ESTADÍSTICA
APLICADA A LA GASTRONOMÍA

CONTENIDO

OBJETIVOS GENERALES

- Ser capaz de reflexionar de forma crítica sobre los datos que se le presenten en diferentes situaciones.
- Sistematizar la información numérica. (Capacidad de análisis y síntesis).
- Resolución de problemas.
- Razonamiento crítico.
- Comunicación oral y escrita.

OBJETIVOS ESPECÍFICOS

- Analizar datos de una y dos variables.
- Entender el uso de métodos gráficos.
- Resolver problemas de probabilidad.
- Entender cómo y por qué se incorpora el concepto de probabilidad en la variable analizada (variable aleatoria).
- Evaluar la posibilidad de que se produzca una determinada situación y saber caracterizar los diferentes tipos de situaciones que pueden producirse a través de los modelos probabilísticos.

UNIDAD 1

PRUEBA DE HIPÓTESIS

La estadística inferencial es el proceso de usar la información de una muestra para describir el estado de una población. Sin embargo es frecuente que usemos la información de una muestra para probar un reclamo o conjetura sobre la población. El reclamo o conjetura se refiere a una hipótesis. El proceso que corrobora si la información de una muestra sostiene o refuta el reclamo se llama prueba de hipótesis.

HIPÓTESIS Y NIVELES DE SIGNIFICANCIA

En la prueba de hipótesis se pone a prueba un reclamo hecho sobra la naturaleza de una población a base de la información de una muestra. El reclamo se llama ***hipótesis estadística.***

Hipótesis Estadística: Una hipótesis estadística es un reclamo hecho sobre la naturaleza de una población.

Por ejemplo, la premisa formulada por un productor de baterías para autos de que su batería dura en promedio 48 meses, es una hipótesis estadística porque el manufacturero no inspecciona la vida de cada batería que él produce.

Si surgieran quejas de parte de los clientes, entonces se pone a prueba el reclamo del manufacturero. La hipótesis estadística sometida a prueba se llama la ***hipótesis nula***, y se denota como H_0.

CÓMO ESTABLECER LA HIPÓTESIS NULA Y LA ALTERNA

Hipótesis Nula **(H_0):** Premisa, reclamo, o conjetura que se pronuncia sobre la naturaleza de una o varias poblaciones.

Por ejemplo, para probar o desaprobar el reclamo pronunciado por el productor de baterías debemos probar la hipótesis estadística de que $\mu \geq 48$. Por lo tanto, la hipótesis nula es:

$$H_0 : \mu \geq 48.$$

Luego se procede a tomar una muestra aleatoria de baterías y medir su vida media. Si la información obtenida de la muestra no apoya el reclamo en la hipótesis nula (H0), entonces otra cosa es cierta. La premisa alterna a la hipótesis nula se llama hipótesis alterna y se representa por H1.

Hipótesis Alterna: Una premisa que es cierta cuando la hipótesis nula es falsa.

Por ejemplo, para el productor de baterías

$$H_0 : \mu \geq 48 \text{ y}$$
$$H_1 : \mu < 48$$

Para probar si la hipótesis nula es cierta, se toma una muestra aleatoria y se calcula la información, como el promedio, la proporción, etc. Esta información muestral se llama estadística de prueba.

Estadística de Prueba: Una estadística de prueba se basa en la información de la muestra como la media o la proporción .

ERROR TIPO 1 Y ERROR TIPO 2

A base de la información de una muestra nosotros podemos cometer dos tipos de errores en nuestra decisión.

1. Podemos rechazar un $\mathbf{H_0}$ que es cierto.
2. Podemos aceptar un $\mathbf{H_0}$ que es falso.

El primero se llama error Tipo 1

Error Tipo 1: Cuando rechazamos una Hipótesis Nula que es cierta cometemos error tipo 1.

Y el segundo error se llama error Tipo 2.

Error Tipo 2: Cuando aceptamos una Hipótesis Nula que es falsa cometemos error tipo 2.

