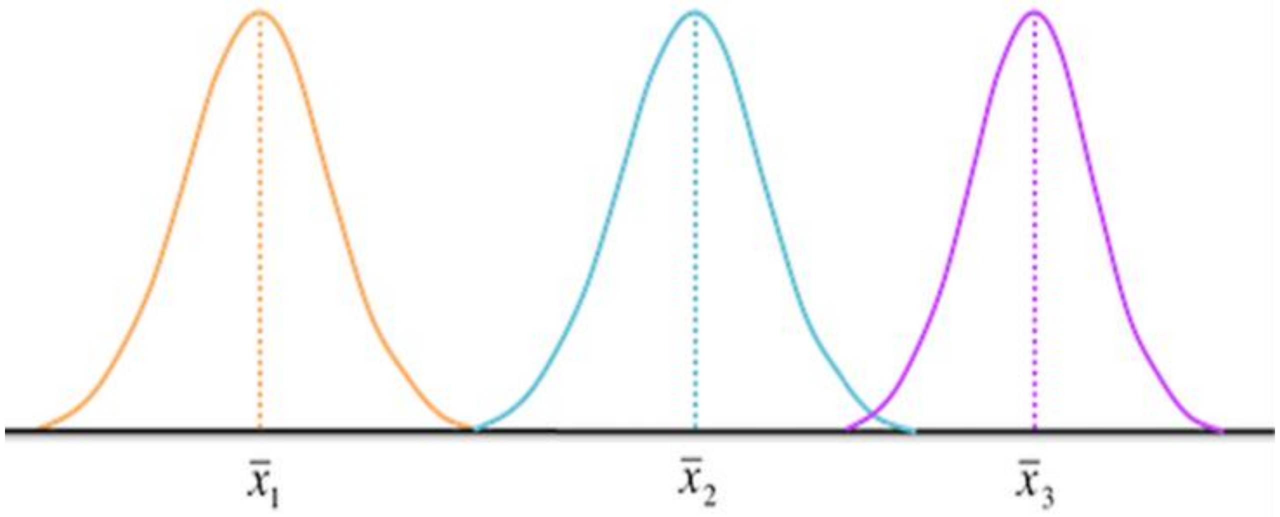
- La diferencia entre sus medias individuales y su gran media no será suficientemente significativa.



Análisis de Varianza - ANOVA

Variabilidad entre grupos

- Ahora considere estas distribuciones de muestras:

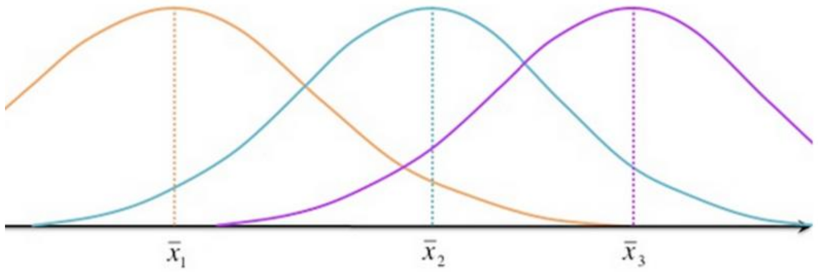


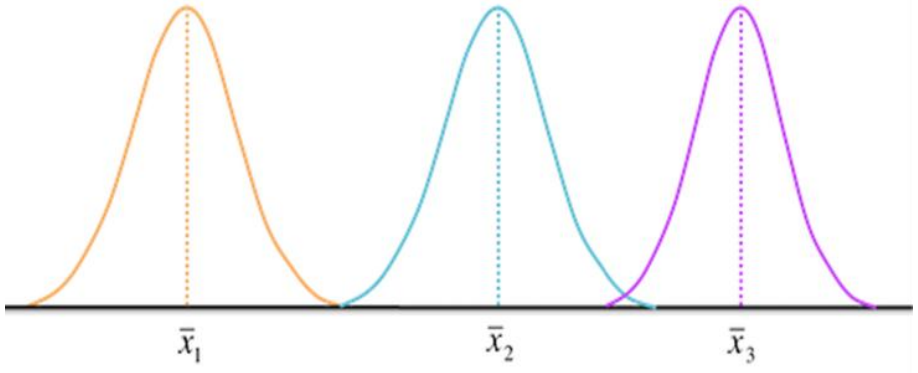
- Las muestras difieren entre sí por un gran margen, así como sus medias individuales.
- La diferencia entre las medias individuales y la gran media también sería significativa.

Análisis de Varianza - ANOVA

Variabilidad entre grupos

- Si hay superposición o cercanía, la gran media será similar a las medias individuales.
- Si las distribuciones están muy separadas, la diferencia entre las medias y la gran media será grande.

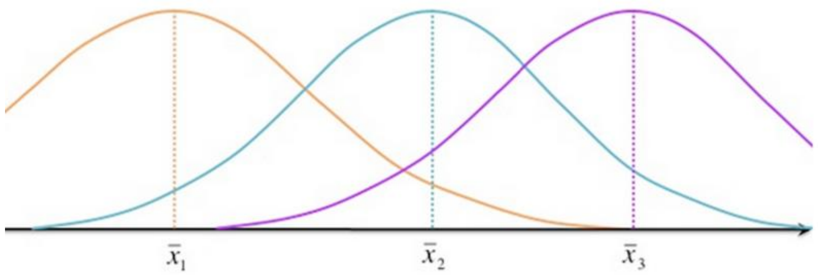


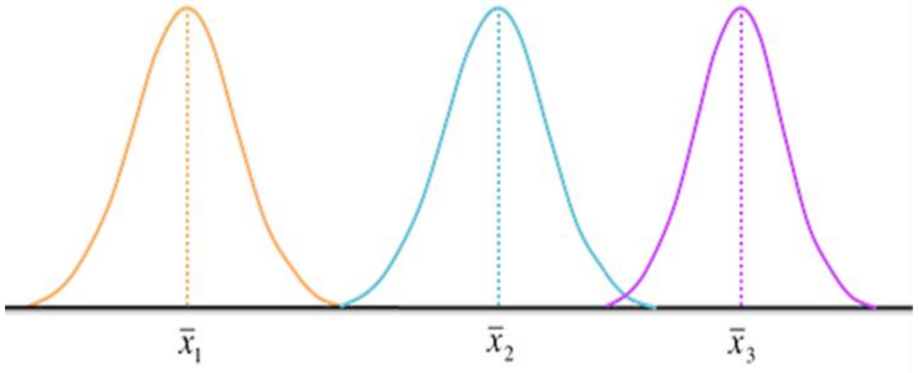


Análisis de Varianza - ANOVA

Variabilidad

- Hay dos fuentes diferentes de variabilidad que mide un ANOVA:
- Variabilidad entre grupos: la variación total entre la media de cada grupo y la gran media.
- Variabilidad dentro del grupo: la variación total en los valores individuales en cada grupo y su media grupal.

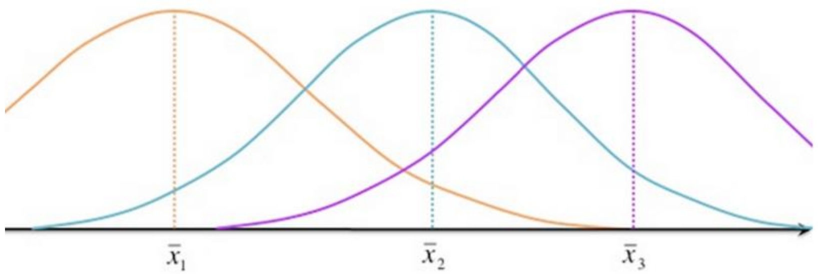


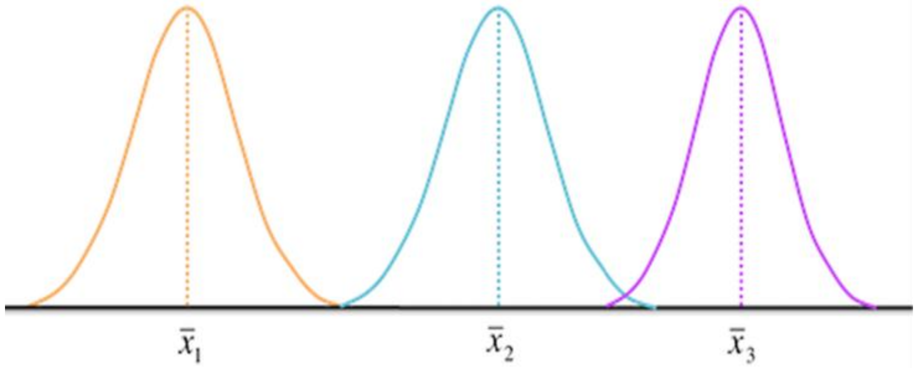


Análisis de Varianza - ANOVA

Variabilidad

- Si la variación entre grupos es alta en relación con la variación dentro del grupo, entonces el estadístico *F* del ANOVA será más alto y el valor *p* correspondiente será más bajo.
- Entonces será más probable que ser rechace la hipótesis nula de que las medias de los grupos son iguales.





