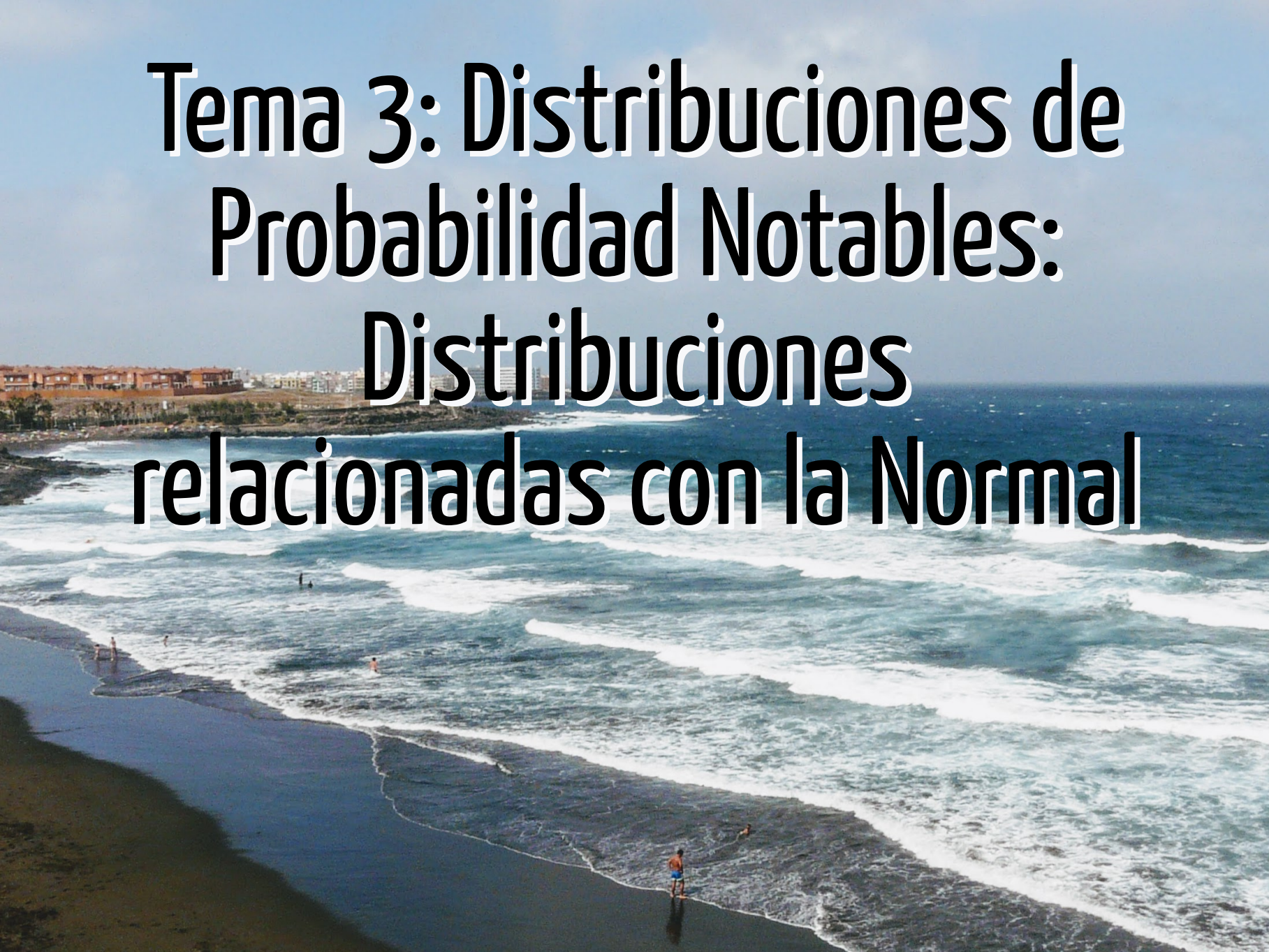
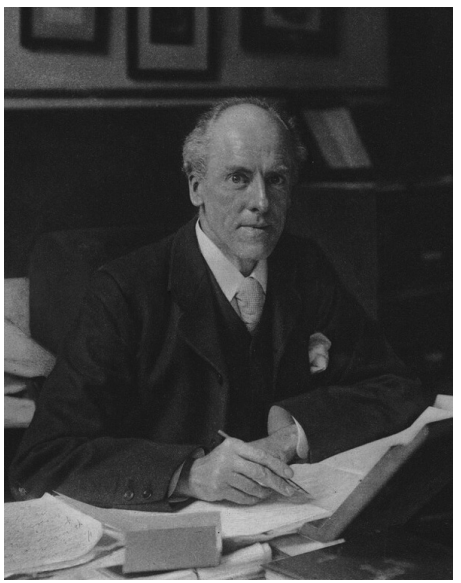


