DISEÑO DE EXPERIMENTOS Y ANÁLISIS DE VARIANZA (ANOVA)

El análisis de varianza (ANOVA) se utiliza para probar la hipótesis nula de que las medias de dos o más poblaciones son iguales. En el contexto de un experimento, ANOVA permite probar la hipótesis nula de que las medias son iguales a dos o más niveles de un factor o variable independiente. La hipótesis alternativa es que al menos una de estas medias es diferente.

Estos conceptos son más fáciles de comprender con un ejemplo concreto. Imaginen que una empresa elaboró un sistema de purificación de aguas en Santiago. Esta empresa utilizó aleatoriamente tres métodos para armar los purificadores (A, B y C) que se diferenciaban en los pasos de ensamblaje. La duda de la empresa era con qué método se podían armar más purificadores en una semana. Aquí la **variable independiente o factor** es el método para armar el purificador. Como hay tres métodos, podemos decir que en este experimento hay tres **tratamientos.** Los trabajadores que emplean los tres tratamientos para armar el sistema son las unidades experimentales que constituyen las tres poblaciones de interés. El objetivo de la empresa es ver si el número promedio producido por semana es igual en las tres poblaciones tratadas. La **variable dependiente** o **variable de respuesta** es entonces **el** número de sistemas de purificación armados por semana.

Supongamos que diseño experimental fue **completamente aleatorizado.** Cada tratamiento se le asigna de forma aleatoria a cada trabajador. En este experimento se obtiene una medición para cada método de armar filtros. *Replicaremos* este proceso para 15 trabajadores y a cada 5 se le asigna un método de forma aleatoria.

A continuación, se muestra el número de unidades armadas por cada empleado en una semana. Calculen la varianza y media muestral y desviación estándar obtenidas con cada método.

Tratamientos	**A**	**B**	**C**	**Total**
Datos muestrales	58	58	48	-
	64	69	57	
	55	71	59	
	66	64	47	
	67	68	49	
Media muestral				
Varianza muestral				
DE muestral				

Como desconocemos los verdaderos valores de μ_1, μ_2 y μ_3, tenemos que hacer una inferencia y utilizar las medias muestrales para probar nuestra hipótesis nula a nivel poblacional:

$$H_0\colon\ \mu_1 = \mu_2 = \mu_3$$

$$H_1\colon No\ todas\ las\ medias\ poblacionales\ son\ iguales$$

ANOVA se emplea para determinar si las diferencias observadas entre las tres medias muestrales son lo suficientemente grande para rechazar nuestra hipótesis nula. Es decir, para probar si hay un método de armado más eficiente que otro.

Los **supuestos para el uso de ANOVA** son los siguientes:

- La variable de respuesta en cada población se distribuye normalmente.
- La varianza de la variable de respuesta (σ^2) es la misma en todas las poblaciones.
- Las observaciones deben ser independientes.

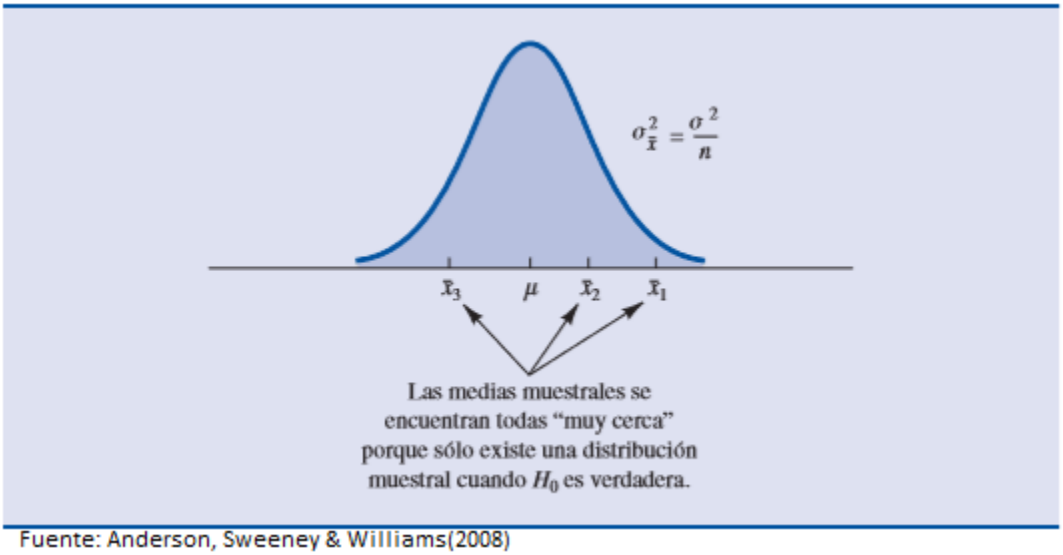
¿Qué significa cada uno de estos supuestos?

MIRADA GENERAL DEL ANALISIS DE VARIANZA

Mientras más similares sean las medias muestrales, tendremos mayor evidencia para concluir que las medias poblacionales son iguales, apoyando la hipótesis nula. Entre mayor sea la diferencia entre las medias muestrales, mayor evidencia tendremos para decir que las medias poblacionales son distintas, favoreciendo la hipótesis alternativa. Es decir, la variabilidad entre medias muestrales puede ser "pequeña" o "grande", favoreciendo H_0 o H_1 respectivamente.

Basándose en que la suposición de la hipótesis nula es verdadera, cada muestra proviene de la misma población y solo hay una distribución muestral de $\bar{x}$.