ANOVA de una vía

- El ANOVA de una vía (*one-way* ANOVA) o unidireccional compara las medias de tres o más grupos independientes para determinar si existe una diferencia estadísticamente significativa entre las medias poblacionales correspondientes.

Análisis de Varianza - ANOVA

ANOVA de una vía

- Por ejemplo, supongamos que se quiere saber si tres programas diferentes de preparación para exámenes conducen o no a puntajes medios diferentes en un examen de ingreso a la universidad.
- Dado que hay miles de estudiantes de secundaria en todo el país, sería demasiado lento y costoso visitar a cada estudiante y que utilicen uno de los programas de preparación para exámenes.
- En lugar de esto, se reclutan 30 estudiantes y se asignan aleatoriamente 10 a cada método de estudio diferente.

Análisis de Varianza - ANOVA

ANOVA de una vía

- La variable de respuesta será el puntaje obtenido en la prueba de admisión.
- La hipótesis nula del experimento es que la media de la variable de respuesta es la misma en los 3 diferentes grupos, $H_0 : \mu_1 = \mu_2 = \mu_3$
- En contraposición, la hipótesis alternativa establece que al menos dos medias difieren de forma significativa. Es decir, $H_A : \mu_i \neq \mu_j$ para algún i ≠ j con i, j = {1, 2, 3}

ANOVA de una vía

Method 1	Method 2	Method 3
75	78	82
77	78	82
78	79	84
78	81	86
79	81	86
81	82	87
81	83	87
83	85	89
86	86	90
87	88	94

- Los puntajes de los exámenes de los estudiantes en cada grupo son:

	Method 1	Method 2	Method 3
	75	78	82
	77	78	82
	78	79	84
	78	81	86
	79	81	86
	81	82	87
	81	83	87
	83	85	89
	86	86	90
	87	88	94
Group Mean	80.5	82.1	86.7
Overall Mean	83.1		

- Luego se pueden calcular las medias de cada grupo y la gran media:

Análisis de Varianza - ANOVA

ANOVA de una vía

- El nombre de análisis de varianza se deriva de una partición de la variabilidad total en sus partes componentes.
- La suma total de cuadrados se utiliza como medida de la variabilidad general de los datos.

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

ANOVA de una vía

- Recordemos la notación punto (k tratamientos y tamaño de muestra n):

Treatment:	**1**	**2**	$\cdots$	$\boldsymbol{i}$	$\cdots$	$\boldsymbol{k}$	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	$\cdots$	$Y_{i.}$	$\cdots$	$Y_{k.}$	$Y_{..}$
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\cdots$	$\bar{y}_{i.}$	$\cdots$	$\bar{y}_{k.}$	$\bar{y}_{..}$

Análisis de Varianza - ANOVA

ANOVA de una vía

- Se puede demostrar la siguiente identidad de la suma de cuadrados:

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

- Donde $n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2$ representa las diferencias entre los promedios de tratamiento y el promedio general (variabilidad entre grupos)
- Y $\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$ las diferencias de observaciones dentro de los tratamientos con el promedio del tratamiento (variabilidad dentro del grupo).

Análisis de Varianza - ANOVA

ANOVA de una vía

- Se calcula la variabilidad entre grupos:

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

- Luego, la variación entre grupos es:

$10(80.5 - 83.1)^2 + 10(82.1 - 83.1)^2 + 10(86.7 - 83.1)^2$

$= 207.2$

	Method 1	Method 2	Method 3
	75	78	82
	77	78	82
	78	79	84
	78	81	86
	79	81	86
	81	82	87
	81	83	87
	83	85	89
	86	86	90
	87	88	94
Group Mean	80.5	82.1	86.7
Overall Mean	83.1		

Análisis de Varianza - ANOVA

ANOVA de una vía

- Luego la variabilidad dentro del grupo: $\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$

Group 1: $(75-80.5)^2 + (77-80.5)^2 + (78-80.5)^2 + (78-80.5)^2 + (79-80.5)^2 + (81-80.5)^2 + (81-80.5)^2 + (83-80.5)^2 + (86-80.5)^2 + (87-80.5)^2 =$ **136.5**

Group 2: $(78-82.1)^2 + (78-82.1)^2 + (79-82.1)^2 + (81-82.1)^2 + (81-82.1)^2 + (82-82.1)^2 + (83-82.1)^2 + (85-82.1)^2 + (86-82.1)^2 + (88-82.1)^2 =$ **104.9**

Group 3: $(82-86.7)^2 + (82-86.7)^2 + (84-86.7)^2 + (86-86.7)^2 + (86-86.7)^2 + (87-86.7)^2 + (87-86.7)^2 + (89-86.7)^2 + (90-86.7)^2 + (94-86.7)^2 =$ **122.1**

	Method 1	Method 2	Method 3
	75	78	82
	77	78	82
	78	79	84
	78	81	86
	79	81	86
	81	82	87
	81	83	87
	83	85	89
	86	86	90
	87	88	94
Group Mean	80.5	82.1	86.7
Overall Mean	83.1		

- Finalmente es: 136.5 + 104.9 + 122.1 = 363.5

ANOVA de una vía

- Si se usa software estadístico para realizar un ANOVA unidireccional (o de una vía) usando este conjunto de datos, se obtiene la siguiente tabla:

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	**207.2**	2	103.6	7.6952	0.0023
Within Groups	**363.5**	27	13.4630		
Total	570.7	29			

- Hasta ahora hemos calculado los valores *SS*, o suma de cuadrados (*sum of squares*) para la variabilidad entre grupos y para la variabilidad dentro del grupo.

Análisis de Varianza - ANOVA

ANOVA de una vía

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	**207.2**	2	103.6	7.6952	0.0023
Within Groups	**363.5**	27	13.4630		
Total	570.7	29			

- El siguiente paso es calcular los grados de libertad (*df*).
- En el caso de la variabilidad entre grupos, df_B es el número de grupos (o medias muestrales) menos 1 = $k - 1$
- Para la variabilidad dentro del grupo, df_W es el número total de observaciones N = $n_1 + n_2 + n_3 + \ldots + n_k$ menos el número total de grupos $k = N - k$
- Otra forma de ver este último es $(n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1) = N - k$

Análisis de Varianza - ANOVA

ANOVA de una vía

ANOVA

Source of Variation	SS	df	MS	F	P-value
Between Groups	**207.2**	2	103.6	7.6952	0.0023
Within Groups	**363.5**	27	13.4630		
Total	570.7	29			

- El cociente SS_B / df_B para la variabilidad entre grupos se conoce como el cuadrado medio de la variabilidad entre grupos (MS_B).
- De forma análoga, el cociente SS_W / df_W para la variabilidad dentro del grupo se conoce como el cuadrado medio de la variabilidad dentro del grupo (MS_W).
- Luego, F es el estadístico que corresponde al cociente MS_B / MS_W que sigue la función de distribución conocida como F de Fisher-Snedecor.
- El valor p corresponde a F para los grados de libertad df_B, df_W

Análisis de Varianza - ANOVA

ANOVA de una vía

ANOVA

Source of Variation	SS	df	MS	F	P-value
Between Groups	207.2	2	103.6	7.6952	0.0023
Within Groups	363.5	27	13.4630		
Total	570.7	29			

- El estadístico $F = MS_B / MS_W$ se compara con el valor teórico asociado a los grados de libertad y al **α** seleccionado con anticipación (por ejemplo 0.05).
- En este caso, el valor estadístico observado 7.6952 es mayor que el valor teórico 3.3541, por lo que la hipótesis nula se rechaza.
- Entonces, se puede concluir que los 3 métodos de estudio no conducen a resultados con medias similares.

	DF1	α = 0.05	
DF2	1	2	3
1	161.45	199.5	215.71
2	18.513	19	19.164
3	10.128	9.5521	9.2766
...			
25	4.2417	3.3852	2.9912
26	4.2252	3.369	2.9752
27	4.21	3.3541	2.9604
28	4.196	3.3404	2.9467

Análisis de Varianza - ANOVA

ANOVA de una vía

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	**207.2**	2	103.6	7.6952	0.0023
Within Groups	**363.5**	27	13.4630		
Total	570.7	29			

- Recordemos que el cociente $F = MS_B \,/\, MS_W$ cuantifica la proporción de la variabilidad entre grupos en comparación con la variabilidad dentro del grupo.
- Cuanto mayor sea el estadístico *F*, mayor será la variabilidad entre las medias de los grupos en relación con la variabilidad dentro de los grupos.
- Por lo tanto, cuanto mayor sea el estadístico *F*, mayor será la evidencia de que existe una diferencia entre las medias de los grupos.