Tema 3: Distribuciones de Probabilidad Notables: Distribuciones relacionadas con la Normal

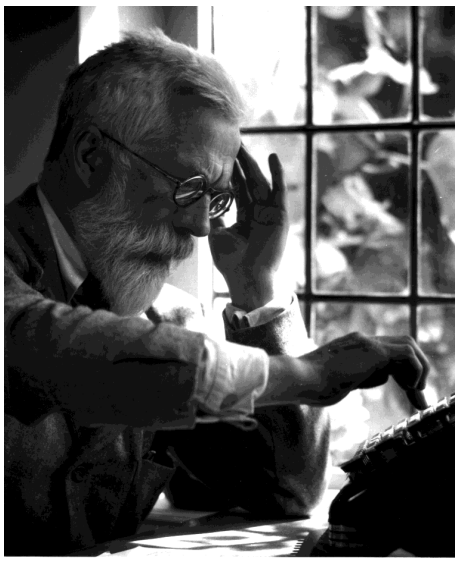
Distribuciones relacionadas con la normal

- t de Student t_n: William S. Gosset (*Student*)
- Chi-cuadrado χ^2_n: Karl Pearson
- F de Fisher F_{n_1,n_2}: Ronald A. Fisher

W. Gosset (1876-1937)



K. Pearson (1857-1936)



R. Fisher (1890-1962)

Distribución t de Student: t_n

Si $\bar{X}$ y s son, respectivamente, la media y desviación típica de una muestra de n observaciones de una variable $X \approx N(\mu, \sigma)$ se cumple que:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx t_{n-1}$$

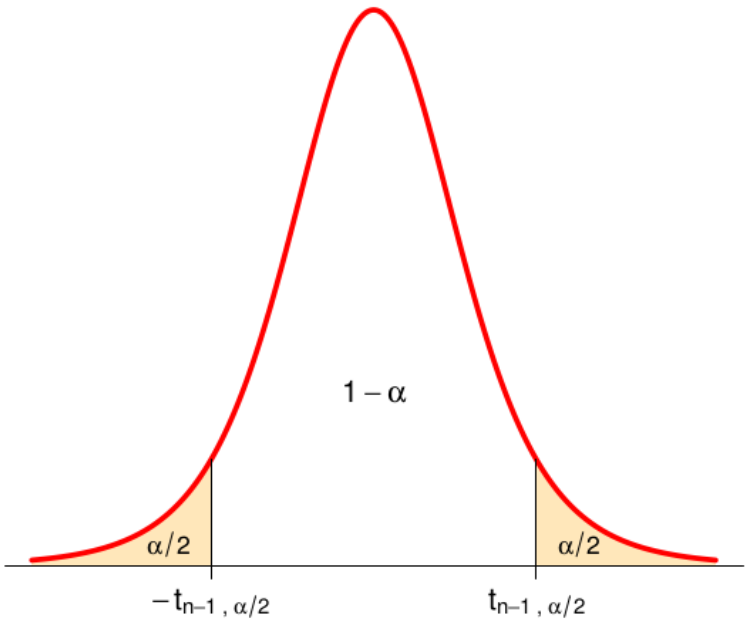
Utilizando la aplicación, podemos encontrar el valor $t_{n,\alpha/2}$ tal que $P(t_n > t_{n,\alpha/2}) = \alpha/2$, de tal forma que:

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1 - \alpha$$

Ver el trabajo original de Gosset en el que se obtiene la distribución t

Distribución t de Student: t_n

Gráficamente:



$$P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1-\alpha$$

El resultado anterior puede expresarse también de la forma:

$$P\left(|\bar{X}-\mu| \leq t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

Una vez tomados los datos, ya $\bar{X}$ no es una variable aleatoria sino un valor fijo, y la expresión anterior suele enunciarse diciendo que:

Con una confianza $1-\alpha$ podemos asegurar que la media poblacional μ se diferencia de la media muestral $\bar{X}$ en menos de $t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$ unidades

o dicho de otra forma:

Con una confianza $1-\alpha$ podemos asegurar que la media poblacional μ se encuentra en el intervalo $\left[\bar{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right]$

Distribución Chi-Cuadrado de Pearson: χ^2_n

Sea s^2 la varianza de una muestra de n observaciones de una variable $X \approx N(\mu, \sigma)$. *Antes de tomar la muestra,* s^2 es una variable aleatoria y se cumple que:

$$\frac{(n-1)s^2}{\sigma^2} \approx \chi^2_{n-1}$$

Utilizando la aplicación, podemos encontrar los valores $\chi_{n-1,\alpha/2}$ y $\chi_{n-1,1-\alpha/2}$ tales que

$$P(\chi_{n-1} > \chi_{n-1,\alpha/2}) = \alpha/2$$

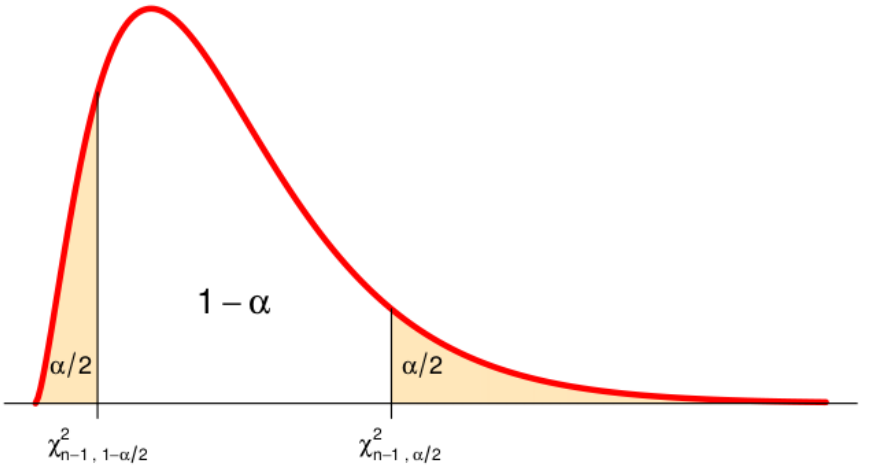
$$P(\chi_{n-1} > \chi_{n-1,1-\alpha/2}) = 1 - \alpha/2$$

de tal forma que:

$$P\left(\chi_{n-1,1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}\right) = 1 - \alpha$$

Distribución Chi-Cuadrado de Pearson: χ^2_n

Gráficamente:



$$P\left(\chi_{n-1,1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}\right) = 1-\alpha$$

Distribución Chi-Cuadrado de Pearson: χ^2_n

La expresión:

$$P\left(\chi_{n-1,1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{n-1,\alpha/2}\right) = 1-\alpha$$

puede expresarse de forma equivalente como:

$$P\left(\frac{\chi^2_{n-1,1-\alpha/2}}{(n-1)s^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{n-1,\alpha/2}}{(n-1)s^2}\right) = 1-\alpha$$

o también, invirtiendo las fracciones:

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha$$

Distribución Chi-Cuadrado de Pearson: χ^2_n

Una vez tomada la muestra s^2 no es una variable aleatoria, sino un valor fijo. En este momento, la expresión anterior se interpreta diciendo que:

Con una confianza $1-\alpha$ *podemos asegurar que la varianza* σ^2 *de la población se encuentra en el intervalo:*