NIVEL DE SIGNIFICANCIA (α)

Para ser muy cuidadosos en no cometer el error tipo 1, debemos especificar la probabilidad de rechazar $\mathbf{H_0}$, denotada por α . A ésta se le llama **nivel de significancia**.

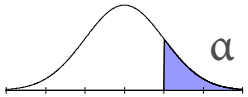
Nivel de Significancia: La probabilidad (α)más alta de rechazar $\mathbf{H_0}$ cuando $\mathbf{H_0}$ es cierto se llama nivel de significancia.

Comentario: Para mantener la probabilidad de cometer el error tipo 1 baja, debemos escoger un valor pequeño de α .

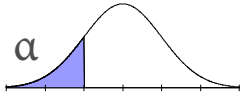
Usando un valor preasignado de α se construye una **región de rechazo** o **región crítica** en la curva normal estándar o en la curva ***t*** que indica si debemos rechazar H_0.

Región Crítica o de Rechazo: Una región crítica o de rechazo es una parte de la curva de ***z*** o de la curva ***t*** donde se rechaza H_0. La región puede ser de una cola o de dos dependiendo de la hipótesis alterna.

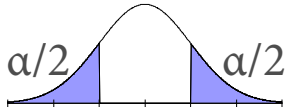
Ejemplos Para H_1: $\mu >$ valor aceptado, la región de rechazo está dada por:
(cola derecha, ***z*** ó ***t***)



Para H_1 : $\mu <$ valor aceptado, la región de rechazo está dada por:
(cola izquierda, ***z*** ó ***t***)



Para H_1 : $\mu \neq$ valor aceptado, la región de rechazo es de dos colas y está dada por:
(2-colas, ***z*** ó ***t***)

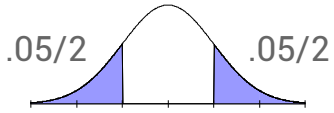


Ejemplo 1: Determine si la región de rechazo es de la cola derecha, de la cola izquierda o de dos colas.

a. $\mathbf{H_0}$: $\mu = 15$, $\mathbf{H_1}$: $\mu \neq 15$, $\alpha = .05$
b. $\mathbf{H_0}$: $p \leq 0.7$, $\mathbf{H_1}$: $p > 0.7$, $\alpha = .02$

Solución: La forma de la región de rechazo está determinada por la hipótesis alterna.

a. $\mathbf{H_1}$: $\mu \neq 15$ significa que la región está en ambas colas.



b. $\mathbf{H_1}$: $p > 7$ significa que la región está en la cola derecha.

USO DE LA TABLA DE LA DISTRIBUCIÓN NORMAL TÍPICA

Sea **Z** una variable aleatoria con distribución normal típica

1) Busca de la función de distribución de un número positivo.

Supongamos que queremos calcular $P\{ Z \leq 0{,}92\}$. Dicha probabilidad está representada por el área sombreada en la figura 1.

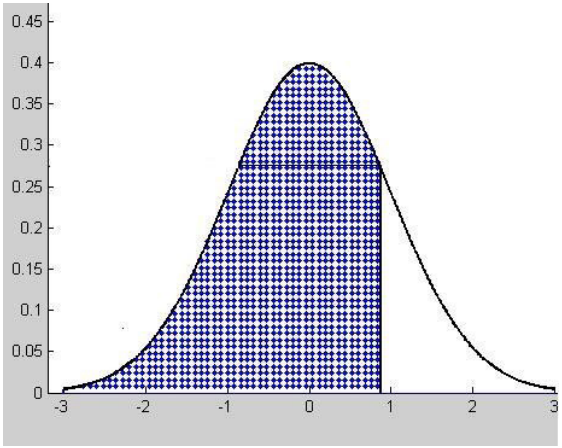


Figura1

Obtendremos la respuesta buscando en la tabla normal (para ello buscamos la fila correspondiente al número truncado en su primera cifra decimal (es decir 0,9) y la columna correspondiente a la segunda cifra decimal (es decir 0,02). La intersección de esa fila y esa columna nos indicará el número buscado).

