Capítulo 5

Inferencia Estadística II: Estimación por Intervalos de Confianza.

5.1. Introducción.

En el capítulo anterior hemos visto cómo podemos obtener un estimador puntual para un parámetro de una distribución de probabilidad. Si se dan las condiciones adecuadas (error cuadrático medio pequeño, tamaño de muestra suficiente) sabemos que el estimador, al ser evaluado sobre distintas muestras, va a producir valores distintos pero siempre próximos al valor del parámetro que se pretende estimar. Ahora bien, en la práctica, una vez que hemos obtenido la muestra, tenemos un solo valor del estimador, pero ¿cuál es el grado de precisión alcanzado en la estimación? ¿Cuánto se parece este valor estimado al verdadero valor del parámetro? En este capítulo aprenderemos a construir intervalos que podemos confiar en que contienen al parámetro desconocido. La amplitud de estos intervalos, como veremos, nos informa de la precisión alcanzada en la estimación.

Objetivos.

Al finalizar este capítulo el alumno deberá:

1. Conocer y comprender el concepto de intervalo de confianza.
2. Entender la necesidad de acompañar la estimación de parámetros de la estimación de su error estándar y su intervalo de confianza.
3. Ser capaz de calcular los intervalos de confianza más frecuentes en la práctica.

4. Ser capaz de deducir intervalos de confianza a partir de funciones pivote.

5. Ser capaz de deducir intervalos de confianza asintóticos para los estimadores de máxima verosimilitud de una distribución arbitraria.

5.2. Definición de intervalo de confianza.

Dado un parámetro desconocido θ, que caracteriza la distribución de probabilidad de una variable aleatoria determinada, y dada una muestra aleatoria $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$ de dicha variable, diremos que un intervalo de la forma $[\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]$, donde $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$ son variables aleatorias que dependen de la muestra, es un *intervalo de confianza a nivel* $1-\alpha$ *para el parámetro* θ si la probabilidad de que el intervalo contenga a dicho parámetro es $1-\alpha$, esto es:

$$P(\theta \in [\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]) = 1 - \alpha$$

De esta forma, si disponemos de un intervalo de confianza para un parámetro θ desconocido, ya no nos limitaremos a decir que θ tiene un valor parecido a $\hat{\theta}$ (su estimador puntual), sino que además podemos afirmar que con probabilidad $1-\alpha$ (donde α es en general un valor pequeño) el valor de θ se encuentra entre $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$. Ello nos da una idea aproximada de la precisión conseguida en la estimación. Nótese que en la definición de intervalo de confianza, los extremos $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$ son variables aleatorias ya que son funciones de la muestra y ésta es aleatoria. Ello significa que muestras distintas de la misma población producirán intervalos de confianza distintos.

5.3. Intervalo de confianza para la esperanza de una variable $X \approx N(\mu, \sigma)$ con σ conocida.

Supongamos que se desea estimar la esperanza μ de una variable X con distribución normal de varianza σ^2 conocida[1]. Aquí X podría ser el peso que alcanzan los peces de un cultivo marino cuando se les alimenta con cierta dieta experimental, la concentración de un contaminante en la boca de un emisario, el peso mensual de las capturas de una flota, o cualquier otra variable cuya distribución de probabilidad pueda razonablemente considerarse normal.

[1]Debemos confesar que, en la práctica, la varianza σ^2 no se conoce nunca, por lo que el intervalo que vamos a construir carece de interés práctico; no obstante, resulta simple e ilustrativo para entender el concepto y modo de construcción de estos intervalos.

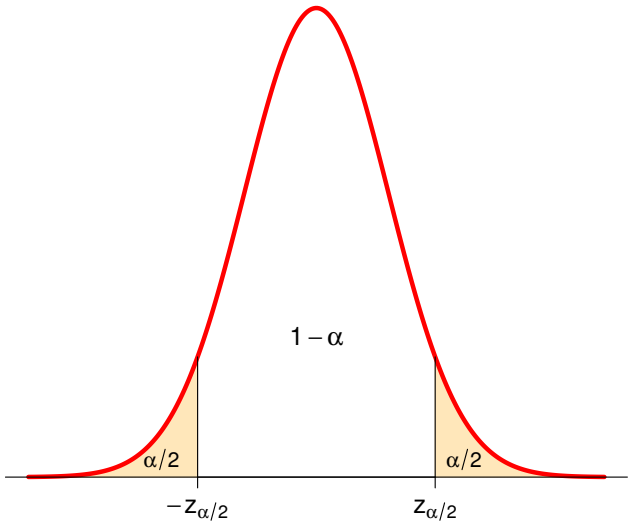

Figura 5.1: Función de densidad de la distribución normal estándar $N(0,1)$. La zona sombreada encierra un área $1-\alpha$. El percentil $z_{\alpha/2}$ es el valor que deja a su derecha un área $\alpha/2$, esto es, $P\left(Z > z_{\alpha/2}\right) = \alpha/2$, por lo que $P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1-\alpha$

En el tema 3 ya hemos visto que, debido a la propiedad reproductiva de la distribución normal, si $\bar{X}$ es la media aritmética de n variables independientes $X_i \approx N(\mu, \sigma)$ entonces:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

Por tanto, si $z_{\alpha/2}$ es el percentil $1-\alpha/2$ de la distribución normal estándar $N(0,1)$ (véase figura 5.1), se tiene que:

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

de donde:

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

que, tras reordenar términos puede escribirse como:

$$P\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

o, de modo análogo:

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

Por tanto, de acuerdo con la definición dada más arriba, el intervalo $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ es un *intervalo de confianza a nivel* $1-\alpha$ para el parámetro μ.

Aplicación a una muestra particular: Se dispone de 20 peces de un cultivo que han estado consumiendo una dieta experimental durante los cuatro últimos meses. Cada pez fue pesado al inicio y al final de este periodo. Los incrementos de peso (en gramos) observados fueron los siguientes:

402	308	261	357	425	378	457	345	372	321
305	370	293	439	363	392	417	452	291	244

Suponiendo que el incremento de peso X experimentado por cada pez en estas condiciones sigue una distribución $N(\mu, \sigma)$, siendo $\sigma = 60$, se desea construir un intervalo de confianza al 95 % para μ.

Para ello basta tener en cuenta que como la confianza buscada es $1-\alpha = 0{,}95$, entonces $\alpha = 0{,}05$ y utilizando la tabla de la $N(0,1)$ encontramos $z_{\alpha/2} = z_{0{,}025} = 1{,}96$. La media aritmética de los 20 valores anteriores es 359.6 gramos, y el intervalo de confianza sería entonces:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[359{,}6 - 1{,}96\frac{60}{\sqrt{20}}, 359{,}6 + 1{,}96\frac{60}{\sqrt{20}}\right] = [333{,}3,\ 385{,}9]$$

Por tanto, con un 95 % de confianza podemos esperar que el incremento medio de peso μ que se obtiene con la citada dieta experimental sea un valor comprendido entre 333.3 y 385.9 gramos.

Cálculo con R : R no incluye ninguna función específica para calcular este intervalo (ya que en la práctica no se presenta nunca una situación en la que se conozca la desviación típica de la población). No obstante, este intervalo de confianza puede calcularse de manera muy sencilla:

```
> incPeso = c(402, 308, 261, 357, 425, 378, 457, 345, 372, 321,
    305, 370, 293, 439, 363, 392, 417, 452, 291, 244)
> sigma = 60
> za2 = qnorm(0.975)
> n = length(incPeso)
> intervalo = mean(incPeso) + c(-1, 1) * za2 * sigma/sqrt(n)
> intervalo

[1] 333.3043 385.8957
```

5.4. Interpretación del intervalo de confianza: ¿por qué el término "confianza"?

Para la determinación del intervalo de confianza que hemos visto en el ejemplo anterior, nos apoyamos en el hecho de que, *antes de obtener la muestra,* la media muestral $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *es una variable aleatoria* con distribución $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. A partir de aquí hemos deducido que:

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

Por tanto, *mientras no se haya obtenido la muestra,* los extremos del intervalo son variables aleatorias y se puede calcular la probabilidad de que dicho intervalo contenga a μ. Ahora bien, una vez que se ha obtenido una muestra, los extremos del intervalo son valores fijos, como 333.3 y 385.9 en el ejemplo anterior. En este momento, el valor de μ estará comprendido entre ellos o no, pero ya no cabe hablar de la probabilidad de que ésto ocurra.

Podemos utilizar el símil del lanzador de cuchillos circense que se dispone a lanzar un cuchillo contra una diana con los ojos vendados. Él sabe, por su experiencia, que la probabilidad de acertar en la diana es del 95 %. Ahora bien, una vez que ha lanzado el cuchillo habrá acertado o no, pero ya no se puede hablar de la probabilidad de que acierte. Si el lanzador continúa con los ojos vendados tras el lanzamiento, *puede confiar* en que ha acertado (incluso, tener mucha confianza en ello, ya que sabe que tiene muy buena puntería), pero no puede estar del todo seguro.

La situación de un investigador que construye un intervalo de confianza a partir de unos datos experimentales es análoga a la del lanzador de cuchillos que nunca se quita la venda de los ojos: *antes de tomar la muestra* sabe que la probabilidad de que el intervalo contenga al parámetro es del 95 %; por tanto, cuando tome los datos y obtenga un intervalo concreto, puede tener mucha confianza (que puede valorar en ese mismo 95 %) en que el intervalo habrá

"*capturado*" al parámetro, pero no puede saber con seguridad si lo ha capturado o no, ya que el valor del parámetro sigue siendo desconocido.

De un modo más general, si para un parámetro θ de una distribución de probabilidad disponemos de dos estadísticos $\theta_1(\mathfrak{X})$ y $\theta_2(\mathfrak{X})$ tales que:

$$P(\theta \in [\theta_1(\mathfrak{X}), \theta_2(\mathfrak{X})]) = 1 - \alpha$$

siendo $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de dicha distribución, entonces cabe esperar que el $100(1-\alpha)$ % de los intervalos construidos de esta manera contengan a θ y, obviamente, que el restante 100α % no lo contengan. Una vez que obtenemos una muestra particular $(x_1, x_2, \ldots, x_n)$ y calculamos los valores $\hat{\theta}_1 = \theta_1(x_1, x_2, \ldots, x_n)$ y $\hat{\theta}_2 = \theta_2(x_1, x_2, \ldots, x_n)$, tenemos un intervalo concreto $\left[\hat{\theta}_1, \hat{\theta}_2\right]$. En realidad *no sabemos* si este intervalo contiene o no a θ, pero *confiamos* en que sea uno de entre el $100(1-\alpha)$ % de intervalos que contienen al parámetro. De ahí que valoremos nuestra confianza en $1-\alpha$.