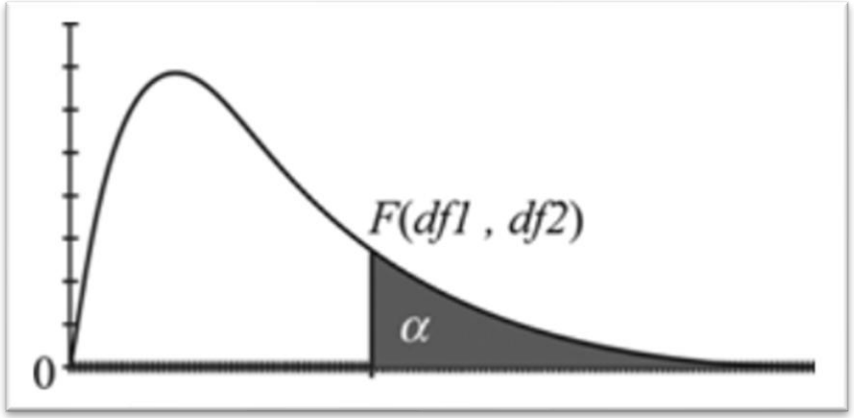
Análisis de Varianza - ANOVA

ANOVA de una vía

ANOVA

Source of Variation	SS	df	MS	F	P-value
Between Groups	**207.2**	2	103.6	7.6952	0.0023
Within Groups	**363.5**	27	13.4630		
Total	570.7	29			

- Una forma equivalente es analizar el valor-p.
- Se puede ver en este ejemplo que el valor-p que corresponde al estadístico *F* de 7.6952 es 0.0023.
- Dado que este valor es menor que α = 0.05, se rechaza la hipótesis nula del ANOVA y se concluye que las tres técnicas de estudio no conducen a la misma calificación en el examen.



ANOVA de una vía - Supuestos

- Para que los resultados sean válidos, se deben cumplir los siguientes supuestos:
- Normalidad: Los residuales (no las observaciones) se distribuyen normalmente dentro de cada grupo.
- Varianzas iguales: La varianza de los residuales es constante entre todos los grupos (homocedasticidad).
- Independencia: Las observaciones (y por ende los residuales) son independientes entre sí (no correlacionadas en el tiempo o entre sujetos).

Análisis de Varianza - ANOVA

Prueba de comparación múltiple

- Si con los resultados del ANOVA de una vía se puede rechazar la hipótesis nula, se puede concluir que existe suficiente evidencia para decir que al menos una de las medias de los grupos es diferente de las demás.
- Sin embargo, esto no nos dice qué grupos son diferentes entre sí. Simplemente nos dice que no todas las medias de los grupos son iguales.
- Para determinar qué grupos son diferentes entre sí, se debe realizar una prueba de comparación múltiple (o *post hoc*), que permitirá explorar la diferencia entre las medias de varios grupos y al mismo tiempo controlar el error tipo I.

Prueba de comparación múltiple

- Es importante tener en cuenta que solo se necesita realizar una prueba de comparación múltiple cuando el valor-p para el ANOVA es estadísticamente significativo.
- Si el valor-p no es estadísticamente significativo, esto indica que no hay evidencia para rechazar la hipótesis de que las medias de todos los grupos son iguales entre sí, por lo que no tiene sentido realizar una prueba post hoc para averiguar qué grupos son diferentes entre sí.

Análisis de Varianza - ANOVA

Prueba de comparación múltiple

- Algunas pruebas de comparación múltiple:
- Prueba de Tukey: Realiza comparaciones por pares entre todas las combinaciones posibles de medias de los grupos. Es potente cuando se cumplen los supuestos. Si no se cumplen o se tienen pocos datos es susceptible de no detectar diferencias significativas.
- Método de Holm: Realiza las pruebas en un orden específico y ajusta los valores p de acuerdo con su posición en el orden. Es útil cuando se tienen muchas comparaciones o cuando se sospecha que los supuestos no se cumplen. El orden en el que se realizan las pruebas puede afectar los resultados.

Análisis de Varianza - ANOVA

Prueba de Tukey

- Por ejemplo, supongamos que tenemos cuatro grupos que queremos comparar sus medias.
- Los grupos son: A, B, C y D y cada grupo tuvo una muestra de 20.

```
#view summary of anova model
summary(anova_model)

#             Df Sum Sq Mean Sq F value   Pr(>F)
#group         3  25.37   8.458   17.66 8.53e-09 ***
#Residuals    76  36.39   0.479
```

Análisis de Varianza - ANOVA

Prueba de Tukey

```
#view summary of anova model
summary(anova_model)

#            Df Sum Sq Mean Sq F value   Pr(>F)
#group        3  25.37   8.458   17.66 8.53e-09 ***
#Residuals   76  36.39   0.479
```

- De la tabla ANOVA, vemos que el estadístico *F* es 17.66 y el valor-p es extremadamente pequeño.
- Se tiene evidencia suficiente para rechazar la hipótesis nula de que todas las medias de los grupos son iguales.
- A continuación, se puede usar una prueba de comparación múltiple.

Análisis de Varianza - ANOVA

Prueba de Tukey

- Al ejecutar la prueba TukeyHSD se obtiene:
- Se debe tener en cuenta que se especificó que el nivel de confianza es del 95%, lo que significa que se quiere que la tasa de error sea de 0.05.
- R brinda dos métricas para comparar cada diferencia por pares:

```
#perform Tukey's Test for multiple comparisons
TukeyHSD(anova_model, conf.level=.95)

#  Tukey multiple comparisons of means
#    95% family-wise confidence level
#
#Fit: aov(formula = amount ~ group, data = data_long)
#
#$group
#          diff          lwr       upr     p adj
#B-A 0.2822630 -0.292540425 0.8570664 0.5721402
#C-A 0.8561388  0.281335427 1.4309423 0.0011117
#D-A 1.4676027  0.892799258 2.0424061 0.0000000
#C-B 0.5738759 -0.000927561 1.1486793 0.0505270
#D-B 1.1853397  0.610536271 1.7601431 0.0000041
#D-C 0.6114638  0.036660419 1.1862672 0.0326371
```

Análisis de Varianza - ANOVA

Prueba de Tukey

- Un intervalo de confianza para la diferencia de medias (dado por los valores de *lwr* y *upr*)
- El valor-p ajustado (con corrección) para la diferencia de medias.
- Tanto el intervalo de confianza como el valor-p conducirán a la misma conclusión.

```
#perform Tukey's Test for multiple comparisons
TukeyHSD(anova_model, conf.level=.95)

#  Tukey multiple comparisons of means
#    95% family-wise confidence level
#
#Fit: aov(formula = amount ~ group, data = data_long)
#
#$group
#          diff          lwr       upr      p adj
#B-A 0.2822630 -0.292540425 0.8570664 0.5721402
#C-A 0.8561388  0.281335427 1.4309423 0.0011117
#D-A 1.4676027  0.892799258 2.0424061 0.0000000
#C-B 0.5738759 -0.000927561 1.1486793 0.0505270
#D-B 1.1853397  0.610536271 1.7601431 0.0000041
#D-C 0.6114638  0.036660419 1.1862672 0.0326371
```

Análisis de Varianza - ANOVA

Prueba de Tukey

- Por ejemplo, el intervalo de confianza para la diferencia de medias entre el grupo C y el grupo A es (0.2813, 1.4309).
- Dado que este intervalo no contiene a 0, se sabe que la diferencia entre las medias de estos dos grupos es estadísticamente significativa.

```
#perform Tukey's Test for multiple comparisons
TukeyHSD(anova_model, conf.level=.95)

#  Tukey multiple comparisons of means
#    95% family-wise confidence level
#
#Fit: aov(formula = amount ~ group, data = data_long)
#
#$group
#          diff          lwr       upr     p adj
#B-A 0.2822630 -0.292540425 0.8570664 0.5721402
#C-A 0.8561388  0.281335427 1.4309423 0.0011117
#D-A 1.4676027  0.892799258 2.0424061 0.0000000
#C-B 0.5738759 -0.000927561 1.1486793 0.0505270
#D-B 1.1853397  0.610536271 1.7601431 0.0000041
#D-C 0.6114638  0.036660419 1.1862672 0.0326371
```