z	0.00	0.01	0.02	0
0.0	0.5000	0.5040	0.5080	0.5
0.1	0.5398	0.5438	0.5478	0.55
0.2	0.5793	0.5832	0.5871	0.59
0.3	0.6179	0.6217	0.6255	0.62
0.4	0.6554	0.6591	0.6628	0.666
0.5	0.6915	0.6950	0.6985	0.701
0.6	0.7257	0.7291	0.7324	0.73
0.7	0.7580	0.7611	0.7642	0.76
0.8	0.7881	0.7910	0.7939	0.79
0.9	0.8159	0.8186	0.8212	0.82

Por lo tanto **P{ Z ≤ 0,92}= 0,8212.**

UNIDAD 2

DISTRIBUCIÓN T

T-STUDENT PARA 2 MUESTRAS INDEPENDIENTES

Uno de los análisis estadísticos más comunes en la práctica es probablemente el utilizado para comparar dos grupos independientes de observaciones con respecto a una variable numérica.

La aplicación de un contraste paramétrico requiere la normalidad de las observaciones para cada uno de los grupos. La comprobación de esta hipótesis puede realizarse tanto por métodos gráficos (por medio de histogramas, diagramas de cajas o gráficos de normalidad) como mediante tests estadísticos. Un número suficiente de observaciones (mayor de 30) justifica la utilización del mismo test.

Así mismo, este tipo de metodología exigirá que la varianza en ambos grupos de observaciones sea la misma. En primer lugar se desarrollará el test t de Student para el caso en el que se verifiquen ambas condiciones, discutiendo posteriormente el modo de abordar formalmente el caso en el que las varianzas no sean similares.

Bajo las hipótesis de normalidad e igual varianza la comparación de ambos grupos puede realizarse en términos de un único parámetro como el valor medio.
El t test para dos muestras independientes se basa en el estadístico:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)\hat{S}_1^2 + (m-1)\hat{S}_2^2}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad 1.$$

donde $\overline{X}$ e $\overline{Y}$ denotan el valor medio en cada uno de los grupos.

Si la hipótesis de partida es cierta el estadístico (1) seguirá una distribución t de Student con (n+m-2 grados de libertad). De ser así, el valor obtenido debería estar dentro del rango de mayor probabilidad según esta distribución.

Usualmente se toma como referencia el rango de datos en el que se concentra el 95% de la probabilidad. El valor-p que usualmente reportan la mayoría de paquetes estadísticos no es más que la probabilidad de obtener, según esa distribución, un dato más extremo que el que proporciona el test. Como ya se dijo, refleja también la probabilidad de obtener los datos observados si fuese cierta la hipótesis inicial. Si el valor-p es muy pequeño (usualmente se considera $p<0.05$) es poco probable que se cumpla la hipótesis de partida y se debería de rechazar. La región de aceptación corresponde por lo tanto a los valores centrales de la distribución para los que $p>0.05$.

En la siguiente tabla se determina los grados de libertad (en la primera columna) y el valor de α (en la primera fila). El número que determina su intersección es el valor crítico correspondiente. De este modo, si el estadístico que se obtiene toma un valor mayor se dirá que la diferencia es significativa.

DOS MUESTRAS INDEPENDIENTES CON VARIANZA DISTINTA

El caso en el que se dispone de dos grupos de observaciones independientes con diferentes varianzas, la distribución de los datos en cada grupo no puede compararse únicamente en términos de su valor medio. Obviamente, el primer problema a resolver es el de encontrar un método estadístico que nos permita decidir si la varianza en ambos grupos es o no la misma. El test de la razón de varianzas viene a resolver este problema. Bajo la suposición de que las dos poblaciones siguen una distribución normal y tienen igual varianza se espera que la razón de varianzas:

$$F = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2} = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

siga una distribución F de Snedecor con parámetros (n-1) y (m-1).