El siguiente código en R simula la obtención de 1000 muestras de tamaño 100 de una variable aleatoria $X \approx N(\mu = 10, \sigma = 2)$. Para cada muestra se calculan la media muestral $\bar{X}$ y el intervalo de confianza para μ obtenido en la sección anterior, calculado de acuerdo con la expresión $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$, siendo $\sigma = 2$ y $1 - \alpha = 0{,}95$:

```
> simulaMuestreo = function(n) {
    muestra = rnorm(n, 10, 2)
    intervalo = mean(muestra) + c(-1, 1) * qnorm(0.975) * 2/sqrt(n)
    return(intervalo)
 }
> intervalos = t(replicate(1000, simulaMuestreo(100)))
```

Mostramos los primeros 10 intervalos:

```
> intervalos[1:10, ]
          [,1]      [,2]
 [1,]  9.214422  9.998408
 [2,]  9.868193 10.652178
 [3,]  9.692417 10.476403
 [4,]  9.546502 10.330488
 [5,]  9.560918 10.344904
 [6,]  9.514950 10.298936
 [7,]  9.672468 10.456454
 [8,] 10.120441 10.904426
 [9,]  9.728458 10.512444
[10,]  9.735197 10.519183
```

(obsérvese que en esta simulación particular el octavo intervalo no contiene a la media $\mu = 10$). Ahora contamos cuántos de los 1000 intervalos contienen a μ. Como hemos elegido una confianza del 95 %, esperamos que aproximadamente el 95 % de los intervalos (esto es, unos 950), contengan al parámetro:

```
> numinterv = 0
> for (k in 1:1000) if ((intervalos[k, 1] <= 10) & (10 <= intervalos[k,
      2])) numinterv = numinterv + 1
> numinterv

[1] 944
```

Como vemos, el 94.4 % (muy cerca del 95 %) de los intervalos contiene al parámetro, tal como esperábamos. Se invita al lector a copiar el código anterior y a repetir el experimento varias veces. Podrá comprobar que, efectivamente, en todos los casos el número de intervalos que contienen a la media está siempre en torno al 95 %.

La figura 5.2 representa los 100 primeros intervalos de confianza de la simulación anterior, La línea vertical corresponde al valor de $\mu = 10$. Como vemos, 94 de los intervalos cubren al parámetro y 6 (marcados en rojo) no lo contienen. Remarquemos una vez más, que en la práctica el investigador *toma una única muestra*, no 100 ni 1000. El investigador *confía* (con un nivel de confianza del 95 %) en haber capturado al parámetro. Pero, si ha ocurrido que esa única muestra le lleva a obtener un intervalo de los que se han marcado en rojo entonces, lamentablemente, el parámetro se le habrá escapado, sin que nuestro investigador tenga ningún medio de saberlo.

5.5. Método general de construcción de intervalos de confianza.

El procedimiento de construcción de un intervalo de confianza para un parámetro θ sigue en líneas generales los pasos dados en la sección anterior para obtener el intervalo de confianza para la media μ de una población normal de varianza σ conocida. Partiendo de una muestra aleatoria $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$:

1. Deberemos disponer de una *función pivote* $T(\theta, \mathfrak{X})$ cuya distribución de probabilidad sea conocida y no dependa de θ.

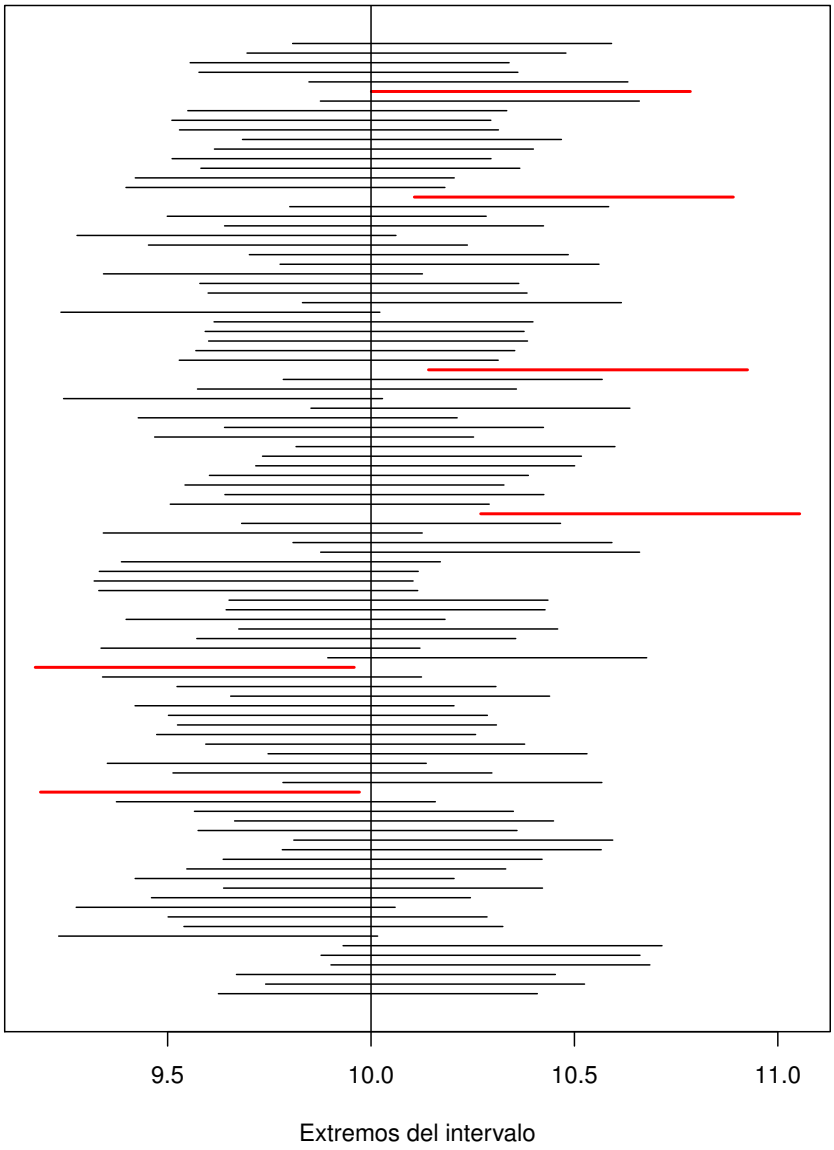

Figura 5.2: 100 intervalos de confianza al 95 % para el parámetro μ de una distribución normal de varianza conocida. En rojo los intervalos que *no* contienen a μ.

2. A partir del pivote y de su distribución de probabilidad deberán encontrarse dos valores $\tau_I(\alpha)$ y $\tau_S(\alpha)$ tales que:

$$P\left(\tau_I(\alpha) \leq T(\theta, X) \leq \tau_S(\alpha)\right) = 1 - \alpha$$

3. Si la función $T(\theta, \mathfrak{X})$ es monótona en θ, las ecuaciones:

$$T(\theta_I, X) = \tau_I(\alpha)$$
$$T(\theta_S, X) = \tau_S(\alpha)$$

tienen solución única. Si $\theta_I(\mathfrak{X}, \alpha)$ y $\theta_S(\mathfrak{X}, \alpha)$ son las respectivas soluciones de estas ecuaciones, se tiene que

$$P(\theta_I(\mathfrak{X}, \alpha) \leq \theta \leq \theta_S(\mathfrak{X}, \alpha)) = 1 - \alpha$$

por lo que el intervalo de confianza a nivel $1 - \alpha$ es $[\theta_I(\mathfrak{X}, \alpha), \theta_S(\mathfrak{X}, \alpha)]$

Ejemplo. Así, para estimar la media μ de una distribución normal de varianza conocida σ^2, la función pivote utilizada fue:

$$T(\mu, \mathfrak{X}) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

cuya distribución de probabilidad es $N(0, 1)$ (y por tanto no depende de μ). En este caso, $\tau_I(\alpha) = -z_{\alpha/2}$ y $\tau_S(\alpha) = z_{\alpha/2}$. Los extremos del intervalo se hallan resolviendo:

$$T(\mu_I, \mathfrak{X}) = \tau_I(\alpha) \Rightarrow \frac{\bar{X} - \mu_I}{\sigma/\sqrt{n}} = -z_{\alpha/2} \Rightarrow \mu_I = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$T(\mu_S, \mathfrak{X}) = \tau_S(\alpha) \Rightarrow \frac{\bar{X} - \mu_S}{\sigma/\sqrt{n}} = z_{\alpha/2} \Rightarrow \mu_s = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

5.6. Intervalo de confianza para la esperanza de una variable $X \approx N(\mu, \sigma)$ con σ desconocida.

Ya hemos visto en la sección 5.3 como contruir un intervalo de confianza para la media de una variable aleatoria con distribución normal de varianza conocida. Este intervalo en la práctica resulta de poca utilidad, toda vez que normalmente la varianza σ^2 es desconocida. Afortunadamente, es posible demostrar que si $X_1, X_2, \ldots$,X es una muestra aleatoria de una distribución $N(\mu, \sigma)$ entonces:

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

siendo $S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$ la desviación típica de la muestra.

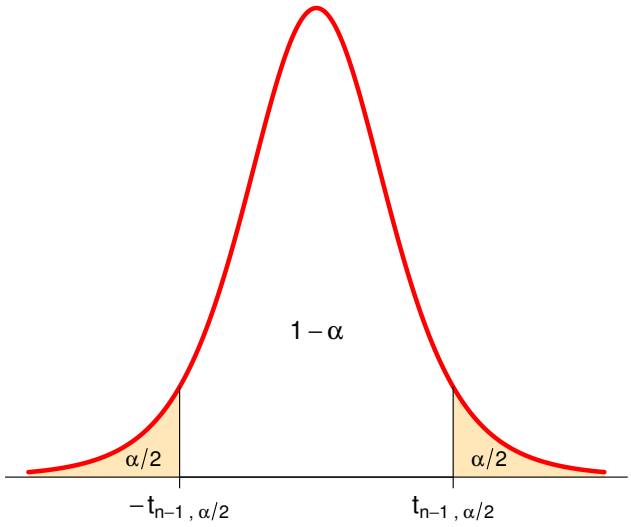

Figura 5.3: Posición de los percentiles $1-\alpha/2$ y $\alpha/2$ de la distribución t de Student denotados, respectivamente, como $t_{n-1,\alpha/2}$ y $-t_{n-1,\alpha/2}$. El área entre estos dos percentiles es $1-\alpha$.