Análisis de Varianza - ANOVA

Prueba de Tukey

- Asimismo, el valor-p para la diferencia de medias entre el grupo C y el grupo A es 0.0011, que es menor que el nivel de significancia definido de 0.05.
- Esto también indica que la diferencia entre las medias de estos dos grupos es estadísticamente significativa.

```
#perform Tukey's Test for multiple comparisons
TukeyHSD(anova_model, conf.level=.95)

#  Tukey multiple comparisons of means
#    95% family-wise confidence level
#
#Fit: aov(formula = amount ~ group, data = data_long)
#
#$group
#          diff          lwr       upr     p adj
#B-A 0.2822630 -0.292540425 0.8570664 0.5721402
#C-A 0.8561388  0.281335427 1.4309423 0.0011117
#D-A 1.4676027  0.892799258 2.0424061 0.0000000
#C-B 0.5738759 -0.000927561 1.1486793 0.0505270
#D-B 1.1853397  0.610536271 1.7601431 0.0000041
#D-C 0.6114638  0.036660419 1.1862672 0.0326371
```

Análisis de Varianza - ANOVA

Prueba de Tukey

- Existen varios ajustes del valor-p propuestos. La corrección de Bonferroni es uno de los más conocidos.
- Si *m* es el número de hipótesis nulas (familia de hipótesis $\mu_1 = \mu_2$, $\mu_1 = \mu_2$, ..., $\mu_{m-1} = \mu_m$) y p_1 , p_2 , ... p_m sus valores p correspondientes, se rechaza cada $hipótesis_i$ con $p_i \leq \alpha / m$
- En R comparamos con 0.05, dado que el valor fue ajustado (no lo hacemos nosotros)

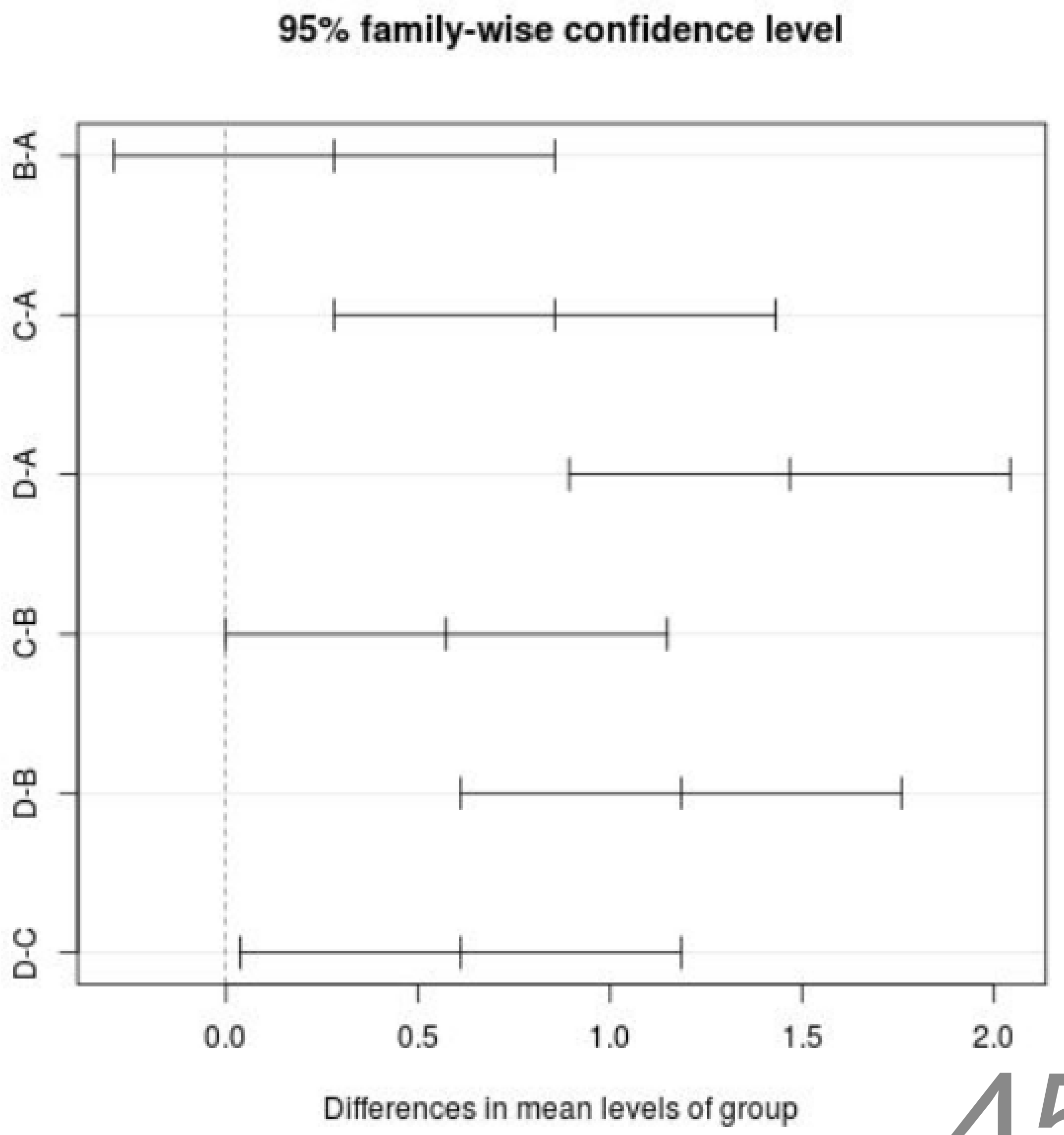
```
#perform Tukey's Test for multiple comparisons
TukeyHSD(anova_model, conf.level=.95)


#  Tukey multiple comparisons of means
#    95% family-wise confidence level
#
#Fit: aov(formula = amount ~ group, data = data_long)
#
#$group
#          diff          lwr       upr     p adj
#B-A 0.2822630 -0.292540425 0.8570664 0.5721402
#C-A 0.8561388  0.281335427 1.4309423 0.0011117
#D-A 1.4676027  0.892799258 2.0424061 0.0000000
#C-B 0.5738759 -0.000927561 1.1486793 0.0505270
#D-B 1.1853397  0.610536271 1.7601431 0.0000041
#D-C 0.6114638  0.036660419 1.1862672 0.0326371
```

Análisis de Varianza - ANOVA

Prueba de Tukey

- También se pueden visualizar los intervalos de confianza del 95% que resultan de la prueba de Tukey usando la función plot() en R:

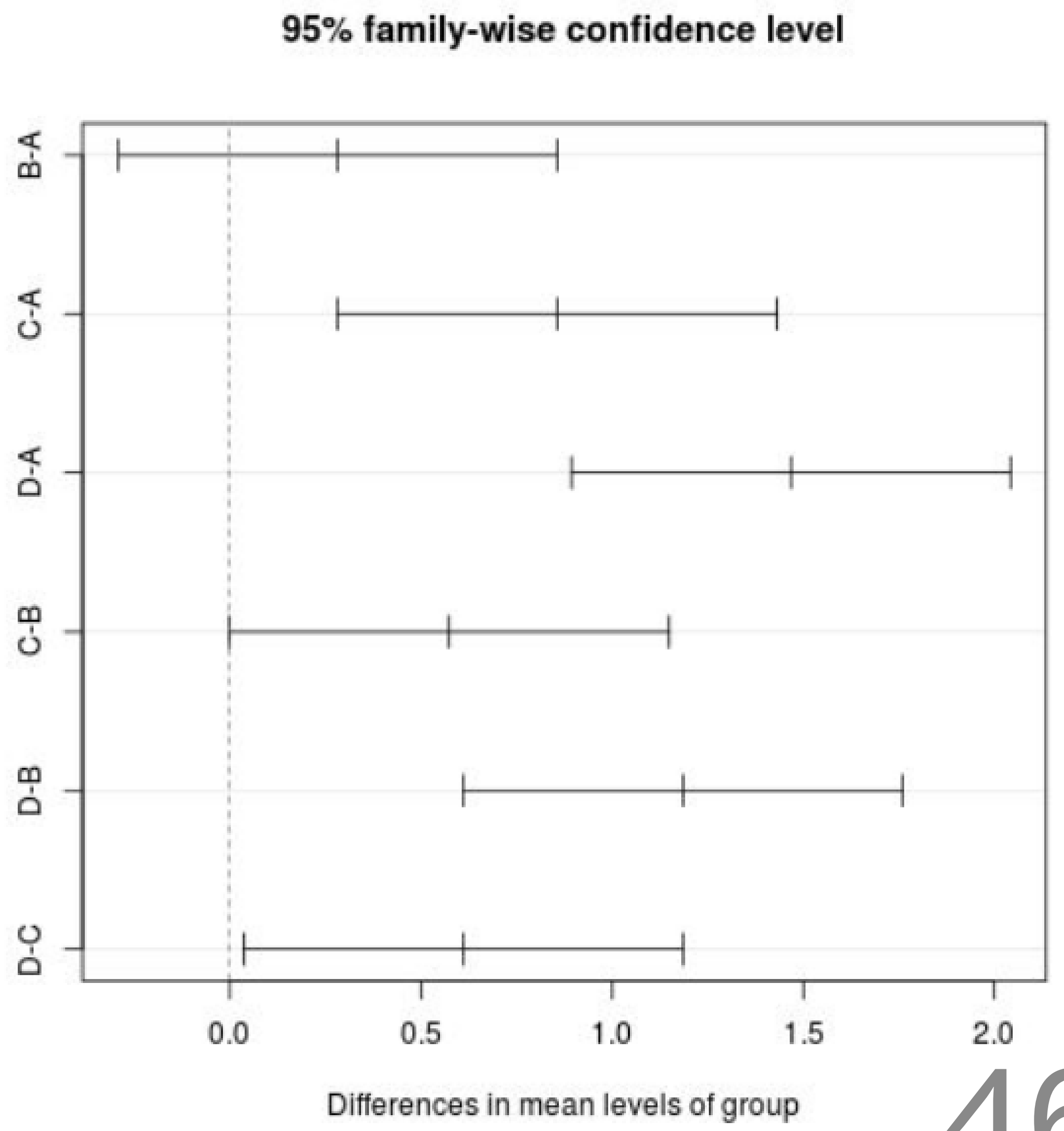
```
plot(TukeyHSD(anova_model, conf.level=.95))
```