En este tipo de situaciones, donde no se debe aplicar el contraste basado en (1), podemos utilizar una modificación del test para el caso de varianzas desiguales, conocido como el test de Welch basada en el estadístico:

$$t = \frac{\left(\overline{X} - \overline{Y}\right)}{\sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}}}$$

que, bajo la hipótesis nula seguirá una distribución t de Student con un número f de grados de libertad que dependerá de las varianzas muestrales según la expresión:

$$f = \frac{\left(\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}\right)}{\frac{1}{n+1}\left(\frac{\hat{S}_1^2}{n}\right)^2 + \frac{1}{m+1}\left(\frac{\hat{S}_2^2}{m}\right)^2} - 2$$

La técnica para realizar el contraste es análoga a la vista anteriormente cuando las varianzas son desconocidas e iguales.

Al igual que en el caso anterior, podrá optarse por calcular el correspondiente 95% intervalo de confianza para la diferencia de medias dado por:

$$\left(\overline{X} - \overline{Y}\right) \pm t_{0.975}^{f} \sqrt{\frac{\hat{S}_1^2}{n} + \frac{\hat{S}_2^2}{m}}$$

UNIDAD 3

PRUEBAS DE HIPÓTESIS DE PROPORCIONES
Y CHI CUADRADA (VARIABLES NO MÉTRICAS)

Como investigadores en muchas ocasiones estamos interesados en un fenómeno cuyo comportamiento es expresado en porcentajes.

Por ejemplo, podemos estar interesados en probar si la proporción de potenciales electores que planean votar por el candidato del PRI es estadísticamente distinta de la proporción que declaró preferir el candidato del PAN.

1. PRUEBA DE HIPÓTESIS DE PROPORCIONES PARA UNA SOLA MUESTRA.

"Una encuesta realizada por Bancomer a 35 clientes indicó que un poco más del 74 por ciento tenían un ingreso familiar de más de $200,000 al año. Si esto es cierto, el banco desarrollará un paquete especial de servicios para este grupo. La administración quiere determinar si el porcentaje verdadero es mayor del 60 por ciento antes de desarrollar e introducir este nuevo paquete de servicios. Los resultados mostraron que 74.29 por ciento de los clientes encuestados reportaron ingresos de $200,000 o más al año".

El procedimiento para la prueba de hipótesis de proporciones es el siguiente:

1. Especifica la hipótesis nula y alternativa.

Hipótesis Nula: $\mathbf{H_0 = P \leq .60}$
Hipótesis Alternativa: $\mathbf{H_a = P > .60}$

donde P = la proporción de clientes con ingresos familiares anuales de $200,000 o más.

2. Específica el nivel de significación, α permitido. Para una $\alpha = .05$, el valor de tabla de Z para una prueba de una sola cola es igual a 1.64.

3. Calcula el error estándar de la proporción específicada en la hipótesis nula.

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

Donde:

p = proporción especificada en la hipótesis nula.

n = tamaño de la muestra.

Por consiguiente:

$$s_p = \sqrt{\frac{0.60(1-0.60)}{35}} = .0828$$

4. Calcula la estadística de prueba.

$$z = \frac{(proporción\_observada) - (proporción\_H_0)}{s_p}$$

$$z = \frac{0.7429 - 0.60}{0.0828} = 1.73$$

5. La hipótesis nula se rechaza porque el valor de la Z calculada es mayor que el valor crítico Z . El banco puede concluir con un 95 por ciento de confianza (1 - α = .95) que más de un 60 por ciento de sus clientes tienen ingresos familiares de $200,000 o más. La administración puede introducir el nuevo paquete de servicios orientado a este grupo.

El presidente del PRI en 1988, basado en su experiencia, sostiene que un 95% de los votos para las elecciones presidenciales han sido a favor de su partido. Los partidos de oposición levantaron una muestra de1,100 electores y encontraron que un 87% de ellos votaría por el PRI. El presidente del PRI quiere probar la hipótesis, con un nivel de significación de 0.05, que el 95% de los votos son para su partido.