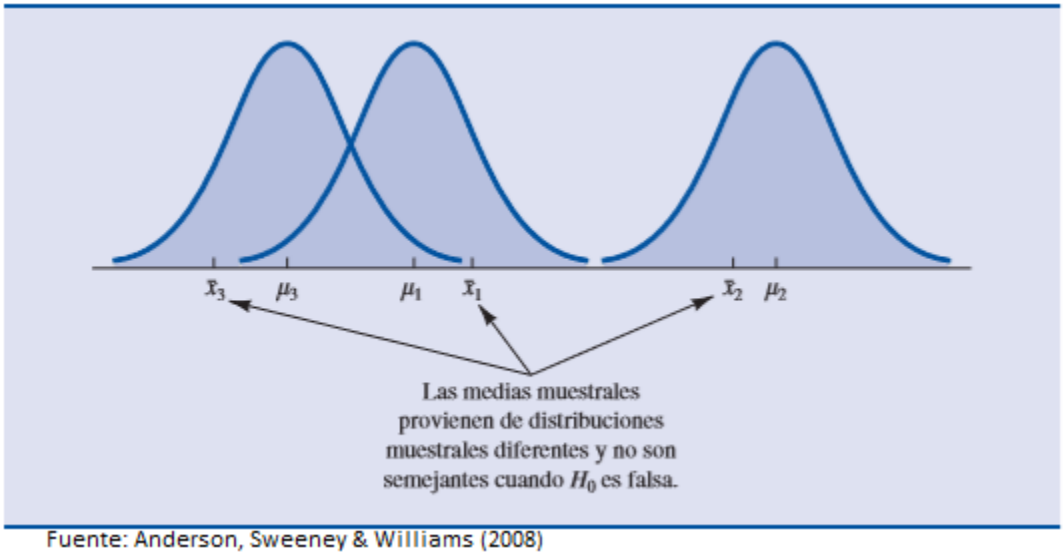
Distribución muestral de $\bar{x}$ cuando H_0 es verdadera



Fuente: Anderson, Sweeney & Williams(2008)

Si asumimos que la hipótesis nula es falsa, las medias muestrales no estarán tan cercas unas de otras. Entonces $s_{\bar{x}}^2$ será mayor, haciendo que la estimación de la varianza poblacional sea mayor (estimación de σ^2). **Generalmente cuando las medias poblacionales no son iguales, la estimación entre tratamientos entre medias sobreestima la varianza poblacional $\boldsymbol{\sigma^2}$.**

Distribución muestral de $\bar{x}$ cuando H_0 es Falsa



Fuente: Anderson, Sweeney & Williams (2008)

Para probar si H_0 es verdadera, estimaremos la variabilidad de las medias entre tratamientos y al interior de cada tratamiento.

Estimación de la varianza poblacional $\boldsymbol{\sigma^2}$ entre tratamientos

Lo primero es observar que se satisfacen los supuestos. Luego consideramos cada una de las tres medias muestrales obtenidas anteriormente. Como los tamaños muestrales son iguales podemos estimar la media de la distribución muestral de $\bar{x}$ como el promedio de las medias muestrales. La v**arianza poblacional entre tratamientos se calcula utilizando las medias muestrales**:

Fórmula $s_{\bar{x}}^2 = \frac{\sum_{i=1}^{n}(\bar{x}_i - \bar{x}_{\bar{x}})^2}{n-1}$

$$s_{\bar{x}}^2 = \frac{(62-60)^2 + (66-60)^2 + (52-60)^2}{3-1} = \frac{104}{2} = 52$$

La fórmula para la varianza poblacional σ^2se obtiene despejando $\sigma_{\bar{x}}^2 = \sigma^2/n$:

$$\sigma^2 = n\sigma_{\bar{x}}^2$$

Entonces, la **varianza poblacional σ^2 entre tratamientos** es igual al tamaño muestral del tratamiento n multiplicado por la Estimación de la varianza de las medias muestrales $\sigma_{\bar{x}}^2$: $5(52) = 260$.

Estimación conjunta o dentro de los tratamientos de la varianza poblacional σ^2

En ANOVA la variación dentro de cada una de las muestras también tiene efecto sobre la conclusión a la que se llega. Cada muestra proporciona un estimador insesgado de σ^2, y se pueden combinar en una estimación general. A esta se le conoce como estimación conjunta o dentro de los tratamientos de σ^2. **En este caso no afecta que las medias poblacionales sean o no iguales**. Si los tamaños muestrales son iguales, obtenemos la estimación de los tratamientos de σ^2 del promedio de las varianzas muestrales. **Es decir, ahora ocupamos las varianzas muestrales en el cálculo**:

$$Estimación\ de\ \sigma^2\ dentro\ de\ los\ tratamientos = \frac{27.5 + 26.5 + 31}{3} = \frac{85}{3} = 28.33$$

En este experimento la estimación de σ^2 entre los tratamientos (260) es mucho mayor que la estimación de σ^2 dentro de los tratamientos (28.33). El cociente entre estas dos estimaciones es 260/28.33=9.18. **Recordar que el método entre tratamientos proporciona una buena estimación de σ^2 si la hipótesis nula es verdadera, si la hipótesis nula es falsa el método entre tratamientos sobreestima σ^2. El método dentro de tratamientos proporciona una buena estimación para ambos casos. Esto no sindica si la hipótesis nula es verdadera tendremos un cociente cercano a 1, en cambio si es falsa nuestra hipótesis nula su cociente será más grande.** Luego veremos que tan grande debe ser el cociente para rechazar la hipótesis nula.

En síntesis, con ANOVA se realiza una doble estimación: una basada en la variabilidad entre las medias muestrales y otra basada en la variabilidad entre los datos dentro de cada muestra. Al comprar estas dos estimaciones de σ^2, podemos determinar si rechazamos o no nuestra hipótesis nula, si las medias poblacionales son iguales o hay al menos una diferente a las otras.