Análisis de Varianza - ANOVA

Prueba de Tukey

- Si el intervalo contiene al 0, entonces la diferencia en las medias de los grupos no es estadísticamente significativa.
- En el ejemplo, las diferencias para B-A y C-B no son estadísticamente significativas, pero las diferencias para las otras cuatro comparaciones por pares sí lo son.



Análisis de Varianza – ANOVA de una vía

Modelo de datos

- Suponga que se tienen *a* tratamientos o niveles diferentes de un solo factor que quieren compararse.
- La respuesta observada de cada uno de los *a* tratamientos es una variable aleatoria.
- Los datos aparecerían como en la tabla:

Treatment (Level)	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots\ \vdots\ \vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

Análisis de Varianza – ANOVA de una vía

Modelo de datos

- Una entrada y_{ij} representa la observación *j*-ésima tomada bajo el nivel del factor o tratamiento *i.*
- Habrá, en general, *n* observaciones bajo el tratamiento *i*-ésimo.

Treatment (Level)	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

Análisis de Varianza – ANOVA de una vía

Modelo de datos

Treatment (Level)	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

- Es útil describir las observaciones de un experimento con un modelo. Una manera de escribir este modelo es el modelo de las medias:

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

- donde y_{ij} representa la observación *ij*-ésima, μ_i es la media del nivel del factor *i*-ésimo y ϵ_{ij} es un componente del error aleatorio que incorpora todas las demás fuentes de variabilidad del experimento.

Modelo de datos

- Una forma alternativa de escribir un modelo de los datos es definiendo:

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

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

- Entonces el modelo (denominado de los efectos) ahora se escribe como:

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

- Donde μ es un parámetro común a todos los tratamientos (la media global) y τ_i es un parámetro único del tratamiento *i*-ésimo (efecto del tratamiento *i*-ésimo).

Análisis de Varianza – ANOVA de una vía

Modelo de datos

- Tanto el modelo de las medias como el de los efectos son modelos estadísticos lineales; es decir, la variable de respuesta y_{ij} es una función lineal de los parámetros del modelo.

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

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

- El modelo de los efectos se encuentra con mayor frecuencia en la literatura.
- Tiene cierto atractivo intuitivo por cuanto μ es una constante y los efectos de los tratamientos τ_i representan desviaciones de esta constante cuando se aplican los tratamientos específicos.

Análisis de Varianza – ANOVA de una vía

Modelo de datos

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

- Al modelo de los efectos se le llama también el modelo del análisis de varianza simple o de un solo factor (o dirección), porque únicamente se investiga un factor.
- Además, será un requisito que el experimento se lleve a cabo en orden aleatorio para que el ambiente en el que se apliquen los tratamientos sea lo más uniforme posible.
- Por lo tanto, el diseño experimental es un diseño completamente aleatorizado.
- Los objetivos serán probar las hipótesis apropiadas acerca de las medias de los tratamientos y estimarlas.

Modelo de datos

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

- Para probar las hipótesis, se supone que los errores del modelo son variables aleatorias que siguen una distribución normal e independiente con media cero y varianza σ^2.
- Se supone asimismo que la varianza $\boldsymbol{\sigma}^2$ es constante para todos los niveles del factor.
- Esto implica que las observaciones son mutuamente independientes y que, además:

$$y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

Análisis de Varianza – ANOVA de una vía

Modelo de datos

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

- El modelo de los efectos describe dos situaciones:
- Primera: los *a* tratamientos pudieron ser elegidos expresamente por el experimentador. Se llama modelo con efectos fijos.
- En esta situación quieren probarse hipótesis acerca de las medias de los tratamientos, y las conclusiones se aplicarán únicamente a los niveles del factor considerados en el análisis. También puede desearse estimar los parámetros (μ , τ_i, σ^2).
- Las conclusiones no pueden extenderse a tratamientos similares que no fueron considerados explícitamente.

Análisis de Varianza – ANOVA de una vía

Modelo de datos

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

- Segunda: los *a* tratamientos podrían ser una muestra aleatoria de una población más grande de tratamientos (modelo con efectos aleatorios o modelo de los componentes de la varianza).
- En este caso sería deseable poder extender las conclusiones (que se basan en la muestra de tratamientos) a la totalidad de los tratamientos de la población, sea que se hayan considerado explícitamente en el análisis o no.
- Aquí las τ_i son variables aleatorias, y el conocimiento de las τ_i particulares que se investigaron es relativamente inútil. Más bien, se prueban hipótesis acerca de la variabilidad de las τ_i y se intenta estimar su variabilidad.

Análisis de Varianza – ANOVA de una vía

Análisis del modelo de efectos fijos

- Si se desea desarrollar el análisis de varianza de un solo factor para el modelo con efectos fijos, el interés se encuentra en probar la igualdad de las *a* medias de los tratamientos.

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

- Las hipótesis apropiadas son:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_1: \mu_i \neq \mu_j \quad \text{para algún par } (i, j)$$

Análisis de Varianza – ANOVA de una vía

Análisis del modelo de efectos fijos

- La media μ_i del tratamiento *i*-ésimo se descompone en dos: $\mu_i = \mu + \tau_i$
- Por lo general, μ se considera como una media global, de tal modo que: $$\frac{\sum_{i=1}^{a} \mu_i}{a} = \mu$$
- Esto implica que: $\sum_{i=1}^{a} \tau_i = 0$
- Es decir, los efectos del tratamiento o factor pueden considerarse como desviaciones de la media global.

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

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_1: \mu_i \neq \mu_j$$

Análisis del modelo de efectos fijos

- Por consiguiente, una forma equivalente de escribir las hipótesis anteriores es en términos de los efectos de los tratamientos τ_i :

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

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_1: \mu_i \neq \mu_j$$

$$H_0: \tau_1 = \tau_2 = \cdots \tau_a = 0$$

$H_1: \tau_i \neq 0$ para al menos una *i*

- Se habla entonces de probar la igualdad de las medias de los tratamientos o de probar que los efectos de los *a* tratamientos (los τ_i) son cero.
- El procedimiento apropiado para ello es el análisis de varianza.

Análisis de Varianza – ANOVA de una vía

Verificación de la adecuación del modelo

- Para probar formalmente que no hay diferencias en las medias de los tratamientos se requiere que se satisfagan ciertos supuestos.
- Específicamente, estos supuestos son que el modelo $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ describe de manera adecuada las observaciones, y que los errores siguen una distribución normal e independiente con media cero y varianza σ^2 constante pero desconocida.
- Si estos supuestos se satisfacen, el procedimiento del análisis de varianza es una prueba exacta de la hipótesis de que no hay diferencias en las medias de los tratamientos.

Análisis de Varianza – ANOVA de una vía

Verificación de la adecuación del modelo

- Sin embargo, es común que en la práctica estos supuestos no se satisfagan exactamente.
- Por consiguiente, en general no es prudente confiar en el análisis de varianza hasta haber verificado estos supuestos.
- Las violaciones de los supuestos básicos y la adecuación del modelo pueden investigarse con facilidad mediante el examen de los residuales.

Verificación de la adecuación del modelo

- El residual de la observación *j*-ésima en el tratamiento *i*-ésimo se define como:

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

- donde $\hat{y}_{ij}$ es una estimación de la observación y_{ij} correspondiente que se obtiene como sigue:

$$\begin{aligned}\hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i \\ &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) \\ &= \bar{y}_{i.}\end{aligned}$$

- El resultado intuitivamente claro es que la estimación de cualquier observación en el tratamiento *i*-ésimo no es sino el promedio del tratamiento correspondiente.