Hipótesis Nula: **H_0 : p = 0.95**
Hipótesis Alternativa: **H_a : p ≠ 0.95**
Tamaño de muestra: **n=1,100**
Nivel de Significación **= 0.05.**

El primer paso es calcular el error estándar de la proporción utilizando el valor hipotético del porcentaje que históricamente vota por el PRI:

$$SE_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.95 * 0.05}{1100}} = 0.0066$$

Ahora sólo es necesario construir el intervalo de confianza:

$$p_o \pm 1.96 * SE_p$$

$$0.95 \pm (1.96 * 0.0066) = 0.937 \rightarrow 0.963$$

La proporción de .87 de votos por el PRI en la encuesta no cae en la región de aceptación, por lo tanto el presidente del PRI debe de "preocuparse" por que la tendencia entre los votantes es a favorecer menos al PRI.

SEXO DEL PATRON

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hombre	1634	83.9	83.9	83.9
	Mujer	314	16.1	16.1	100.0
	Total	1948	100.0	100.0	

Probemos la hipótesis de que el porcentaje de microempresas cuyos dueños son hombres captado por la ENAMIN es distinto de 88 por ciento.

Hipótesis Nula: **H_0 = P = 0.88**
Hipótesis Alternativa: **H_a = P ≠ 0.88**

$$s_p = \sqrt{\frac{0.88(1-0.88)}{1948}} = .0074$$

$$z = \frac{0.839 - 0.88}{0.0074} = -5.54$$

La hipótesis nula se rechaza porque el valor de la Z calculada es menor que el valor crítico Z de 1.96. Podemos concluir con un 95 por ciento de confianza (1 - α = .95) que la proporción captada por la ENAMIN es estadísticamente distinta de 0.88.

2. PRUEBA DE HIPÓTESIS PARA DIFERENCIAS ENTRE DOS PROPORCIONES (MUESTRAS INDEPENDIENTES)

Algunas veces estamos interesados en analizar la diferencia entre las proporciones de poblaciones de grupos con distintas características. Por ejemplo, pensemos que la administración de las tiendas Oxxo cree, sobre la base de una investigación, que el porcentaje de hombres que visitan sus tiendas 9 o más veces al mes (clientes frecuentes) es mayor que el porcentaje de mujeres que hacen lo mismo. Las especificaciones requeridas y el procedimiento para probar esta hipótesis es la siguiente:

1. Las hipótesis nula y alternativa son las siguientes:

$H_o = P_H - P_M \leq 0$ la proporción de hombres que reportan 9 o más visitas por mes es la misma o menor que la proporción de mujeres que hacen lo mismo.

$H_a = P_H - P_M > 0$ la proporción de hombres que reportan 9 o más visitas por mes es mayor a la proporción de mujeres que hacen lo mismo.

La información proporcionada es:

$$n_H = 45 \quad n_M = 71$$

$$P_H = .58 \quad P_M = .42$$

$$P_H - P_M = .58 - .42 = .16$$

2. Especifica el nivel de significación de $\alpha = .05$.
   El valor crítico para la prueba de una sola cola es de 1.64.

3. Estima el error estándar de la diferencia de las dos proporciones:

$$s_{p_{h-m}} = \sqrt{P(1-P)\left(\frac{1}{n_H} + \frac{1}{n_M}\right)}$$

Donde.

$$P = \frac{n_H P_H + n_M P_M}{n_H + n_M}$$

P_H = proporción muestra de hombres (H)
P_M = proporción muestra de mujeres (M)
N_H = tamaño de muestra hombres
N_M = tamaño de muestra mujeres

Por lo tanto:

$$P = \frac{45(.58) + 71(.42)}{45 + 71} = 0.48$$

y

$$s_{p_{h-m}} = \sqrt{.48(1-.48)\left(\frac{1}{45} + \frac{1}{71}\right)} = 0.10$$

4. Cálculo de prueba estadística.

$$Z = \frac{(diferencia\_entre\_proporciones\_observadas) - (diferencia\_entre\_proporciones\_H_o)}{s_{p_{h-m}}}$$

$$Z = \frac{(.58 - .42) - (0)}{.10} = 1.60$$

La hipótesis nula es aceptada porque el valor de la Z calculada es menor que el valor crítico Z. La administración no puede concluir con un 95 por ciento de confianza que la proporción de hombres que visita 9 o más veces los Oxxo es mayor que la proporción de mujeres.