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

Distribución F de Fisher: F_{n_1,n_2}

Supongamos que se van a tomar dos **muestras aleatorias independientes** de tamaños respectivos n_1 y n_2, de dos distribuciones normales con varianzas respectivas σ_1^2 y σ_2^2. Sean s_1^2 y s_2^2 las varianzas de estas muestras. *Antes de realizar el muestreo* s_1^2/s_2^2 *es una variable aleatoria que cumple*:

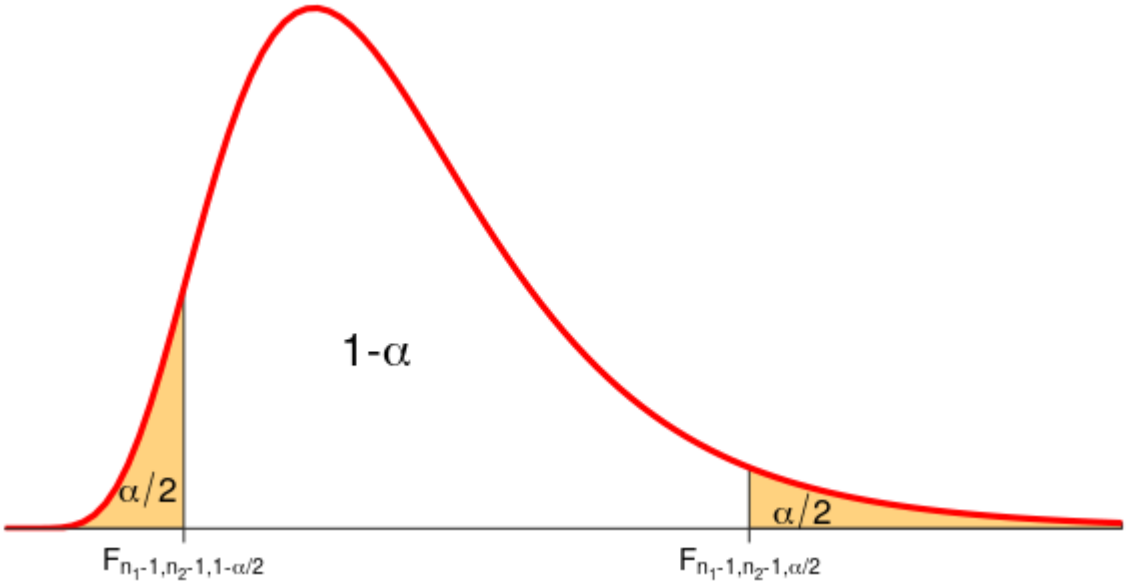
$$\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \approx F_{n_1-1,n_2-1}$$

de donde se sigue que:

$$P\left(F_{n_1-1,n_2-1,1-\alpha/2} \leq \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \leq F_{n_1-1,n_2-1,\alpha/2}\right) = 1-\alpha$$

Distribución F de Fisher: F_{n_1,n_2}

Gráficamente:



$$P\left(F_{n_1-1,n_2-1,1-\alpha/2} \leq \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \leq F_{n_1-1,n_2-1,\alpha/2}\right) = 1-\alpha$$

Esta expresión también puede escribirse como:

$$P\left(\frac{1}{F_{n_1-1,n_2-1,\alpha/2}} \leq \frac{\sigma_1^2/\sigma_2^2}{s_1^2/s_2^2} \leq \frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right) = 1-\alpha$$

o lo que es lo mismo:

$$P\left(\frac{s_1^2/s_2^2}{F_{n_1-1,n_2-1,\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2/s_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right) = 1-\alpha$$

Una vez tomada la muestra, $\frac{s_1^2}{s_2^2}$ no es una variable aleatoria sino un valor fijo, y el intervalo anterior se interpreta diciendo que:

Con probabilidad $1-\alpha$ *el cociente* $\frac{\sigma_1^2}{\sigma_2^2}$ *se encuentra comprendido en el intervalo*:

$$\left[\frac{s_1^2/s_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{s_1^2/s_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right]$$