Podemos ahora utilizar las tablas de la t de Student (o R) para encontrar el percentil $t_{n-1,\alpha/2}$ de esta distribución, de tal forma que

$$P\left(-t_{n-1,\alpha/2} \leq t_{n-1} \leq t_{n-1,\alpha/2}\right) = 1-\alpha$$

(ver figura 5.3). Podemos escribir entonces:

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

de donde, operando en el interior del intervalo:

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

o, expresado de otra forma:

$$P\left(\mu \in \left[\bar{X} - \frac{S}{\sqrt{n}}t_{n-1,\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}}t_{n-1,\alpha/2}\right]\right) = 1-\alpha$$

Así pues, el intervalo de confianza a nivel $1-\alpha$ para la media μ de una distribución $N(\mu,\sigma)$ con σ desconocida es

$$\left[\bar{X}-\frac{S}{\sqrt{n}}t_{n-1,\alpha/2}, \bar{X}+\frac{S}{\sqrt{n}}t_{n-1,\alpha/2}\right]$$

Aplicación a una muestra particular: Consideremos nuevamente los incrementos de peso (en gramos) observados en 20 peces de un cultivo cuando son alimentados con una dieta experimental:

402	308	261	357	425	378	457	345	372	321
305	370	293	439	363	392	417	452	291	244

Si el incremento de peso X experimentado por cada pez en estas condiciones sigue una distribución $N(\mu,\sigma)$, considerando ahora que σ es desconocida, para construir un intervalo de confianza al 95 % para μ, debemos buscar en la tabla de la t de Student el valor $t_{19,0,025} = 2{,}093$. Asimismo, calculamos :

$$s=\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}=\sqrt{\frac{\sum_{i=1}^{20}\left(X_i-359{,}6\right)^2}{19}}=62{,}8$$

El intervalo de confianza es entonces:

$$\left[359{,}6-\frac{62{,}8}{\sqrt{20}}\cdot 2{,}093,\ 359{,}6+\frac{62{,}8}{\sqrt{20}}\cdot 2{,}093, 4{,}8\right]=[359{,}6\pm 29{,}39]=[330{,}21,\ 388{,}99]$$

Por tanto podemos concluir, con una confianza del 95 %, que el incremento medio de peso (en gramos) obtenido en peces alimentados con la dieta experimental se encuentra en el intervalo [330,21, 388,99]; dicho de otro modo, podemos afirmar con una confianza del 95 % que el incremento medio de peso es aproximadamente de 359.6 gramos, con un margen de error de ±29,39 gramos.

Cálculo en R : en R el cálculo del intervalo de confianza es tan simple como escribir el comando:

```
> t.test(incPeso)
```

```
        One Sample t-test

data:  incPeso
t = 25.6066, df = 19, p-value = 3.42e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 330.2072 388.9928
sample estimates:
mean of x
    359.6
```

Como vemos, R proporciona aquí mucha más información que el intervalo de confianza. Además de calcular la media muestral de la variable (mostrada en la última linea) y el intervalo de confianza, R lleva a cabo un *contraste de hipótesis* sobre la media de la población. Explicaremos este concepto en el siguiente capítulo.

Nota: si deseamos que R calcule un intervalo con otro nivel de confianza, por ejemplo 0.9, utilizaríamos la opción *conf.level*:

```
> t.test(incPeso, conf.level = 0.9)
```

5.7. Intervalo de confianza para la varianza σ^2 de una población normal.

Ya hemos visto en el capítulo anterior que la varianza muestral:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

es un estimador centrado de la varianza de la variable aleatoria X cualquiera que sea su distribución de probabilidad. En el caso particular de que $X \approx N(\mu, \sigma)$, dada una muestra aleatoria $\{X_1, X_2, \ldots, X_n\}$ de X, es posible probar que:

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

Por tanto, utilizando la tabla de la distribución χ^2_{n-1} (o R) podemos encontrar los percentiles $\chi^2_{n-1,1-\alpha/2}$ y $\chi^2_{n-1,\alpha/2}$ (ver figura 5.4) para los que:

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

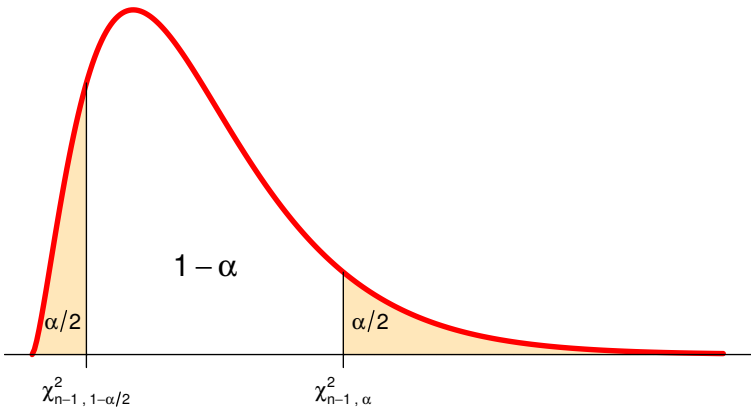

Figura 5.4: Posición de los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución χ^2_{n-1} (denotados, respectivamente, como $\chi^2_{n-1,1-\alpha/2}$ y $\chi^2_{n-1,\alpha/2}$). El área entre estos dos percentiles es $1-\alpha$.

Operando en el interior del intervalo podemos despejar σ^2:

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

Por tanto el intervalo de confianza a nivel $1-\alpha$ para la varianza de una variable aleatoria X con distribución normal $N(\mu,\sigma)$ es:

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

Aplicación a una muestra particular: Siguiendo con nuestro ejemplo del incremento de peso (en gramos) que se obtiene en peces alimentados con una dieta experimental, si deseamos calcular un intervalo de confianza al 95 % para la varianza de esta variable, asumiendo que sigue una distribución normal, y partiendo de la anterior muestra de $n = 20$ peces, en la tabla de la χ^2 encontramos los valores $\chi^2_{19,\,0{,}975} = 8{,}906$ y $\chi^2_{19,\,0{,}025} =$

32,852. La varianza muestral es:

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1} = \frac{\sum_{i=1}^{20} \left(X_i - 359{,}6\right)^2}{19} = 3944{,}25$$

Por tanto, el intervalo de confianza para σ^2 es:

$$\left[\frac{19 \cdot 3944{,}25}{32{,}852}, \frac{19 \cdot 3944{,}25}{8{,}906}\right] = [2281{,}16,\ 8414{,}64]$$

Si queremos calcular el intervalo para la desviación típica $\sigma = \sqrt{\sigma^2}$ basta con aplicar la raiz cuadrada a los extremos del intervalo anterior:

$$\left[\sqrt{\frac{19 \cdot 3944{,}25}{32{,}852}}, \sqrt{\frac{19 \cdot 3944{,}25}{8{,}906}}\right] = [47{,}76,\ 91{,}73]$$

Por tanto podemos concluir, con una confianza del 95 %, que la desviación típica del incremento de peso (en gramos) obtenido por peces alimentados con la nueva dieta experimental se encuentra en el intervalo [47,76, 91,73].

Cálculo en R : en R podemos calcular fácilmente un intervalo de confianza para la varianza del siguiente modo:

```
> n = length(incPeso)
> (n - 1) * var(incPeso)/qchisq(c(0.975, 0.025), n - 1)

[1] 2281.141 8414.154
```

(**Nota:** las diferencias que se observan con el intervalo calculado más arriba obedecen a que en aquel caso hemos utilizado los valores de la tabla de la χ^2, que están redondeados a 3 decimales, mientras que aquí R ha hecho el cálculo con mayor precisión).

En R podemos utilizar también la librería *TeachingDemos*, que implementa la función *sigma.test()* que también calcula el intervalo de confianza para la varianza de una población normal. Para utilizar esta librería debemos cargarla previamente:

```
> library(TeachingDemos)
> sigma.test(incPeso)

        One sample Chi-squared test for variance

data:  incPeso
```

```
X-squared = 74940.8, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
 2281.141 8414.154
sample estimates:
var of incPeso
      3944.253
```

Tal como ocurría también con `t.test()` esta función, además del intervalo de confianza para la varianza, también lleva a cabo un contraste de hipótesis, que se explicará en el siguiente capítulo.

5.8. Intervalo de confianza para el cociente de varianzas de poblaciones normales

En el capítulo 3 hemos visto que si Y_1 e Y_2 son $v$ariables aleatorias independientes con distribuciones de probabilidad respectivas $Y_1 \approx \chi^2_{n_1}$ e $Y_2 \approx \chi^2_{n_2}$, entonces:

$$\frac{Y_1/n_1}{Y_2/n_2} \approx F_{n_1,n_2}$$

Asimismo, en la sección anterior hemos visto también que:

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

Así pues, si se dispone de 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, llamando $Y_i = (n_i-1)S_i^2/\sigma_i^2$, $i = 1, 2$, de los dos resultados anteriores se sigue que:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \approx F_{n_1-1,n_2-1}$$

Por tanto, utilizando la tabla de la distribución F, podemos encontrar los percentiles $\alpha/2$ y $1-\alpha/2$ de modo que:

$$P\left(F_{n_1-1,n_2-1,1-\alpha/2} \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq F_{n_1-1,n_2-1,\alpha/2}\right) = 1-\alpha$$

Ordenando términos en la desigualdad:

$$P\left(\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right) = 1-\alpha$$

Por tanto el intervalo de confianza a nivel $1-\alpha$ para el cociente de varianzas σ_1^2/σ_2^2 de poblaciones normales es:

$$\left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right]$$

Nota: si sólo se dispone de la tabla F para el nivel $\alpha/2$ utilizaremos la propiedad:

$$F_{n_1-1,n_2-1,1-\alpha/2} = \frac{1}{F_{n_2-1,n_1-1,\alpha/2}}$$

Ejemplo de aplicación: Siguiendo con nuestro ejemplo del incremento de peso (en gramos) que se obtiene en peces alimentados con una dieta experimental, supongamos que se ensaya una segunda dieta en otro tanque con 24 peces, y que los incrementos de peso observados en este caso son:

439	425	345	368	390	424	448	332	452	420	422	311
382	383	419	387	456	500	436	446	385	391	368	405

Obviamente estos incrementos de peso presentan variabilidad (no todos los peces con la misma dieta ganan el mismo peso). Se desea estimar la diferencia entre esta variabilidad y la que se produce cuando se utiliza la primera dieta (ver datos en la página 4).

Las variabilidades de los incrementos de peso con ambas dietas pueden cuantificarse mediante las varianzas muestrales respectivas. Si denotamos por $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$ y $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$ las dos muestras, siendo $n_1 = 20$, $n_2 = 24$, y las medias muestrales respectivas $\bar{X}_1 = 359{,}6$ y $\bar{X}_2 = 405{,}58$, tenemos:

$$s_1^2 = \frac{\sum_{i=1}^{n_1}\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1} = \frac{\sum_{i=1}^{20}\left(X_{1i} - 359{,}6\right)^2}{19} = 3944{,}25$$

$$s_2^2 = \frac{\sum_{i=1}^{n_2}\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1} = \frac{\sum_{i=1}^{24}\left(X_i - 405{,}58\right)^2}{23} = 1895{,}91$$

El cociente entre ambos valores es:

$$\frac{s_1^2}{s_2^2} = \frac{3944{,}25}{1895{,}91} = 2{,}08$$

por lo que la variabilidad *observada* cuando se administra la primera dieta es el doble que cuando se administra la segunda. El intervalo de confianza al 95 % nos ayuda a poner este dato en perspectiva ya que nos proporciona el margen de error probable en esta estimación:

$$\left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}}\right] = \left[\frac{2{,}08}{F_{19,23,0{,}025}}, \frac{2{,}08}{1/F_{23,19,0{,}025}}\right] =$$
$$= \left[\frac{2{,}08}{2{,}374}, \frac{2{,}08}{1/2{,}465}\right] = [0{,}88, 5{,}13]$$

De esta forma vemos que, con la información que tenemos, y con un 95 % de confianza el valor (desconocido) del cociente σ_1^2/σ_2^2 podría llegar a ser tan pequeño como 0.88 o tan grande como 5.13. Nótese que el hecho de que 0.88 sea menor que 1, significa que podría ser que $\sigma_1^2 < \sigma_2^2$; como el valor 1 también está incluído en el intervalo, ello significa que podría ser $\sigma_1^2/\sigma_2^2 = 1$ y por tanto $\sigma_1^2 = \sigma_2^2$; y como el intervalo contiene también valores mayores que 1, ello implicaría que podría ocurrir también que $\sigma_1^2 > \sigma_2^2$. Evidentemente las tres cosas no pueden ocurrir al mismo tiempo, y el resultado que hemos obtenido, en definitiva, nos indica que *no tenemos información suficiente para* distinguir de una manera clara entre las tres situaciones. Por tanto, aunque en las muestras disponibles la varianza observada con la dieta 1 duplique a la varianza observada con la dieta 2, no hay evidencia suficiente para generalizar este resultado, pudiendo achacarse la diferencia observada al puro azar.