SPSS no cuenta con procedimientos para hacer pruebas de hipótesis de proporciones. Probemos si el porcentaje de hombres dueños de microempresas es estadísticamente diferente del porcentaje de mujeres.

$$P = \frac{1634(83.9) + 314(16.1)}{1634 + 314} = 72.97$$

y

$$s_{p_{h-m}} = \sqrt{..73(1 - ..73)\left(\frac{1}{1634} + \frac{1}{314}\right)} = 0.0274$$

$$Z = \frac{(.839 - .161) - (0)}{.0274} = 24.74$$

La hipótesis nula es rechazada porque el valor de la Z calculada es mayor que el valor crítico Z. Podemos concluir que el porcentaje de hombres dueños de microempresas es estadísticamente superior al porcentaje de mujeres propietarias de microempresas.

3. CHI - CUADRADA

La mayoría de la información que se trabaja en las ciencias sociales o administrativas es de carácter no-métrico nominal. Por lo mismo, muchas de las técnicas multivariadas más populares, como la regresión lineal de mínimos cuadrados, presentan serias limitaciones analíticas.

¿Cómo analizar información nominal o categórica?

χ^2 es una prueba estadística no paramétrica para diferencias entre dos o más muestras donde frecuencias esperadas son comparadas en relación con frecuencias obtenidas.

χ^2 se utiliza para hacer comparaciones entre frecuencias y no entre valores medios.

Prueba No Paramétrica: procedimiento estadístico que no adopta ningún supuesto acerca de cómo se distribuye la característica bajo estudio en la población, y que sólo requiere datos nominales u ordinales.

Estas medidas son importantes porque la mayoría de la información en la investigación social y administrativa es de carácter nominal u ordinal, y porque no siempre estamos seguros que la característica que deseamos estudiar se distribuye normalmente en la población.

La prueba de significación χ^2 se refiere esencialmente a la distinción entre frecuencias esperadas y frecuencias obtenidas.

Las frecuencias esperadas f_e se refieren a los términos de la hipótesis nula, según la cual la frecuencia relativa (o proporción) se supone es la misma entre los dos grupos.

Por ejemplo, si se espera que un 50% de los negocios que llevan una contabilidad formal hayan iniciados sus actividades con ahorros personales, entonces también esperamos un 50% de aquellos que empezaron con financiamiento externo.

Las frecuencias obtenidas f_o se refieren a los resultados obtenidos en el estudio y que, por consiguiente, pueden variar o no de un grupo a otro.

Sólo si la diferencia entre las frecuencias observadas y obtenidas es suficientemente grande, se rechaza la hipótesis nula, y se concluye que existe una diferencia real en la población.

Como resultado, la **hipótesis nula** para la χ^2 señala que las poblaciones o grupos no difieren con respecto a la frecuencia de ocurrencia de una característica dada. Mientras que l**a hipótesis de investigación** señala que las diferencias entre las muestras reflejan diferencias reales en la población con respecto a la frecuencia relativa de una característica dada.

Ejemplo:

Hipótesis Nula: la frecuencia relativa de microempresas que llevan una contabilidad formal y que iniciaron su actividad con un financiamiento externo, es la misma que la frecuencia relativa de microempresas que llevan una contabilidad formal y que iniciaron su actividad con ahorros personales.

ó

Hipótesis Nula: la proporción de microempresas con contabilidad formal y cuyo inicio fue gracias a financiamiento externo, es la misma que la de microempresas con contabilidad formal cuyo inicio fueron ahorros personales.

UNIDAD 4

DISTRIBUCIÓN "F" FISHER

La necesidad de disponer de métodos estadísticos para comparar las varianzas de dos poblaciones es evidente a partir del análisis de una sola población. Frecuentemente se desea comparar la precisión de un instrumento de medición con la de otro, la estabilidad de un proceso de manufactura con la de otro o hasta la forma en que varía el procedimiento para calificar de un profesor universitario con la de otro.