Ejemplo

Se ha desarrollado un nuevo pienso para alimentar a peces criados en cultivos marinos. Para valorar la eficiencia de este pienso se han seleccionado aleatoriamente 40 alevines de similares características, se han pesado, se han marcado para poder identificarlos y se han devuelto al tanque de cultivo. Transcurridos seis meses, los 40 peces son vueltos a pesar y se anota para cada uno de ellos el incremento de peso experimentado durante este periodo. Los incrementos de peso (en gramos) registrados fueron los siguientes:

1226	988	1326	1246	1346	1258	1216	925	1156	1158
1328	988	1277	1242	1141	939	1179	1370	1139	1321
1331	1356	1087	1257	1156	1402	1209	1020	1108	1060
1109	1144	1143	1301	1308	1537	1511	1061	981	1306

- Media: $\bar{X} = 1203.9$
- Desviación típica: $s = 146.91$

Ejemplo

1. Con la información aportada por este experimento, ¿en qué intervalo podemos esperar que se encuentre el incremento medio de peso en la población de peces? (calcula el intervalo con una confianza del 95%)

El intervalo en este caso es:

$$\left[\bar{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right] = \left[1203.9 \pm t_{39,0.025}\frac{146.91}{\sqrt{40}}\right]$$

Mediante la aplicación podemos obtener $t_{39,0.025} = 2.0227$, y por tanto:

$$t_{39,0.025}\frac{146.91}{\sqrt{40}} = 46.98$$

Sustituyendo obtenemos el intervalo:

$$[1156.92, 1250.88]$$

Por tanto, con una confianza del 95% el incremento medio de peso en la población de peces está entre 1156.92 y 1250.88 gramos.

Ejemplo

2. Con la información aportada por este experimento, ¿en qué intervalo podemos esperar que se encuentre la desviación típica del peso en la población de peces? (calcula el intervalo con una confianza del 95%)

El intervalo a utilizar ahora es:

$$\sigma^2 \in \left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \right] = \left[\frac{39 \cdot 146.91^2}{\chi^2_{39,0.025}}, \frac{39 \cdot 146.91^2}{\chi^2_{39,0.975}} \right]$$

Mediante la aplicación obtenemos:

$$\chi^2_{39,0.025} = 58.12 \qquad \chi^2_{39,0.975} = 23.65$$

y sustituyendo:

$$\sigma^2 \in [14482.62, 35591.10]$$

Tomando raíces cuadradas:

$$\sigma \in [120.34, 188.66]$$

Ejemplo

3. Se ha realizado un experimento similar con otros 50 peces, utilizando otro pienso con un mayor contenido en proteínas e hidratos de carbono. En este segundo experimento la desviación típica observada en el incremento de peso fue de 205.62 gramos. Con esta información ¿Contamos con evidencia suficiente para asegurar que los dos piensos producen distinta variabilidad en la ganancia en peso de los peces que los consumen?

Para responder a esta pregunta calculamos un intervalo para el cociente de las varianzas observadas en ambos experimentos:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{s_1^2/s_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{s_1^2/s_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}} \right] = \left[\frac{146.91^2/205.62^2}{F_{39,49,0.025}}, \frac{146.91^2/205.62^2}{F_{49,39,0.975}} \right]$$

Mediante la aplicación obtenemos:

$$F_{39,49,0.025} = 1.8082 \qquad F_{39,49,0.975} = 0.5416$$

y sustituyendo:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{0.51}{1.8082}, \frac{0.51}{0.5416} \right] = [0.28, 0.94]$$

Ejemplo

Como el intervalo obtenido es:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{0.51}{1.8082}, \frac{0.51}{0.5416}\right] = [0.28, 0.94]$$

podemos estamos "seguros" con un 95% de confianza de que el cociente de estas varianzas es algún valor entre 0.28 y 0.94, esto es, **un valor menor que 1**. De esta forma podemos decir (con esa confianza) que el experimento contiene evidencia suficiente para afirmar que la ganancia de peso con el pienso 1 presenta menor variabilidad que con el pienso 2.