Cálculo en R : en R es posible calcular fácilmente un intervalo de confianza para el cociente de varianzas del siguiente modo:

```
> incPeso2 = c(439, 425, 345, 368, 390, 424, 448, 332, 452, 420,
    422, 311, 382, 383, 419, 387, 456, 500, 436, 446, 385, 391,
    368, 405)
> var.test(incPeso, incPeso2)

        F test to compare two variances

data:  incPeso and incPeso2
F = 2.0804, num df = 19, denom df = 23, p-value = 0.0957
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
```

```
 0.8761571 5.1277598
sample estimates:
ratio of variances
          2.080405
```

Al igual que hemos visto en casos anteriores, R no se limita sólo a calcular un intervalo para el cociente de varianzas, sino que presenta además un contraste de hipótesis que se explica en el siguiente capítulo.

5.9. Intervalos de confianza para la diferencia de medias de poblaciones normales.

En muchas ocasiones resulta de interés estimar un intervalo de confianza para la diferencia entre las medias de dos distribuciones normales $X_1 \approx N(\mu_1, \sigma_1)$ y $X_2 \approx N(\mu_2, \sigma_2)$. La diferencia entre las medias muestrales $\bar{X}_1 - \bar{X}_2$ nos permite estimar $\mu_1 - \mu_2$, y el intervalo de confianza nos dará una idea de la precisión conseguida en la estimación. Para ello será preciso disponer de sendas muestras aleatorias de ambas variables. Denotaremos a dichas muestras como $\{X_{11}, X_{12}, \ldots, X_{1n_1}\}$ y $\{X_{21}, X_{22}, \ldots, X_{2n_2}\}$. El diseño del muestreo puede llevarse a cabo de dos formas:

- **Muestras independientes:** las variables X_1 y X_2 son independientes: el conocimiento de X_1 no aporta información sobre X_2. En general, cuando se utilizan muestras independientes, los sujetos u objetos sobre los que se mide X_1 no tienen relación ni asociación alguna con aquellos sobre los que se mide X_2. Por ejemplo, en un ensayo sobre la ganancia de peso que se consigue con dos dietas distintas, si la primera dieta se experimenta sobre una muestra de n_1 peces en un tanque, y la segunda sobre otros n_2 peces en otro tanque diferente, ambas muestras son independientes. Los valores de n_1 y n_2 pueden ser iguales o distintos.

- **Muestras emparejadas:** las variables X_1 y X_2 están asociadas, y por tanto, el conocimiento de los valores de una aporta información sobre los valores de la otra. En un diseño de muestras emparejadas ambas muestras son del mismo tamaño. Las variables X_1 y X_2 se suelen medir sobre los mismos sujetos u objetos, o bien sobre objetos que han sido cuidadosamente emparejados según características comunes. Por ejemplo, si se desea conocer el incremento medio de peso que se consigue en una semana con una dieta concreta, se pueden utilizar n peces, siendo X_{1i} el peso del pez i-ésimo al inicio del experimento y X_{2i} su peso al final; de esta forma las variables X_1 y X_2 están emparejadas.

5.9.1. Muestras Independientes: Varianzas conocidas.

Si $X_1 \approx N(\mu_1, \sigma_1)$ y $X_2 \approx N(\mu_2, \sigma_2)$, y se toma una muestra de tamaño n_1 de X_1, y una muestra de tamaño n_2 de X_2, siendo ambas muestras independientes, entonces $\bar{X}_1 \approx N\left(\mu_1, \sigma_1/\sqrt{n_1}\right)$ y $\bar{X}_2 \approx N\left(\mu_2, \sigma_2/\sqrt{n_2}\right)$. De acuerdo con la propiedad reproductiva de la distribución normal, se tiene que

$$\bar{X}_1 - \bar{X}_2 \approx N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

por lo que:

$$\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

A partir de aquí podemos proceder de modo análogo al caso del intervalo de confianza para la media de una población normal con varianza conocida.

El intervalo de confianza a nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones normales con varianzas conocidas es entonces:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

5.9.2. Muestras Independientes: Varianzas desconocidas e iguales.

Si $X_1 \approx N(\mu_1, \sigma)$ y $X_2 \approx N(\mu_2, \sigma)$, y se dispone de sendas muestras aleatorias independientes de ambas variables, de tamaños respectivos n_1 y n_2 entonces:

$$\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t_{n_1+n_2-2}$$

donde:

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

El intervalo de confianza a nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones normales con la misma varianza (desconocida) es entonces:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

Ejemplo: Siguiendo con nuestro ejemplo del incremento de peso (en gramos) que se obtiene en peces alimentados con dos dietas, sea X_1 el incremento de peso cuando se utiliza la dieta 1 y X_2 el incremento cuando se usa la dieta 2. En este caso podemos asumir que las dos muestras son independientes ya que los datos para cada dieta han sido obtenidos con peces distintos en tanques distintos, sin que haya habido relación ni influencia alguna entre ambos tanques. Si asumimos además que $X_1 \approx N(\mu_1, \sigma_1)$ y $X_2 \approx N(\mu_2, \sigma_2)$, con $\sigma_1 = \sigma_2$, utilizando los datos que hemos visto en las páginas 4 y 16 tenemos:

$$s_p = \sqrt{\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}} = \sqrt{\frac{19 \cdot 3944{,}25 + 23 \cdot 1895{,}91}{42}} = 53{,}13$$

$$\bar{X}_1 = 359{,}6, \qquad \bar{X}_2 = 405{,}58, \qquad \bar{X}_1 - \bar{X}_2 = -45{,}98$$

y por tanto el intervalo de confianza al 95 % es:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = \left[-45{,}98 \pm 2{,}018 \cdot 53{,}13 \cdot \sqrt{\frac{1}{20} + \frac{1}{24}}\right]$$
$$= [-78{,}44, -13{,}52]$$

Así pues, en las muestras disponibles el incremento de peso ha sido, por término medio, casi 46 gramos mayor cuando se usa la dieta 2. Ahora bien, a la hora de generalizar este resultado, con un 95 % de confianza podemos afirmar que con la dieta 2 se ganan, por término medio, entre 13.52 y 78.44 gramos más de peso que con la dieta 1. Por tanto, la dieta 2 produce (con un 95 % de confianza) mayor incremento de peso que la dieta 1.

Cálculo con R : en R es posible calcular fácilmente un intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas iguales utilizando el siguiente comando (nótese el uso del argumento `var.equal=TRUE` con el que se indica que asumimos que las varianzas son iguales):

```
> t.test(incPeso, incPeso2, var.equal = T)

        Two Sample t-test

data:  incPeso and incPeso2
t = -2.8587, df = 42, p-value = 0.006594
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -78.44452 -13.52214
sample estimates:
mean of x mean of y
 359.6000  405.5833
```

Nuevamente, R no se limita sólo a calcular un intervalo para el cociente de varianzas, sino que presenta además un contraste de hipótesis que se explica en el siguiente capítulo.

5.9.3. Muestras Independientes: Varianzas desconocidas y distintas.

En el caso anterior hemos supuesto que las varianzas de las variables X_1 y X_2 son iguales. En la práctica, lo más frecuente es que ambas varianzas sean diferentes. En este caso es posible demostrar que:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_n$$

con

$$n = REDONDEO \left[\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1 - 1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2 - 1}} \right]$$

El intervalo de confianza a nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones normales con varianzas desconocidas y distintas es entonces:

$$\left[(\bar{X}_1 - \bar{X}_2) \pm t_{n,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Ejemplo: En el caso anterior hemos supuesto la igualdad de las varianzas σ_1^2 y σ_2^2 de los incrementos de peso obtenidos al administrar dos dietas distintas al cultivo de peces de

una misma especie. En la página 17 hemos visto, a partir del cálculo de un intervalo de confianza para el cociente σ_1^2/σ_2^2, que con la evidencia disponible no es posible estar seguros de si ambas varianzas son iguales o distintas. Por ello resulta cuando menos prudente calcular el intervalo de confianza para la diferencia de medias suponiendo que las varianzas son distintas. Bajo este supuesto calculamos en primer lugar:

$$n = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \frac{1}{n_1-1} + \left(\frac{s_2^2}{n_2}\right)^2 \frac{1}{n_2-1}} = \frac{\left(\frac{3944{,}25}{20} + \frac{1895{,}91}{24}\right)^2}{\left(\frac{3944{,}25}{20}\right)^2 \frac{1}{19} + \left(\frac{1895{,}91}{24}\right)^2 \frac{1}{23}} = 32{,}91 \cong 33$$

El intervalo de confianza para la diferencia de medias es entonces:

$$\left[(\bar{X}_1 - \bar{X}_2) \pm t_{n,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] = \left[359{,}6 - 405{,}58 \pm t_{33,0{,}025}\sqrt{\frac{3944{,}25}{20} + \frac{1895{,}91}{24}}\right] =$$
$$= [-79{,}79, -12{,}17]$$

Cálculo con R : en R el intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas distintas se calcula mediante el siguiente comando (nótese que ahora NO utilizamos el argumento *var.equal=TRUE*; por defecto R siempre asume que las varianzas de las poblaciones que se comparan son distintas):

```
> t.test(incPeso, incPeso2)

        Welch Two Sample t-test

data:  incPeso and incPeso2
t = -2.7668, df = 32.908, p-value = 0.009215
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -79.79960 -12.16706
sample estimates:
mean of x mean of y
 359.6000  405.5833
```

¿Varianzas iguales o varianzas distintas?: en la práctica, y tal como ha ocurrido en el ejemplo que acabamos de desarrollar, cuando se desea comparar las medias de dos poblaciones normales a partir de la información suministrada por sendas muestras independientes, quien toma los datos no sabe si proceden de poblaciones con varianzas

iguales o de poblaciones con varianzas distintas. ¿Cuál de los dos intervalos anteriores resulta entonces más adecuado?. En general, ambos intervalos resultan muy similares y de hecho, si las muestras son de gran tamaño, ambos intervalos resultan indistinguibles. Cuando las muestras son pequeñas, el intervalo que asume varianzas distintas es siempre algo más amplio que el que asume varianzas iguales. Por tanto el primer intervalo nos garantiza que siempre se alcanza *al menos* la confianza deseada, por lo que resulta preferible. Así, salvo que tengamos razones muy fundadas para pensar que ambas varianzas deban ser iguales, las consideraremos distintas y aplicaremos el intervalo correspondiente a este caso. Como ya hemos mencionado, este es el intervalo que R siempre aplica por defecto.