Intuitivamente, podríamos comparar las varianzas de dos poblaciones, σ_1^2 y σ_2^2, utilizando la razón de las varianzas muestrales s^2_1/s^2_2. Si s^2_1/s^2_2 es casi igual a 1, se tendrá poca evidencia para indicar que σ_1^2 y σ_2^2 no son iguales. Por otra parte, un valor muy grande o muy pequeño para s^2_1/s^2_2, proporcionará evidencia de una diferencia en las varianzas de las poblaciones.

La variable aleatoria F se define como el cociente de dos variables aleatorias ji-cuadrada independientes, cada una dividida entre sus respectivos grados de libertad. Esto es:

$$F = \frac{U/\nu_1}{V/\nu_2}$$

donde U y V son variables aleatorias ji-cuadrada independientes con grados de libertad ν_1 y ν_2 respectivamente.

Sean U y V dos variables aleatorias independientes que tienen distribución ji cuadradas con ν_1 y ν_2 grados de libertad, respectivamente. Entonces la distribución de la variable aleatoria $F = \frac{U/\nu_1}{V/\nu_2}$ está dada por:

$$f(x) = \frac{\Gamma[(\nu_1+\nu_2)/2](\nu_1/\nu_2)^{\nu_1/2}x^{(\nu_1/2)-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(1+\nu_1 x/\nu_2)^{(\nu_1+\nu_2)/2}}$$

$0 < x < \infty$

y se dice que sigue la distribución F con ν_1 grados de libertad en el numerador y ν_2 grados de libertad en el denominador.
La media y la varianza de la distribución F son:

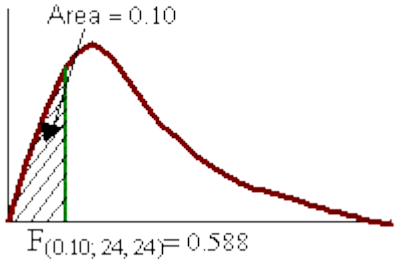
$$\mu = \frac{\nu_2}{\nu_2 - 2} \quad \text{para} \quad \nu_2 > 2$$

$$\sigma^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \quad \text{para} \quad \nu_2 > 4$$

La variable aleatoria F es no negativa, y la distribución tiene un sesgo hacia la derecha. La distribución F tiene una apariencia muy similar a la distribución ji-cuadrada; sin embargo, se encuentra centrada respecto a 1, y los dos parámetros ν_1 y ν_2 proporcionan una flexibilidad adicional con respecto a la forma de la distribución.

Si s_1^2 y s_2^2 son las varianzas muestrales independientes de tamaño n_1 y n_2 tomadas de **poblaciones normales** con varianzas σ_1^2 y σ_2^2, respectivamente, entonces:

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{s_1^2 \sigma_2^2}{s_2^2 \sigma_1^2} = \left(\frac{s_1}{s_2}\right)^2 \left(\frac{\sigma_2}{\sigma_1}\right)^2$$



ANÁLISIS DE LA VARIANZA **ANOBA**

Del mismo modo que la t de Student, la prueba **ANOVA** es una prueba paramétrica y como tal requiere una serie de supuestos para poder ser aplicada correctamente. Denominada ANOVA o análisis de la varianza, en realidad nos va a servir no solo para estudiar las dispersiones o varianzas de los grupos, sino para estudiar sus medias y la posibilidad de crear subconjuntos de grupos con medias iguales. Se puede decir que la prueba ANOVA es la generalización de la t de Student, ya que si realizamos una prueba ANOVA en la comparación de solo dos grupos, obtenemos los mismos resultados.