Por último, señalemos que si deseáramos un intervalo de confianza para el cociente de desviaciones típicas, bastaría con tomar raíces cuadradas en la expresión anterior:

$$\frac{\sigma_1}{\sigma_2} \in \left[\sqrt{\frac{0.51}{1.8082}}, \sqrt{\frac{0.51}{0.5416}}\right] = [0.53, 0.97]$$

es decir, el valor de σ_1 es entre un 53% y un 97% del valor de σ_2.

Teorema Central del Límite: Aplicaciones

El Teorema Central del Límite establece que dada una colección de variables aleatorias **independientes** $X_1, X_2, \ldots, X_n$ tales que $E[X_i] = \mu_i$ y $Var(X_i) = \sigma_i^2$, cuando $n \to \infty$ la distribución de probabilidad de la suma de estas variables es aproximadamente normal:

$$\sum_{i=1}^{n} X_i \approx N\left(\sum_{i=1}^{n} \mu_i,\ \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

En el caso particular de que todas las X_i tengan la misma distribución, esto es, $E[X_i] = \mu$ y $Var(X_i) = \sigma^2\ \forall i$ se tiene que:

$$\sum_{i=1}^{n} X_i \approx N\left(n\mu,\ \sigma\sqrt{n}\right)$$

Aplicaciones del TCL: aproximación de la Binomial por la Normal:

Si X es una variable $B(n,p)$, su valor representa el número de éxitos en n experimentos independientes en cada uno de los cuales la probabilidad de éxito es p.

Si definimos el resultado de cada experimento como:

$$X_i = \begin{cases} 0 & 1-p \ \text{(fracaso)} \\ 1 & p \qquad \text{(éxito)} \end{cases} \qquad \text{(Variable de Bernoulli)}$$

podemos expresar la binomial como suma de variables de Bernoulli:

$$X = X_1 + X_2 + \cdots + X_n$$

Obsérvese que:

$$\mu_i = E[X_i] = 1 \cdot p + 0 \cdot (1-p) = p$$

$$\sigma_i^2 = Var(X_i) = E[X_i^2] - \left(E[X_i]\right)^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = p(1-p)$$

Aplicaciones del TCL: aproximación de la Binomial por la Normal

Entonces, si el valor de n es grande:

$$B(n,p) \approx X = \sum_{i=1}^{n} X_i \underset{n\to\infty}{\approx} N\left(n\mu,\ \sigma\sqrt{n}\right) = N\left(np, \sqrt{np(1-p)}\right)$$

- En general la aproximación es razonablemente buena cuando $n \geq 30$, $np \geq 5$ y $n(1-p) \geq 5$
- Como la normal es continua y la binomial es discreta, en el cálculo aproximado se considera que el valor (discreto) k es equivalente al intervalo $\left[k - \frac{1}{2}, k + \frac{1}{2}\right]$

Ejemplo:

Si $X \approx B(120, 0.35)$, se puede aproximar por

$$X_N \approx N(120 \cdot 0.35, \sqrt{120 \cdot 0.35 \cdot 0.65} = N(42, 5.2249)$$

Entonces:

- $P(X = 40) = 0.0716$ (Valor exacto)
- $P(X = 40) \cong P(39.5 < X_N < 40.5) = P(X_N < 40.5) - P(X_N < 39.5) =$
  $= 0.387 - 0.3162 = 0.0709$ (Valor aproximado)
- $P(X \leq 40) = 0.3905$ (Valor exacto)
- $P(X \leq 40) \cong P(X_N \leq 40.5) = 0.387$ (Valor aproximado)
- $P(X \geq 40) = 0.6811$ (Valor exacto)
- $P(X \geq 40) \cong P(X_N \geq 39.5) = 0.6838$ (Valor aproximado)