Variables no normales: Otra cuestión es si las variables cuyas medias se comparan tienen o no distribución normal. Por efecto del teorema central del límite:

En caso de que se disponga de muestras de gran tamaño, aún cuando la distribución de las variables no sea normal, un intervalo de confianza a nivel $1-\alpha$ para la diferencia de medias es:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

En la práctica este intervalo se suele utilizar si n_1 y n_2 son ambos mayores que 30.

En caso de que las variables cuyas medias se comparan no tengan distribución normal, y los tamaños de muestra sean pequeños los intervalos de confianza mostrados en este curso no son de aplicación y debe recurrirse a otras técnicas como el *bootstrap*.

5.10. Muestras emparejadas.

Los intervalos de confianza para las diferencias de medias vistos hasta ahora son de aplicación cuando la comparación se realiza sobre muestras independientes. En el caso de que se utilice un diseño de muestras emparejadas, los valores de X_1 no son independientes de los de X_2. La construcción del intervalo de confianza, no obstante, es sencilla sin más que considerar que si $X_1 \approx N(\mu_1, \sigma_1)$, $X_2 \approx N(\mu_2, \sigma_2)$ y $cov(X_1, X_2) = \sigma_{12}$, entonces la variable $D = X_1 - X_2$ sigue una distribución $N(\mu_D, \sigma_D)$ donde

$$\mu_D = \mu_1 - \mu_2$$
$$\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

Nótese que si $\{X_{11}, X_{12}, \ldots, X_{1n}\}$ y $\{X_{21}, X_{22}, \ldots, X_{2n}\}$, son las muestras de X_1 y X_2, respectivamente, se dispone entonces de una muestra de D, dada por

$$\{D_1, D_2, \ldots, D_n\} = \{X_{11} - X_{21}, X_{12} - X_{22}, \ldots, X_{1n} - X_{1n}\}$$

Por tanto, construir un intervalo para $\mu_1 - \mu_2$ en estas condiciones es equivalente a construir un intervalo de confianza para la media μ_D de una variable normal $N(\mu_D, \sigma_D)$ a partir de la muestra anterior. Si σ_D es desconocida, como suele ser habitual en la práctica, este intervalo según hemos visto en la sección 5.6 es de la forma:

$$\left[\bar{D} - \frac{S_D}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{D} + \frac{S_D}{\sqrt{n}} t_{n-1,\alpha/2}\right]$$

donde

$$\bar{D} = \bar{X}_1 - \bar{X}_2$$

y

$$\begin{aligned} S_D &= \sqrt{\frac{\sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} \left((X_{1i} - X_{2i}) - \left(\bar{X}_1 - \bar{X}_2\right)\right)^2}{n-1}} = \\ &= \sqrt{\frac{\sum_{i=1}^{n} \left(\left(X_{1i} - \bar{X}_1\right) - \left(X_{2i} - \bar{X}_2\right)\right)^2}{n-1}} = \\ &= \sqrt{\frac{\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)^2 + \sum_{i=1}^{n} \left(X_{2i} - \bar{X}_2\right)^2 - 2\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)\left(X_{2i} - \bar{X}_2\right)}{n-1}} = \\ &= \sqrt{S_1^2 + S_2^2 - 2S_{12}} \end{aligned}$$

Por tanto el intervalo de confianza a nivel $1 - \alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de poblaciones normales en muestras emparejadas de tamaño n es:

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n-1,\alpha/2} \frac{S_D}{\sqrt{n}}\right]$$

Ejemplo: Se dispone de una muestra de 12 tortugas. De cada ejemplar se han medido las variables X_1=*Longitud* y X_2=*Anchura* del caparazón (en centímetros), con los resultados que se muestran a continuación:

Longitud	82.2	74.5	81.4	81.7	85.8	81.6	82.7	74	78.6	85.9	78	80.3
Anchura	78.4	71.5	74.9	80.1	85.6	80.8	77.5	71.3	76.3	82.7	79.5	79.6

Suponiendo que ambas variables siguen sendas distribuciones normales, se desea calcular un intervalo de confianza al 95 % para la diferencia $\mu_1 - \mu_2$.

Obviamente estos datos corresponden a un diseño de muestras emparejadas, ya que cada pareja de valores Longitud-Anchura se ha medido sobre un mismo ejemplar, por lo que cabe esperar que ambas medidas estén asociadas. Las diferencias entre longitud y anchura observadas para cada tortuga son:

D	3.8	3	6.5	1.6	0.2	0.8	5.2	2.7	2.3	3.2	-1.5	0.7

Se tiene entonces:

$$\bar{X}_1 = 80{,}56 \text{ (Longitud media)}, \qquad \bar{X}_2 = 78{,}18 \text{ (Anchura media)}$$

$$\bar{D} = \bar{X}_1 - \bar{X}_2 = 2{,}38, \qquad S_D = \sqrt{\frac{\sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2}{n-1}} = 2{,}21$$

$$t_{11,0{,}025} = 2{,}201$$

Por tanto, el intervalo de confianza para $\mu_1 - \mu_2$ es

$$\left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}}\right] = \left[2{,}38 \pm 2{,}201\frac{2{,}21}{\sqrt{12}}\right] = [2{,}38 \pm 1{,}4] = [0{,}97, 3{,}78]$$

Dicho de otra forma, se estima que la longitud de estas tortugas es, por término medio, 2.38 centímetros mayor que su anchura; con un 95 % de confianza el verdadero valor de esta diferencia se encuentra entre 0.97 y 3.78 centímetros.

Cálculo con R : en R el intervalo de confianza para la diferencia de medias en poblaciones normales con muestras emparejadas se obtiene también con el comando *t.test*, especificando en este caso la opción *paired=TRUE*:

```
> long = c(82.2, 74.5, 81.4, 81.7, 85.8, 81.6, 82.7, 74, 78.6,
    85.9, 78, 80.3)
> anch = c(78.4, 71.5, 74.9, 80.1, 85.6, 80.8, 77.5, 71.3, 76.3,
    82.7, 79.5, 79.6)
> t.test(long, anch, paired = T)
```

```
        Paired t-test

data:  long and anch
t = 3.7187, df = 11, p-value = 0.003390
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9692996 3.7807004
sample estimates:
mean of the differences
                  2.375
```

5.11. Intervalo de confianza para una proporción

La estimación de una proporción es un problema frecuente en la práctica: qué proporción de los huevos de tortuga depositados en una playa eclosionan con éxito, cuál es la proporción de hembras en una especie, qué proporción de los objetos producidos en una fábrica tiene defectos, qué proporción pasa el control de calidad, qué proporción de enfermos experimenta mejoría cuando se les aplica un tratamiento concreto, etc.

Podemos tratar este problema desde un punto de vista general considerando que en estos casos se observa una variable aleatoria X con distribución de Bernoulli de parámetro desconocido π. Recordemos que la variable aleatoria de Bernouilli se caracteriza por tomar uno de dos posibles valores, 1 (éxito*)* ó 0 (*fracaso*), siendo π la probabilidad de éxito. En cada caso particular, el éxito corresponderá a aquel suceso cuya probabilidad queremos estimar: que un huevo de tortuga eclosione, que un ejemplar sea hembra o que un objeto de la producción tenga defectos, por ejemplo.

Sea $\{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de la variable de Bernoulli. Si $N_E = \sum_{i=1}^{n} X_i$ es el número observado de éxitos en la muestra, un estimador de π es:

$$\hat{\pi} = \frac{N_E}{n}$$

esto es, la proporción de éxitos en la muestra. En el capítulo anterior ya hemos visto que este estimador es el que se obtiene tanto por el método de los momentos como por máxima verosimilitud. Sabemos además que el número de éxitos en n pruebas N_E sigue una distribución binomial $B(n, \pi)$, por lo que:

$$E[\hat{\pi}] = E\left[\frac{N_E}{n}\right] = \frac{1}{n}E[N_E] = \frac{1}{n}n\pi = \pi$$

y por tanto $\hat{\pi}$ es un estimador centrado de π.

Ejemplo 5.1. Se han elegido al azar 60 huevos de tortuga en una playa inmediatamente tras la puesta. Transcurrido el periodo de incubación se observa que sólo de 23 de estos huevos nacen tortugas vivas. De esta forma, la proporción de huevos que eclosionan en tortugas vivas puede estimarse como $\hat{\pi} = 23/60 = 0{,}3833 \cong 38{,}33\,\%$.

Para calcular un intervalo de confianza para la proporción π existen varios métodos, que describimos a continuación.

5.11.1. Método de Wilson.

Como $N_E = \sum_{i=1}^{n} X_i \approx B(n, \pi)$, si el valor de n es suficientemente grande (en la práctica si $n\hat{\pi} > 5$ y $n(1-\hat{\pi}) > 5$), entonces, por efecto del teorema central del límite tal como vimos en el capítulo 3:

$$\frac{N_E - n\pi}{\sqrt{n\pi(1-\pi)}} \approx N(0,1)$$

Si observamos que:

$$\frac{N_E - n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{(N_E - n\pi)/n}{\left(\sqrt{n\pi(1-\pi)}\right)/n} = \frac{\frac{N_E}{n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

entonces:

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

Por tanto:

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

Ahora bien:

$$-z_{a/2} \leq \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leq z_{\alpha/2} \Leftrightarrow \left|\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right| \leq z_{\alpha/2} \Leftrightarrow \left(\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right)^2 \leq z_{\alpha/2}^2$$

$$\Leftrightarrow n(\hat{\pi} - \pi)^2 \leq z_{\alpha/2}^2 \pi(1-\pi) \Leftrightarrow \left(n + z_{\alpha/2}^2\right)\pi^2 - \left(2n\hat{\pi} + z_{\alpha/2}^2\right)\pi + n\hat{\pi}^2 \leq 0$$

Si tenemos en cuenta que la función $g(\pi) = \left(n + z_{\alpha/2}^2\right)\pi^2 - \left(2n\hat{\pi} + z_{\alpha/2}^2\right)\pi + n\hat{\pi}^2$ representa una parábola con los brazos abiertos hacia arriba, la desigualdad anterior se verificará para los valores de π comprendidos entre los dos puntos en que esa parábola corta al eje de abcisas.