Verificación de la adecuación del modelo

- El examen de los residuales deberá ser una parte automática de cualquier ANOVA.
- Si el modelo es adecuado, los residuales deberán estar sin estructura; es decir, no deberán contener patrones obvios.
- A través de un estudio de los residuales, pueden descubrirse muchos tipos de inadecuaciones del modelo y violaciones de los supuestos subyacentes.

Análisis de Varianza – ANOVA de una vía

Supuesto de normalidad

- La verificación del supuesto de normalidad puede hacerse graficando un histograma de los residuales.
- Si se satisface el supuesto de $NID(0, \sigma^2)$ para los errores, esta gráfica deberá aparecer como una muestra de una distribución normal con centro en cero.
- Cuando se trabaja con muestras pequeñas, suelen ocurrir fluctuaciones significativas. La aparición de una desviación moderada de la normalidad no implica necesariamente una violación seria de los supuestos. Las desviaciones marcadas de la normalidad son potencialmente serias y requieren análisis adicional.

Análisis de Varianza – ANOVA de una vía

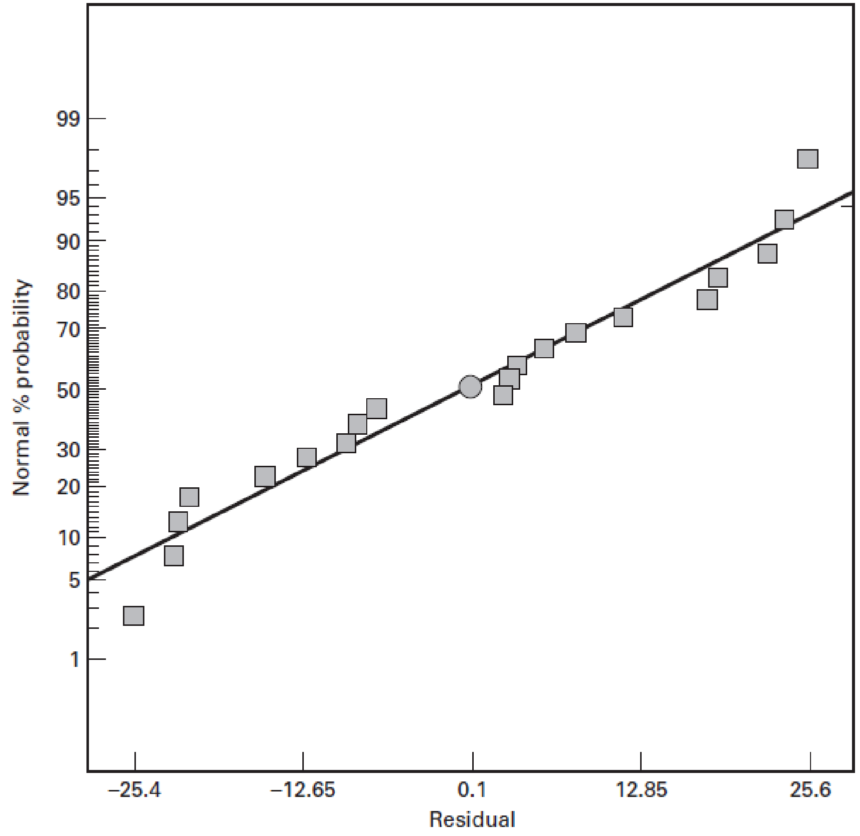
Supuesto de normalidad

- Un procedimiento en extremo útil es construir una gráfica de probabilidad normal de los residuales.
- Recuerde que se utilizó una gráfica de probabilidad normal de los datos originales para verificar el supuesto de normalidad cuando se usó la prueba t.
- En el análisis de varianza, esta comprobación se realiza con los residuales, dado que esto es lo que indica el modelo de datos.

Análisis de Varianza – ANOVA de una vía

Supuesto de normalidad

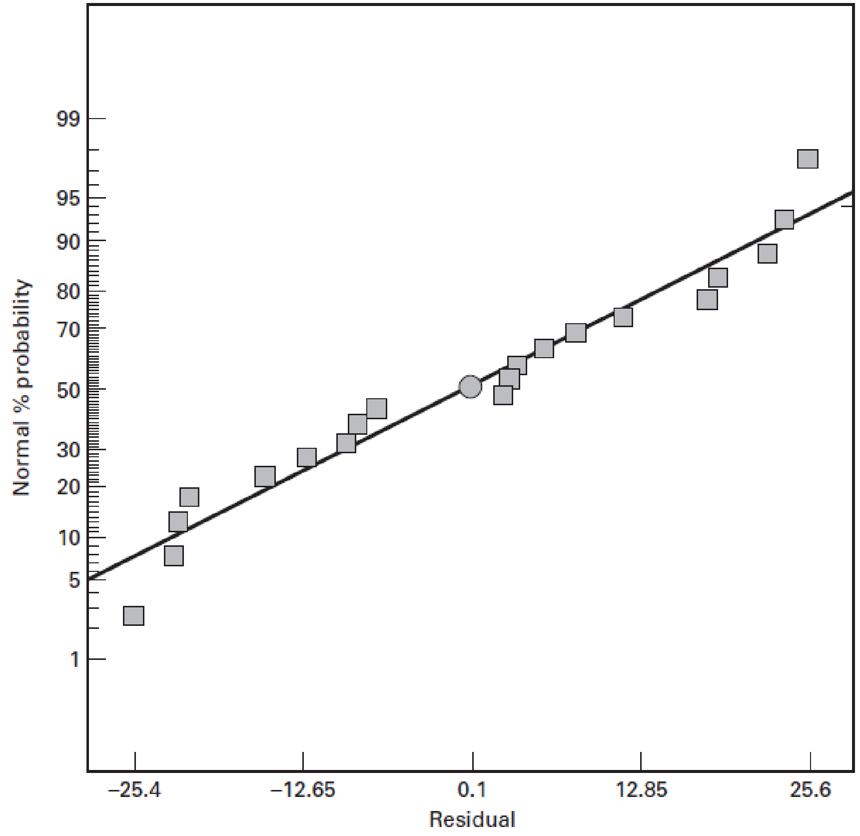
- Si la distribución fundamental de los errores es normal, la gráfica de residuales tendrá la apariencia de una línea recta.
- Para visualizar la línea recta, deberá prestarse más atención a los valores centrales de la gráfica que a los valores extremos.



Análisis de Varianza – ANOVA de una vía

Supuesto de normalidad

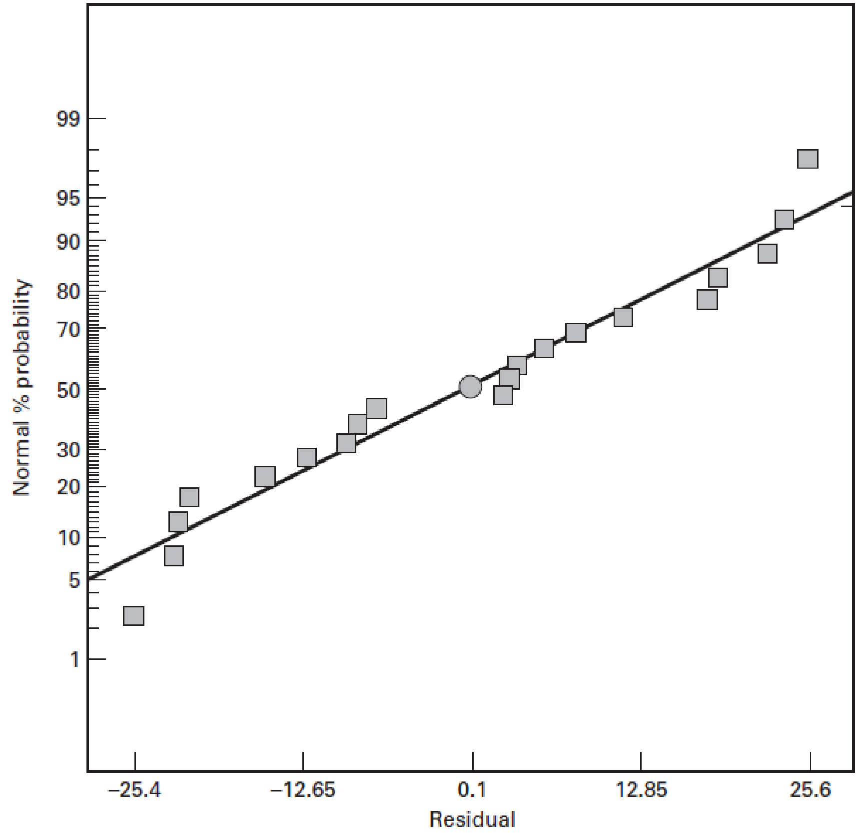
- La impresión general al examinar este ejemplo es que la distribución de errores es aproximadamente normal.
- La tendencia a doblarse ligeramente hacia abajo en el lado izquierdo y hacia arriba en el lado derecho implica que las colas de la distribución de error son algo más delgadas de lo que se esperaría en una distribución normal.