Aproximación de la Binomial por la Normal

Supongamos que $X \approx B(n,p)$ es el número de éxitos en n pruebas independientes; si realizáramos efectivamente este experimento, $\hat{p} = \frac{X}{n}$ sería la **proporción observada** de éxitos en esas n pruebas. Si n es grande:

$$X \approx N\left(np, \sqrt{np\left(1-p\right)}\right)$$

y por tanto:

$$\hat{p} = \frac{X}{n} \approx N\left(\frac{np}{n}, \frac{\sqrt{np\left(1-p\right)}}{n}\right) = N\left(p, \sqrt{\frac{p\left(1-p\right)}{n}}\right)$$

De aquí se sigue que:

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \approx N\left(0,1\right)$$

Aproximación de la Binomial por la Normal

Utilizando la distribución normal $N(0,1)$, podemos encontrar el valor $z_{\alpha/2}$ tal que:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = 1-\alpha$$

De aquí se deduce que la diferencia entre el valor (desconocido) de p y el valor (observado) $\hat{p}$ en la muestra cumple:

$$P\left(|\hat{p}-p| \leq z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

Este resultado es aproximado, y sólo es válido si $n \geq 30$, $np \geq 5$ y $n(1-p) \geq 5$.

Aproximación de la Binomial por la Normal

La probabilidad anterior puede expresarse también como:

$$P\left(|p-\hat{p}| \leq z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

o lo que es lo mismo:

$$P\left(-z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}} \leq p-\hat{p} \leq z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

y de aquí:

$$P\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}} \leq p \leq \hat{p}+z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

esto es,

$$P\left(p \in \left[\hat{p}-z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}, \hat{p}+z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right]\right) = 1-\alpha$$

Aproximación de la Binomial por la Normal

Una vez realizado el experimento, $\hat{p}$ *no es una variable aleatoria, sino un valor fijo*. Podemos decir entonces que tenemos una confianza $1-\alpha$ en que el valor (desconocido) de p cae en el intervalo:

$$p \in \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right]$$

Ahora bien, este intervalo es poco útil en la práctica, ya que sus extremos dependen de p, que es desconocido. Una opción es sustituirlo por el valor observado $\hat{p}$:

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right]$$

(*intervalo de Wald*) aunque, como es obvio, no podemos garantizar entonces que se consigue la confianza deseada $1-\alpha$. **Por ello no es recomendable utilizar este intervalo en la práctica**.

Aproximación de la Binomial por la Normal

En Agresti & Caffo, 2000 se señala como la aproximación del intervalo de Wald anterior mejora notablemente si se añaden 4 pseudo-observaciones a la muestra, 2 éxitos y 2 fracasos. De esta forma:

- n se sustituye por $\tilde{n} = n + 4$
- El número de éxitos X se sustituye por $X + 2$. Por tanto la proporción observada de éxitos $\hat{p}$ se sustituye por $\tilde{p} = \frac{X+2}{n+4}$
- El intervalo ajustado para p (*intervalo de Agresti-Coull*) es entonces:

$$\left[\tilde{p} - z_{\alpha/2}\sqrt{\frac{\tilde{p}\,(1-\tilde{p})}{\tilde{n}}}, \tilde{p} + z_{\alpha/2}\sqrt{\frac{\tilde{p}\,(1-\tilde{p})}{\tilde{n}}}\right]$$

Ejemplo

Se desea conocer la proporción de hembras en una población de peces. Con este fin se obtiene una muestra de 200 peces elegidos aleatoriamente en esta población. En la muestra 140 peces son hembras. ¿Cuál es la proporción de hembras en esa población?

La proporción de hembras en la muestra es $\frac{140}{200} = 0.7 \cong 70\%$. Para evaluar el margen de error con que la proporción de hembras en la población se aproxima a este valor calculamos el intervalo de Agresti-Coull:

- $\widetilde{n} = 200 + 4$
- $\tilde{p} = \frac{142}{204} = 0.6961$
- $\left[\tilde{p} - z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\widetilde{n}}}, \tilde{p} + z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\widetilde{n}}}\right] = [0.6961 \pm 1.96 \cdot 0.0322] =$

$$= [0.633, 0.759]$$

Por tanto a partir de estos datos podemos tener una confianza aproximada del 95% en que la proporción de hembras en la población es un valor comprendido entre el 63.3% y el 75.9%.