Estos puntos son las soluciones de la ecuación $\left(n + z^2_{\alpha/2}\right)\pi^2 - \left(2n\hat{\pi} + z^2_{\alpha/2}\right)\pi + n\hat{\pi}^2 = 0$, que se obtienen fácilmente como:

$$\pi = \frac{\left(2n\hat{\pi} + z^2_{\alpha/2}\right) \pm \sqrt{\left(2n\hat{\pi} + z^2_{\alpha/2}\right)^2 - 4\left(n + z^2_{\alpha/2}\right)n\hat{\pi}^2}}{2\left(n + z^2_{\alpha/2}\right)} =$$

$$= \frac{\left(2n\hat{\pi} + z^2_{\alpha/2}\right) \pm \sqrt{4nz^2_{\alpha/2}\hat{\pi}\left(1-\hat{\pi}\right) + z^4_{\alpha/2}}}{2\left(n + z^2_{\alpha/2}\right)} =$$

$$= \frac{\left(n\hat{\pi} + z^2_{\alpha/2}/2\right)}{\left(n + z^2_{\alpha/2}\right)} \pm \frac{z_{\alpha/2}\sqrt{n}}{\left(n + z^2_{\alpha/2}\right)}\sqrt{\hat{\pi}\left(1-\hat{\pi}\right) + z^2_{\alpha/2}/4n}$$

Por tanto, utilizando que $n\hat{\pi} = N_E$:

$$P\left(\pi \in \left[\frac{\left(N_E + z^2_{\alpha/2}/2\right)}{\left(n + z^2_{\alpha/2}\right)} \pm \frac{z_{\alpha/2}\sqrt{n}}{\left(n + z^2_{\alpha/2}\right)}\sqrt{\hat{\pi}\left(1-\hat{\pi}\right) + z^2_{\alpha/2}/4n}\right]\right) = 1 - \alpha$$

Ejemplo de aplicación: Para calcular un intervalo de confianza al 95 % para la proporción de huevos de tortuga que eclosionan con éxito a partir de los datos del ejemplo 5.1, calculamos $\hat{\pi} = 23/60 = 0{,}3833$ y obtenemos $z_{\alpha/2} = z_{0{,}025} = 1{,}96$ en la tabla de la distribución normal. Sustituyendo estos valores en la expresión anterior obtenemos el intervalo:

$$[0{,}39035 \pm 0{,}11947] = [0{,}27088,\ 0{,}50982]$$

por lo que, con un 95 % de confianza dicha proporción se encuentra entre el 27,088 % y el 50,982 % de los huevos que se ponen en esa playa.

Cálculo con R : En el paquete base de R no se encuentra implementado este intervalo. Sí que se encuentra, no obstante, en la librería *binom*, utilizando el comando *binom.confint*. Para los datos de nuestro ejemplo:

```
> library(binom)
> binom.confint(23, 60, method = "wilson")

  method  x  n      mean     lower    upper
1 wilson 23 60 0.3833333 0.2708827 0.509824
```

5.11.2. Método de Agresti-Coull

Este método proporciona un intervalo de confianza para la proporción con una expresión algo más sencilla que la anterior, si bien requiere tamaños muestrales mayores que 40. En estas condiciones se puede utilizar la aproximación:

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

Por tanto:

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

y despejando π:

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

Así pues, el intervalo de confianza aproximado a nivel $1-\alpha$ para π es:

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

(*Intervalo de Wald*). Este intervalo tiene, no obstante, mal comportamiento para muy diversos valores de n y π, por lo que su uso es desaconsejable. Agresti y Coull han propuesto una modificación de este intervalo que resuelve estos problemas. La modificación consiste en definir:

$$\begin{aligned}\tilde{N}_E &= N_E + z^2_{\alpha/2}/2\\ \tilde{n} &= n + z^2_{\alpha/2}\\ \tilde{\pi} &= \tilde{N}_E/\tilde{n}\end{aligned}$$

y recalcular el intervalo de confianza de Wald sustituyendo $\hat{\pi}$ por $\tilde{\pi}$ y n por $\tilde{n}$. El intervalo de confianza a nivel $1-\alpha$ es entonces de la forma:

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

(*Intervalo de Agresti y Coull*)

Ejemplo de aplicación: Calculamos de nuevo un intervalo de confianza al 95 % para la proporción de huevos de tortuga que eclosionan con éxito a partir de los datos del ejemplo 5.1, utilizando ahora el método de Agresti-Coull (podemos hacerlo ya que $n > 40$). En este caso se tiene $\tilde{\pi} = 0{,}39035$, $z_{0{,}025} = 1{,}96$ y $\tilde{n} = 63{,}84$. Sustituyendo se obtiene el intervalo:

$$[0{,}39035 \pm 1{,}96 \cdot 0{,}06105] = [0{,}39035 \pm 1{,}96 \cdot 0{,}11964] = [0{,}27069, 0{,}51002]$$

que como puede apreciarse es muy similar al obtenido por el método de Wilson (los extremos se diferencian en menos de una milésima). De hecho, a medida que n aumenta los métodos de Agresti y Coull, y Wilson tienden a producir el mismo intervalo.

Cálculo con R : En el paquete base de R tampoco se encuentra implementado este intervalo, pero al igual que el anterior, podemos encontrarlo en la librería ***binom***, utilizando el comando ***binom.confint*** y especificando el método *"agresti"*. Para los datos de nuestro ejemplo:

```
> library(binom)
> binom.confint(23, 60, method = "agresti")

         method  x  n      mean     lower     upper
1 agresti-coull 23 60 0.3833333 0.2706890 0.5100177
```

Por cierto, que el intervalo de Wald también obtenerse en R con la librería ***binom*** especificando el método "*asymptotic":*

```
> binom.confint(23, 60, method = "asymptotic")

      method  x  n      mean     lower     upper
1 asymptotic 23 60 0.3833333 0.2603104 0.5063562
```

5.11.3. Método de Clopper y Pearson

En el caso de que el tamaño n de la muestra o el valor de la proporción estimada $\hat{\pi}$ sean tan pequeños que no se dan las condiciones para aplicar los métodos de Wilson o Agresti y Coull, puede probarse que el siguiente intervalo garantiza un nivel de confianza de al menos $1 - \alpha$ para la estimación del parámetro π:

$$\left[\frac{N_E}{(n-N_E+1)F_1+N_E}, \frac{(N_E+1)F_2}{(n-N_E)+(N_E+1)F_2}\right]$$

(*Intervalo de Clopper-Pearson*) donde:

$$F_1 = F_{2(n-N_E+1),2N_E,\alpha/2}, \quad F_2 = F_{2(N_E+1),2(n-N_E),\alpha/2}$$

son percentiles de la distribución F de Fisher. Conviene señalar que al ser un intervalo que garantiza que la confianza es al menos $1-\alpha$, en muchas ocasiones el nivel de confianza real será mayor, por lo cual este intervalo resulta en general más amplio y por tanto más impreciso que los anteriores, y sólo debe emplearse si no se dan las condiciones para utilizar alguno de aquéllos.

Ejemplo de aplicación: Si con los datos del ejemplo anterior calculamos el intervalo de Clopper-Pearson, obtenemos:

$$F_1 = F_{2(60-23+1),2\cdot 23,0,025} = F_{76,46,0,025} = 1{,}71636,$$
$$F_2 = F_{2(23+1),2(60-23),0,025} = F_{48,74,0,025} = 1{,}65605$$

y el intervalo es entonces: $\left[\frac{23}{(60-23+1)1{,}71636+23}, \frac{(23+1)\cdot 1{,}65605}{(60-23)+(23+1)\cdot 1{,}65605}\right] = [0{,}26071,\ 0{,}51789]$

Como puede apreciarse este intervalo es similar a los anteriores, aunque algo más amplio. Esta mayor amplitud se debe, como hemos señalado, a que el nivel de confianza de este intervalo es algo mayor que el 95 %.

Cálculo con R : en R el intervalo de Clopper y Pearson se obtiene mediante la función *binom.test*. En la sintaxis debe especificarse primero el número de éxitos N_E, y a continuación el número de pruebas (tamaño de la muestra) n. Así, para los datos del ejemplo anterior utilizaríamos:

```
> binom.test(23, 60)

        Exact binomial test

data:  23 and 60
number of successes = 23, number of trials = 60, p-value = 0.09246
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2607071 0.5178850
sample estimates:
```

```
probability of success
            0.3833333
```

5.12. Intervalos de confianza para la comparación de proporciones en poblaciones independientes.

En ocasiones se desean comparar los parámetros π_1 y π_2 de sendas distribuciones de Bernoulli en poblaciones independientes. Por ejemplo: ¿cuál es la diferencia entre las proporciones de machos en dos especies distintas? ¿Cuál es la diferencia entre las proporciones de enfermos que mejoran con dos tratamientos alternativos? ¿La proporción de microchips defectuosos difiere mucho entre dos técnicas diferentes de fabricación de microchips?. La comparación de dos proporciones puede llevarse a cabo mediante su diferencia $\pi_1 - \pi_2$ o mediante su cociente π_1/π_2. Cada una de las dos proporciones se estima mediante la proporción muestral, por lo que el estimador de la diferencia será $\hat{\pi}_1 - \hat{\pi}_2$ y el del cociente será $\hat{\pi}_1/\hat{\pi}_2$. Como en todos los casos anteriores, en la práctica será conveniente acompañar la estimación por un intervalo de confianza.

Si los tamaños muestrales son grandes, el teorema central del límite nos indica que, aproximadamente:

$$\pi_k \approx N\left(\hat{\pi}_k, \sqrt{\frac{\hat{\pi}_k\left(1-\hat{\pi}_k\right)}{n}}\right), \quad k = 1, 2$$

por lo que

$$\pi_1 - \pi_2 \approx N\left(\hat{\pi}_1 - \hat{\pi}_2, \frac{\hat{\pi}_1\left(1-\hat{\pi}_1\right)}{n_1} + \frac{\hat{\pi}_2\left(1-\hat{\pi}_2\right)}{n_2}\right)$$

de donde se deduce fácilmente que un intervalo de confianza aproximado a nivel $1-\alpha$ para $\pi_1 - \pi_2$ sería de la forma:

$$\left[\left(\hat{\pi}_1 - \hat{\pi}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1\left(1-\hat{\pi}_1\right)}{n_1} + \frac{\hat{\pi}_2\left(1-\hat{\pi}_2\right)}{n_2}}\right]$$

(*intervalo de* Wald*)*. El comportamiento de este intervalo mejora si se introduce una *corrección por continuidad*, tal como se vio en el capítulo 3, en la aproximación de la distribución binomial por la normal. Se obtiene así el *intervalo de Wald corregido*:

$$\left[(\hat{\pi}_1 - \hat{\pi}_2) \pm \left(z_{\alpha/2} \sqrt{\frac{\hat{\pi}_1 (1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2 (1 - \hat{\pi}_2)}{n_2}} + \frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right) \right]$$

Cuando la comparación de las proporciones se realiza a través del cociente, puede probarse que el siguiente intervalo, con muestras grandes, proporciona una confianza aproximada de $1 - \alpha$ para la estimación del logaritmo de π_1/π_2:

$$\ln \left(\frac{\pi_1}{\pi_2} \right) \in \left[\ln \left(\frac{\hat{\pi}_1}{\hat{\pi}_2} \right) \pm z_{\alpha/2} \sqrt{\frac{(1 - \hat{\pi}_1)}{n_1 \hat{\pi}_1} + \frac{(1 - \hat{\pi}_2)}{n_2 \hat{\pi}_2}} \right]$$

Ejemplo: En una playa situada al norte de una isla se han elegido al azar 160 huevos de tortuga, de los cuales 30 habían sido depredados por cangrejos. En otra playa situada al sur, de 125 huevos, 28 presentaban señales de depredación por cangrejos. Se desean calcular intervalos de confianza al 95 % para la diferencia y para el cociente de las proporciones de huevos depredados en ambas playas.