Análisis de Varianza – ANOVA de una vía

Supuesto de normalidad

- En general, las desviaciones moderadas de la normalidad no son motivo de gran preocupación en el análisis de varianza de efectos fijos.
- Puesto que la prueba *F* sólo se afecta ligeramente, se dice que el análisis de varianza (y los procedimientos relacionados como las comparaciones múltiples) es robusto con respecto al supuesto de normalidad.



Análisis de Varianza – ANOVA de una vía

Supuesto de normalidad

- Una anomalía muy común que suele ponerse de manifiesto en las gráficas de probabilidad normal es un residual que es mucho más grande que cualquier otro. (punto atípico o *outlier*).
- La presencia de uno o más puntos atípicos puede introducir serias distorsiones en el análisis de varianza, por lo que cuando se localiza un punto atípico potencial, se requiere una investigación atenta.
- En muchas ocasiones, la causa del punto atípico es un error en los cálculos o un error al codificar o copiar los datos.

Análisis de Varianza – ANOVA de una vía

Supuesto de normalidad

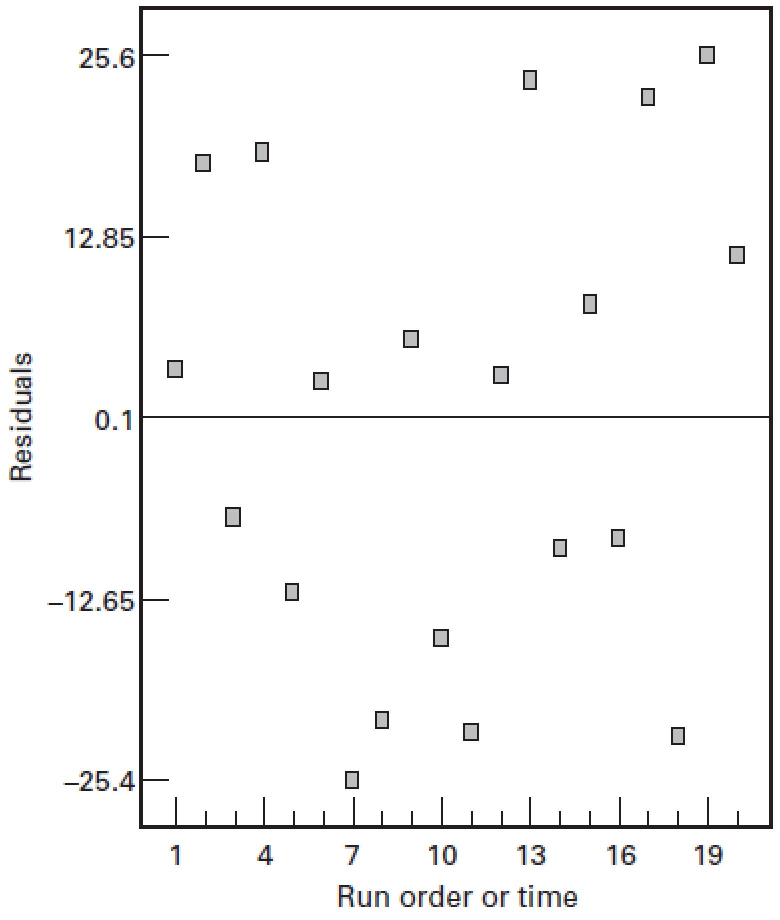
- Si no es ésta la causa, las circunstancias experimentales que rodean esta corrida particular deben estudiarse con atención.
- Si la respuesta atípica ocurre en un valor particularmente deseable, el punto atípico puede ser más informativo que el resto de los datos.
- Deberá tenerse cuidado de no rechazar o descartar una observación atípica a menos que se tengan razones no estadísticas de peso para hacerlo.
- En el peor de los casos, puede terminarse con dos análisis; uno con el punto atípico y uno sin él.

Supuesto de independencia

- El graficar los residuales en el orden temporal de la recolección de los datos es útil para detectar correlaciones entre los residuales.
- Si se identifican correlaciones en los residuales (como una dispersión mayor en un extremo), implicaría que el supuesto de independencia de los errores ha sido violado.
- Se trata de un problema potencialmente serio y cuya solución es difícil, por lo que de ser posible es importante evitar el problema cuando se colecten los datos.
- La aleatorización adecuada es importante para conseguir la independencia.

Supuesto de independencia

- En la gráfica se muestran los residuales contra el tiempo de la recolección de los datos para el experimento del algoritmo de compresión.
- No hay razón para sospechar alguna violación de los supuestos de independencia o de una varianza constante.



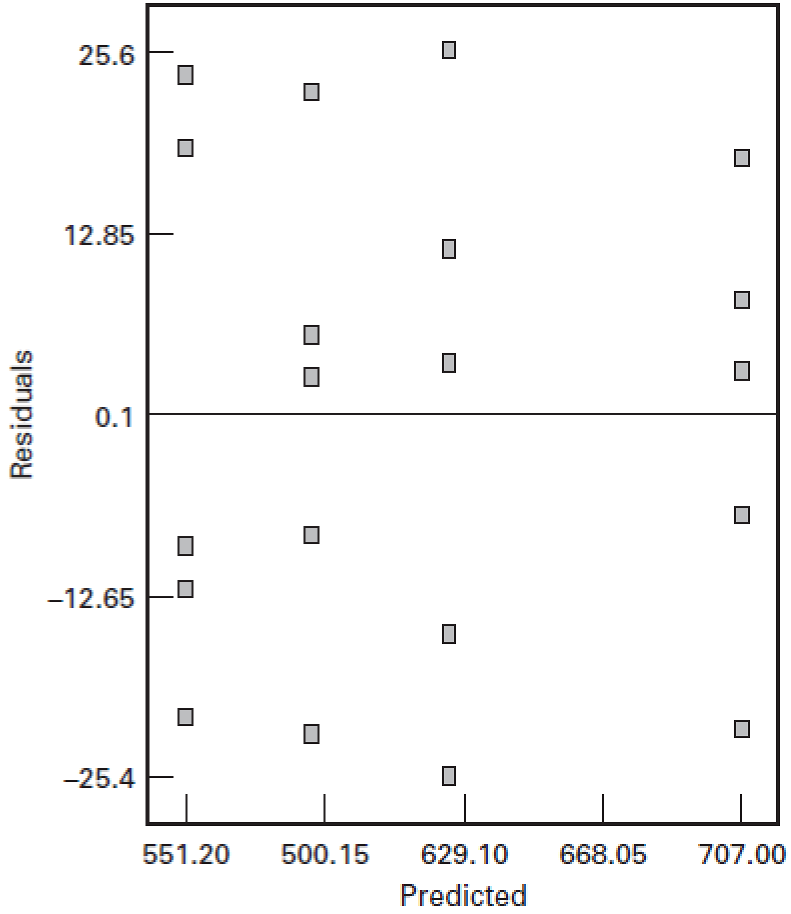
Plot of residuals versus run order or time

Supuesto de igualdad de varianza

- Si el modelo es correcto y se satisfacen los supuestos, los residuales deberán estar sin estructura; en particular, no deberán estar relacionados con ninguna otra variable, incluyendo la respuesta predicha.
- Una verificación simple es graficar los residuales contra los valores ajustados $\hat{y}_{ij}$ (predicciones que hace el modelo para cada observación, valores estimados de las medias para cada grupo)
- Para el modelo de un experimento con un solo factor, recuerde que $\hat{y}_{ij} = \bar{y}_{i.}$ (el promedio del tratamiento *i*-ésimo).

Supuesto de igualdad de varianza

- En la figura se grafican los residuales contra los valores ajustados para el ejemplo de los datos del algoritmo de compresión .
- No es evidente ninguna estructura inusual.



Plot of residuals versus fitted values

Análisis de Varianza – ANOVA de una vía

Supuesto de igualdad de varianza

- Un defecto que puede salir a relucir en esta gráfica es una varianza no constante.
- En ocasiones la varianza de las observaciones se incrementa cuando la magnitud de la observación se incrementa.
- Éste sería el caso si el error o ruido de fondo del experimento fuera un porcentaje constante de la magnitud de la observación.
- Esto ocurre comúnmente con muchos instrumentos de medición; el error es un porcentaje de la escala de medición.