Al igual que la t de Student, se requiere que cada uno de los grupos a comparar tenga distribuciones normales, o lo que es más exacto, que lo sean sus residuales. Los residuales son las diferencias entre cada valor y la media de su grupo. Además debemos estudiar la dispersión o varianzas de los grupos, es decir estudiar su homogeneidad. Cuando mayor sean los tamaños de los grupos, menos importante es asegurar estos dos supuestos, ya que el **ANOVA** suele ser una técnica bastante "robusta" comportándose bien respecto a transgresiones de la normalidad. No obstante, si tenemos grupos de tamaño inferior a 30, es importante estudiar la normalidad de los residuos para ver la conveniencia o no de utilizar el análisis de la varianza. Si no fuera posible utilizar directamente el **ANOVA**, podemos recurrir al uso de pruebas no paramétricas, como la de *Kruskal-Wallis.*

Como ya hemos dicho, el **ANOVA** es la generalización de la t de Student, y sus hipótesis nula y alternativa se pueden formular del siguiente modo:

· **Hipótesis nula** (H_o): $\mu_1 = \mu_2 = \ldots = \mu_k$

Las medias de los k grupos son iguales y por tanto las diferencias encontradas pueden explicarse por el azar. Dicho de otro modo, los grupos proceden de poblaciones con medias iguales.

· **Hipótesis alternativa** (H_1): al menos uno de los grupos tiene una media distinta del resto de grupos.

En la prueba **ANOVA** las comparaciones son siempre bilaterales (a dos colas) ya que estudiamos globalmente si los grupos tienen medias distintas, y no si un grupo tiene una media menor o mayor que otro por separado. Si se rechaza la hipótesis nula, no sabremos entre qué grupos están las diferencias.

El análisis de la varianza permite contrastar la hipótesis nula de que las medias de K poblaciones (K >2) son iguales, frente a la hipótesis alternativa de que por lo menos una de las poblaciones difiere de las demás en cuanto a su valor esperado. Este contraste es fundamental en el análisis de resultados experimentales, en los que interesa comparar los resultados de K 'tratamientos' o 'factores' con respecto a la variable dependiente o de interés.

$$H_0\colon\ \mu_1 = \mu_2 = \ldots = \mu_K = \mu$$
$$H_1\colon\ \exists \mu_j \neq \mu \quad j = 1, 2, \ldots, K$$

El **Anova** requiere el cumplimiento los siguientes supuestos:

- Las poblaciones (distribuciones de probabilidad de la variable dependiente correspondiente a cada factor) son normales.
- Las K muestras sobre las que se aplican los tratamientos son independientes.
- Las poblaciones tienen todas igual varianza (homoscedasticidad).

El **ANOVA** se basa en la descomposición de la variación total de los datos con respecto a la media global (SCT), que bajo el supuesto de que H0 es cierta es una estimación de obtenida a partir de toda la información muestral, en dos partes:

- Variación dentro de las muestras (SCD) o Intra-grupos, cuantifica la dispersión de los valores de cada muestra con respecto a sus correspondientes medias.
- Variación entre muestras (SCE) o Inter-grupos, cuantifica la dispersión de las medias de las muestras con respecto a la media global.

Las expresiones para el cálculo de los elementos que intervienen en el **Anova** son las siguientes:

Media Global: $\overline{X} = \frac{\sum_{j=1}^{K} \sum_{i=1}^{nj} x_{ij}}{n}$.

Variación Total: $SCT = \sum_{j=1}^{K} \sum_{i=1}^{nj} (x_{ij} - \overline{X})^2$

Variación Intra-grupos: $SCD = \sum_{j=1}^{K} \sum_{i=1}^{nj} (x_{ij} - \overline{X_j})^2$

Variación Inter-grupos: $SCE = \sum_{j=1}^{K} (\overline{X_j} - \overline{X})^2 n_j$

Siendo X_{ij} el i-ésimo valor de la muestra j-ésima; nj el tamaño de dicha muestra y $\overline{X}_j$ su media.

Cuando la hipótesis nula es cierta SCE/K-1 y SCD/n-K son dos estimadores insesgados de la varianza poblacional y el cociente entre ambos se distribuye según una F de Snedecor con K-1 grados de libertad en el numerador y N-K grados de libertad en el denominador. Por lo tanto, si H_0 es cierta es de esperar que el cociente entre ambas estimaciones será aproximadamente igual a 1, de forma que se rechazará H_0 si dicho cociente difiere significativamente de 1.

La secuencia para realizar un ANOVA es:

Analizar

Comparar medias