En este caso las proporciones de huevos depredados en cada playa son, respectivamente, $\hat{\pi}_1 = \frac{30}{160} = 0{,}1875 \cong 18{,}75\,\%$ y $\hat{\pi}_2 = \frac{28}{125} = 0{,}224 \cong 22{,}4\,\%$. El intervalo para la diferencia de proporciones es entonces:

$$\left[(0{,}1875 - 0{,}224) \pm \left(1{,}96 \sqrt{\frac{0{,}1875\,(1 - 0{,}1875)}{160} + \frac{0{,}224\,(1 - 0{,}224)}{125}} + \frac{1}{2} \left(\frac{1}{160} + \frac{1}{125} \right) \right) \right]$$

$$= [-0{,}0365 \pm 0{,}1020] = [-0{,}1385, 0{,}0655]$$

Así pues, se estima que en la playa del norte el porcentaje de cangrejos depredados es un 3,65 % inferior al de la playa del sur, si bien el margen de error para esta cifra es tal que con un 95 % de confianza el porcentaje podría oscilar desde un 13.85 % menos a un 6.55 % más, de huevos depredados en el norte que en el sur.

Si deseamos estimar el cociente de proporciones, tenemos que $\hat{\pi}_1/\hat{\pi}_2 = 0{,}1875/0{,}224 = 0{,}8371 \cong 83{,}71\,\%$, esto es, por cada 100 huevos depredados en el sur, sólo se depredan 83.71 en el norte (la tasa de depredación en el norte es un 83.71 % de la del sur). El

intervalo de confianza al 95 % para el logaritmo de este cociente es:

$$\left[\ln(0{,}8371) \pm 1{,}96\sqrt{\frac{(1-0{,}1875)}{30}+\frac{(1-0{,}224)}{28}}\right] = [-0{,}1779 \pm 0{,}4588] =$$

$$= [-0{,}6367, 0{,}2809]$$

y el intervalo al 95 % de confianza para el cociente puede obtenerse sencillamente como:

$$= \left[e^{-0{,}6367}, e^{0{,}2809}\right] = [0{,}5290, 1{,}3244]$$

Por tanto, con un 95 % de confianza podemos decir que, con la incertidumbre que presentan estos datos, la tasa de depredación en el norte podría ser desde poco más de la mitad que la del sur, hasta una vez y un tercio esta última.

Nótese que el intervalo para la diferencia contiene al cero, lo que indica que, con la información que tenemos no es descartable que las tasas de depredación sean iguales en ambas playas. Idéntica conclusión podemos alcanzar observando que el intervalo para el cociente contiene al 1.

Cálculo con R : El intervalo para la diferencia de proporciones puede obtenerse fácilmente en R mediante la función *prop.test(x,n)* donde *x* es un vector con el número de éxitos en cada muestra, y *n* es un vector con los tamaños muestrales. En este caso:

```
> prop.test(c(30, 28), c(160, 125))

        2-sample test for equality of proportions with continuity correction

data:  c(30, 28) out of c(160, 125)
X-squared = 0.3736, df = 1, p-value = 0.5411
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.13849091  0.06549091
sample estimates:
prop 1 prop 2
0.1875 0.2240
```

En cuanto al cociente de proporciones, el paquete base de R no cuenta con ninguna función para la estimación del intervalo de confianza. Aunque es sencillo de calcular, podemos descargar e instalar el paquete *PropCIs*, que incluye la función *riskscoreci(x1,n1,x2,n2,conf)* que sí que implementa este intervalo (con alguna mejora adicional en la aproximación):

```
> library(PropCIs)
> riskscoreci(30, 160, 28, 125, conf = 0.95)

[1] 0.5316599 1.3224231
```

Señalemos, por último, que el cociente de proporciones en la literatura médica se conoce como *Riesgo Relativo*.

5.13. Intervalo de confianza para el parámetro de una distribución exponencial.

Para obtener este intervalo recordemos que si $\{X_1, X_2, \ldots, X_n\}$ es una muestra aleatoria de una distribución $exp(\lambda)$, su suma $T = \sum_{i=1}^{n} X_i$ sigue una distribución gamma $\mathcal{G}\left(n, \frac{1}{\lambda}\right)$ con

$$E[T] = n \cdot \frac{1}{\lambda}$$
$$var(T) = n \cdot \frac{1}{\lambda^2}$$

Si consideramos ahora la variable $V = 2\lambda T = 2\lambda \sum_{i=1}^{n} X_i = 2\lambda n \bar{X}$, como se ha obtenido a partir de T por un simple cambio de escala, entonces V seguirá también una distribución gamma con los parámetros modificados por el mismo factor de misma escala, esto es:

$$E[V] = 2\lambda E[T] = 2\lambda n \frac{1}{\lambda} = 2n$$
$$var(V) = 4\lambda^2 var(T) = 4\lambda^2 n \cdot \frac{1}{\lambda^2} = 4n$$

Por tanto $V = 2\lambda n \bar{X} \approx \mathcal{G}(n, 2) = \chi^2_{2n}$. La tabla de la distribución χ^2 nos permite entonces obtener los percentiles $\chi^2_{2n,1-\alpha/2}$ y $\chi^2_{2n,\alpha/2}$ de forma que:

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

Por tanto:

$$P\left(\chi^2_{2n,1-\alpha/2} \leq 2n\lambda\bar{X} \leq \chi^2_{2n,\alpha/2}\right) = 1 - \alpha$$

Dividiendo todos los términos del interior del intervalo por $2n\bar{X}$:

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

De esta forma el intervalo de confianza a nivel $1-\alpha$ para el parámetro λ de una distribución exponencial calculado a partir de una muestra aleatoria $\{X_1, X_2, \ldots, X_n\}$ con media $\bar{X}$ es:

$$\left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}, \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right]$$

Ejemplo: En una instalación eléctrica, cada vez que se funde un fusible, es reemplazado por otro de iguales características. El tiempo entre reemplazamientos se supone exponencial. A partir de los datos de los últimos 20 fusibles que se han reemplazado, se ha obtenido un tiempo medio entre reemplazamientos de 23 días. Se desea estimar el valor del parámetro λ, así como obtener un intervalo de confianza al 95 % para dicho parámetro.

El estimador de λ es simplemente $\hat{\lambda} = \frac{1}{\bar{X}} = \frac{1}{23} = 0{,}0435$. En la tabla de la distribución χ^2 encontramos los valores $\chi^2_{40,0,975} = 24{,}433$, $\chi^2_{40,0,025} = 59{,}342$. Por tanto el intervalo de confianza al 95 % es:

$$\left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}, \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}}\right] = \left[\frac{24{,}433}{2 \cdot 20 \cdot 23}, \frac{59{,}342}{2 \cdot 20 \cdot 23}\right] = [0{,}0266,\ 0{,}0645]$$

Cálculo con R : R no dispone de ninguna función específica para el cálculo de este intervalo de confianza; no obstante su obtención es elemental. Con los datos del ejemplo anterior:

```
> n = 20
> x = 23
> qchisq(c(0.025, 0.975), 2 * n)/(2 * n * x)
[1] 0.02655765 0.06450186
```

5.14. Intervalo de confianza para el parámetro de una distribución de Poisson

Otra situación frecuente en la práctica es que los datos disponibles procedan de una distribución de Poisson de parámetro λ. Si se dispone de una muestra aleatoria $\{X_1, X_2, \ldots, X_n\}$ de una distribución de Poisson, llamando $T = \sum_{i=1}^{n} X_i$, puede demostrarse que el siguiente intervalo garantiza un nivel de confianza de al menos 1 - α para la estimación del parámetro:

$$\lambda \in \left[\frac{1}{2n}\chi^2_{n_1,1-\alpha/2}, \frac{1}{2n}\chi^2_{n_2,\alpha/2}\right] , \quad n_1 = 2T, \quad n_2 = 2(T+1)$$

Ejemplo. Se realiza un estudio del número de tortugas que acceden diariamente a una playa. Para ello se han seleccionado al azar $n = 40$ días del último año y se ha contado el número de tortugas llegadas a la playa cada día. Durante ese periodo se observó un total de $T = 134$ tortugas. Suponiendo que el número de tortugas diarias sigue una distribución de Poisson, se desea estimar el parámetro de dicha distribución con un intervalo de confianza del 95 %.

El estimador puntual del parámetro, tal como hemos visto en el capítulo anterior es $\hat{\lambda} = \bar{x} = \frac{134}{40} = 3{,}35$. Para obtener el intervalo de confianza calculamos:

$$n_1 = 2T = 2 \cdot 134 = 268, \qquad n_2 = 2\,(134 + 1) = 270$$
$$\chi^2_{268,0,975} = 224{,}5465 \qquad \chi^2_{270,0,025} = 317{,}4092$$

Por tanto, el intervalo de confianza al 95 % es:

$$\left[\frac{1}{80}224{,}5465, \frac{1}{80}317{,}4092\right] = [2{,}807,\ 3{,}968]$$

Cálculo con R : R no dispone de una función específica para el cálculo de este intervalo. No obstante, su cálculo directo es muy simple. Utilizando los datos del ejemplo:

```
> n = 80
> T = 134
> c(qchisq(0.025, 2 * T), qchisq(0.975, 2 * (T + 1)))/(2 * n)

[1] 1.403416 1.983807
```

5.15. Intervalos de confianza aproximados basados en estimadores de máxima verosimilitud.

En todos los casos vistos hasta ahora, la obtención de los intervalos de confianza se ha realizado a través de funciones pivote cuya distribución de probabilidad es conocida y no depende del parámetro a estimar θ, tal como se explicó en la sección 5.5. La obtención de estos pivotes es elemental en algunos casos y más compleja en otros. Pero hay muchos casos en la práctica en que no es posible deducir una función pivote para un parámetro de interés, bien sea por la propia complejidad de la distribución de probabilidad de la variable que se estudia, por la presencia de datos censurados en la muestra[2], o por otras circunstancias. En tales casos, si

[2]Recuérdese del capítulo anterior que un dato censurado es un dato que ofrece sólo información parcial sobre la variable: sabemos de un sujeto que mide más de cierta cantidad, pero no sabemos su longitud exacta;

se dispone de un estimador de máxima verosimilitud para ese parámetro, el siguiente teorema permite utilizarlo para construir intervalos de confianza asintóticos (intervalos de confianza que resultan válidos para tamaños de muestra grandes).