Análisis de Varianza – ANOVA de una vía

Supuesto de igualdad de varianza

- Si éste fuera el caso, los residuales se harían mayores conforme y_{ij} se hiciera más grande, y la gráfica de los residuales contra y_{ij} se vería como un embudo con la boca hacia afuera.
- Una varianza no constante también surge en los casos en que los datos siguen una distribución no normal, sesgada, porque en las distribuciones sesgadas la varianza tiende a ser una función de la media.

Supuesto de igualdad de varianza

- Si se viola el supuesto de la homogeneidad de las varianzas, la prueba *F* sólo resulta afectada ligeramente en el modelo balanceado (el mismo tamaño de la muestra en todos los tratamientos) con efectos fijos.
- Sin embargo, en diseños no balanceados o en casos en que una de las varianzas es considerablemente más grande que las demás, el problema es más grave.
- Ésta es una buena razón para escoger tamaños de las muestras iguales siempre que sea posible.

Análisis de Varianza – ANOVA de una vía

Supuesto de igualdad de varianza

- Aun cuando es frecuente el uso de las gráficas residuales para diagnosticar la desigualdad de la varianza, se han propuesto también varias pruebas estadísticas.
- Estas pruebas pueden considerarse como pruebas formales de las hipótesis:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2$$

H_1: el enunciado anterior no es verdadero para al menos una σ^2_i

Supuesto de igualdad de varianza

- Un procedimiento muy utilizado es la prueba de Bartlett.
- Esta prueba es muy sensible al supuesto de normalidad. Por consiguiente, cuando la validez de este supuesto está en duda, no deberá usarse la prueba de Bartlett.
- La prueba de Levene es un procedimiento muy útil que es robusto en cuanto a las desviaciones de la normalidad.

Gráficas de residuales contra otras variables

- Si se han recolectado datos de otras variables que posiblemente pudieran afectar la respuesta, los residuales deberán graficarse contra estas variables.
- Los patrones en tales gráficas residuales implican que la variable afecta la respuesta.
- Esto sugiere que la variable debería controlarse con mayor atención en experimentos futuros o que debería incluirse en el análisis.

Análisis de Varianza – ANOVA de una vía

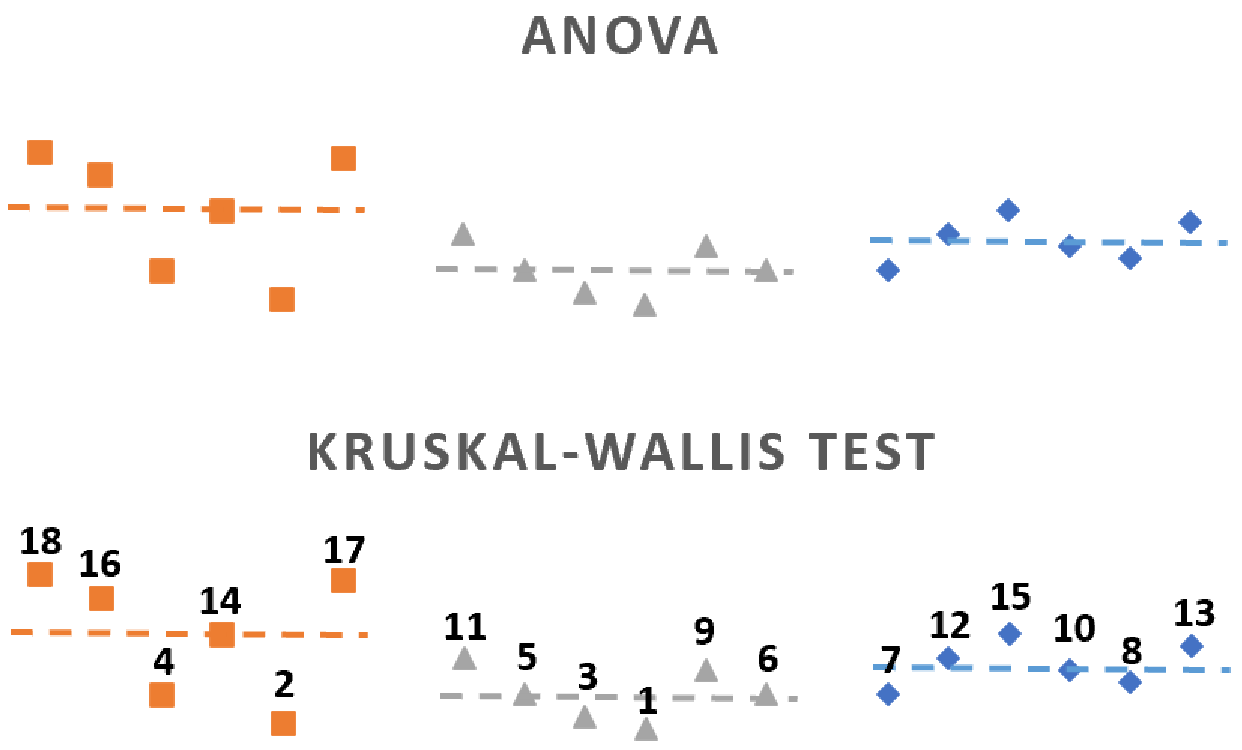
Métodos no paramétricos

- En situaciones en las que el supuesto de normalidad no está justificado, el experimentador puede desear utilizar un procedimiento alternativo al análisis de varianza de la prueba F que no dependa de este supuesto.
- Kruskal y Wallis (1952) desarrollaron un procedimiento de este tipo.
- La prueba de Kruskal-Wallis se utiliza para probar la hipótesis nula de que los tratamientos son idénticos frente a la hipótesis alternativa de que algunos de los tratamientos generan observaciones que son mayores que otros.

Análisis de Varianza – ANOVA de una vía

Métodos no paramétricos

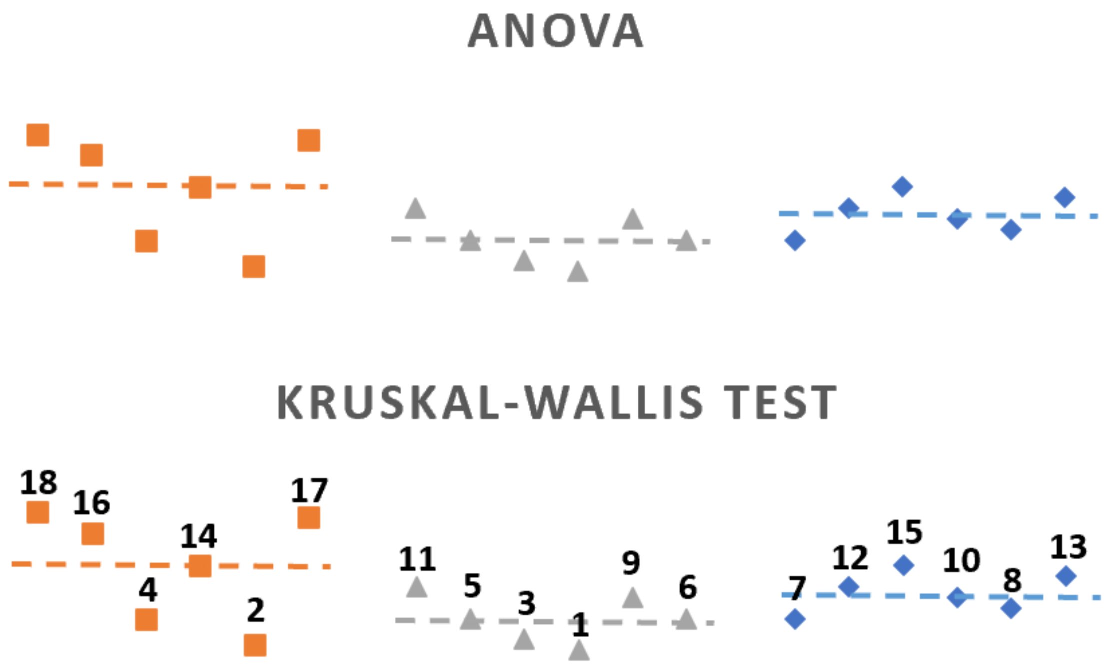
- La prueba de Kruskal-Wallis es una alternativa no paramétrica al análisis de varianza habitual.
- La prueba se basa en la clasificación (*ranks*) de todas las observaciones de todos los grupos en una sola lista ordenada de menor a mayor.



Análisis de Varianza – ANOVA de una vía

Métodos no paramétricos

- Luego, se calculan las posiciones para cada observación en esta lista ordenada, y se suman las posiciones de cada grupo.
- La idea subyacente es que, si no hay diferencias significativas entre los grupos, entonces las posiciones deberían distribuirse de manera similar en todos los grupos.



Referencias

- Montgomery, D.C. (2013). Design of Experiments. John Wiley & sons.
- Statology.org (2019). How to Conduct a One-Way ANOVA in R. https://www.statology.org/one-way-anova-r/