Teorema 5.1. *Sea* $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$ *una muestra aleatoria de una variable* X *con función de densidad* $f_\theta(x)$*, que depende de un parámetro* $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$*. Sea* $L_\mathfrak{X}(\theta)$ *la función de verosimilitud de* θ *dada la muestra* $\mathfrak{X}$*, y sea* $H_\mathfrak{X}(\theta) = \frac{\partial^2 \ln L_X(\theta)}{\partial\theta\,\partial\theta'}$ *la matriz hessiana de segundas derivadas de la log-verosimilitud,* $\ell_\mathfrak{X}(\theta) = \ln(L_\mathfrak{X}(\theta))$*. Bajo las suficientes condiciones de regularidad*[3]*, el estimador de máxima verosimilitud (EMV)* $\hat{\theta}$ *de* θ *es consistente. Además, cuando* $n \to \infty$ *:* $\hat{\theta}_j \approx N\left(\theta_j, \sqrt{\nu_{jj}}\right)$ *siendo* ν_{jj} *el* j*-ésimo elemento de la diagonal de* $-\left(H_\mathfrak{X}(\theta)\right)^{-1}$ *(inversa de la matriz hessiana).*

En la práctica, como el valor de θ no se conoce, la matriz $-\left(H_\mathfrak{X}(\theta)\right)^{-1}$ debe sustituirse por su estimación $-\left(H_\mathfrak{X}\left(\hat{\theta}\right)\right)^{-1}$.

En estas condiciones, el intervalo de confianza aproximado a nivel $1-\alpha$ para el parámetro θ_j, basado en el estimador de máxima verosimilitud $\hat{\theta}$ sería:

$$\left[\hat{\theta}_j \pm z_{\alpha/2}\sqrt{\hat{\nu}_{jj}}\right]$$

siendo $\hat{\nu}_{jj}$ el j-ésimo elemento de la diagonal de $-\left(H_\mathfrak{X}\left(\hat{\theta}\right)\right)^{-1}$

Este resultado, por basarse en la normalidad asintótica de $\hat{\theta}_j$, tal como establece el teorema anterior, es válido sólo cuando $n \to \infty$. En muchas ocasiones se consigue una aproximación razonable a la normalidad para valores de n del orden de 30, si bien ello depende de la distribución de probabilidad de X. Para tamaños de muestra pequeños deben utilizarse otros métodos (*bootstrap, Montecarlo)* que quedan fuera del alcance de este curso.

Nota: la matriz $-\left(H_\mathfrak{X}\left(\hat{\theta}\right)\right)^{-1}$ es un estimador de la *matriz de varianzas-covarianzas* de la *variable aleatoria* $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$. No olvidemos que en muestras distintas se obtienen valores estimados distintos de $\left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\right)$. La variabilidad conjunta de estos valores queda precisamente descrita por su matriz de varianzas-covarianzas. Si $\hat{\nu}_{ij}$ es el término (i,j)-ésimo

sabemos que una célula ha sobrevivido a la acción de un veneno más de 24 horas, pero no sabemos exactamente cuánto ha vivido. Si se utilizan de manera ingenua estos valores censurados para estimar longitudes medias o tiempos medios de supervivencia sin tener en cuenta la presencia de la censura, podemos incurrir en importantes sesgos en la estimación. En el capítulo anterior se señaló como puede construirse una función de verosimilitud que utilice adecuadamente la información de los datos censurados, de forma que el estimador de máxima verosimilitud obtenido a partir de dicha función evita el problema del sesgo.

[3] Condiciones para que exista $H(\Theta)$

de dicha matriz, entonces $\hat{\nu}_{ij}$ es un estimador de $cov\left(\hat{\theta}_i, \hat{\theta}_j\right)$. Asimismo $\hat{\nu}_{jj}$ es un estimador de $var\left(\hat{\theta}_j\right)$.

5.15.1. Ejemplo: cálculo de intervalos de confianza asintóticos para los parámetros de la distribución de Weibull.

Obviamente, calcular los intervalos de confianza asintóticos para los parámetros de una distribución de probabilidad a partir de sus estimadores de máxima verosimilitud puede ser una tarea ardua: calcular la log-verosimilitud, calcular sus derivadas, igualar a cero, despejar los parámetros, calcular las segundas derivadas, ... Afortunadamente R nos permite simplificar enormemente la tarea. Veamos, a modo de ejemplo, como construir intervalos de confianza asintóticos para los parámetros de una distribución de Weibull $W(k, \lambda)$.

Vamos a hacerlo primero de la manera "*difícil*", aplicando paso a paso el teorema anterior. Comenzamos ajustando los parámetros de la distribución $W(\kappa, \lambda)$ por máxima verosimilitud a la variable $X =$"*Altura de ola*". Para ello:

1. Partimos de los datos correspondientes a las alturas medidas en 30 olas:

```
> olas = c(2.1, 2.82, 4.2, 6.34, 2.4, 3.1, 2.15, 2.73, 3.12, 2.41,
    4.59, 2.81, 2.61, 3.81, 3.13, 3.06, 5.85, 3.57, 2.64, 4.08,
    3.38, 1.88, 1.94, 3.24, 1.98, 3.29, 0.21, 2.68, 1.74, 4.25)
```

2. Construimos la función de log-verosimilitud de Weibull, dependiente del vector de parámetros *parms*=(κ, λ), y de la muestra *x*:

```
> logver = function(parms, x) {
    k = parms[1]
    eta = parms[2]
    n = length(x)
    lv = n * log(k) - n * k * log(eta) + (k - 1) * sum(log(x)) -
        sum((x/eta)^k)
    return(lv)
 }
```

3. Hallamos el máximo de esta función de log-verosimilitud mediante la función *optim()*. En este caso, como nos interesa además calcular intervalos de confianza, pediremos a esta función que nos calcule el hessiano mediante la opción *hessian=TRUE*:

```
> EMV = optim(par = c(1, 1), logver, x = olas,
    control = list(fnscale = -1), hessian = TRUE)
> EMV$par

[1] 2.622085 3.426517

> EMV$hessian

          [,1]       [,2]
[1,] -8.571555   3.725557
[2,]  3.725557 -17.562615
```

4. Obtenemos la matriz inversa del hessiano cambiada de signo, $-\left(H_{\mathfrak{x}}\left(\hat{\theta}\right)\right)^{-1}$, y calculamos la raiz de los elementos de su diagonal:

```
> Hinv = solve(EMV$hessian)
> -Hinv

           [,1]       [,2]
[1,] 0.12851401 0.02726167
[2,] 0.02726167 0.06272215

> se = sqrt(diag(-Hinv))
> se

[1] 0.3584885 0.2504439
```

5. Por último construimos los intervalos de confianza para los parámetros:

```
> conf = 0.95
> z = qnorm(1 - (1 - conf)/2)
> EMV$par[1] + c(-1, 1) * z * se[1]

[1] 1.919461 3.324710

> EMV$par[2] + c(-1, 1) * z * se[2]

[1] 2.935656 3.917378
```

Y ahora de la manera *"fácil"* utilizando la función `fitdistr()` de la librería MASS:

```
> library(MASS)
> estimacion = fitdistr(olas, "weibull")
> estimacion
```

```
     shape       scale
 2.6213967   3.4261091
(0.3584319) (0.2504596)

> confint(estimacion)

         2.5 %   97.5 %
shape 1.918883 3.323910
scale 2.935217 3.917001
```

Esta función también proporciona la estimación de la matriz de varianzas-covarianzas $-\left(H_{\mathfrak{x}}\left(\hat{\theta}\right)\right)^{-1}$:

```
> estimacion$vcov

            shape      scale
shape 0.12847341 0.02727454
scale 0.02727454 0.06273002
```

Las ligeras diferencias que se observan entre estos intervalos y los hallados más arriba se deben a errores de redondeo asociados al uso de distintos algoritmos.

5.15.2. Cálculo de intervalos de confianza asintóticos para los parámetros de otras distribuciones.

El procedimiento a seguir es el mismo que acabamos de ver con la distribución de Weibull. El uso de la función *fitdistr()* facilita enormemente esta tarea. Permite estimar los parámetros (e intervalos de confianza) de las siguientes distribuciones de probabilidad: *beta, cauchy, chi-squared, exponential, f, gamma, geometric, log-normal, lognormal, logistic, negative binomial, normal, Poisson, t* y *weibull*.

5.15.3. Intervalos de confianza para funciones de los estimadores de máxima verosimilitud.

En muchas ocasiones el objetivo de la estimación no son los parámetros de la distribución de probabilidad de la variable de interés, sino alguna otra función de los mismos. Si la altura de ola del ejemplo anterior sigue una distribución de Weibull podemos estar interesados no en los parámetros de dicha distribución, sino en estimar cuál es la altura media de ola; o en estimar qué proporción de las olas superará los cuatro metros o quedará por debajo de un metro. Estas cantidades, en general, podrán ponerse como función de los parámetros de la distribución de probabilidad de la altura de ola. Si la estimación de los parámetros de la distribución se

ha llevado a cabo mediante el método de máxima verosimilitud, los siguientes teoremas nos permiten obtener estimaciones de las funciones de interés, e intervalos de confianza, a partir de los estimadores MV (de máxima verosimilitud) de los parámetros.

Teorema 5.2. *Sea $\mathfrak{X} = \{X_1, X_2, \ldots, X_n\}$ una muestra de n observaciones independientes de una variable aleatoria con función de densidad $f(x)$, que depende de un parámetro $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$. Sea $L_{\mathfrak{X}}(\theta)$ la función de verosimilitud de θ dada la muestra $\mathfrak{X}$, y sea $g(\theta)$ una función de $\mathbb{R}^p$ en $\mathbb{R}^k$, $(1 \leq k \leq p)$. Si $\hat{\theta}$ es un estimador MV de θ, entonces $g\left(\hat{\theta}\right)$ es un estimador MV de $g(\theta)$.*

Teorema 5.3. *En las condiciones del teorema anterior, si el valor de parámetro $g(\theta)$ es una función continua y diferenciable, cuando $n \to \infty$:*

$$g\left(\hat{\theta}\right) \approx N\left(g\left(\theta\right), \sigma_g\left(\hat{\theta}\right)\right)$$

siendo $\hat{\theta}$ el estimador MV de θ, y

$$\sigma_g^2\left(\hat{\theta}\right) = \Delta g\left(\hat{\theta}\right)\left\{-H\left(\hat{\theta}\right)\right\}^{-1}\Delta g\left(\hat{\theta}\right)^t$$

$$\Delta g\left(\hat{\theta}\right) = \left(\frac{\partial g}{\partial \theta_1}, ..., \frac{\partial g}{\partial \theta_p}\right)\Big|_{\theta=\hat{\theta}}$$

En estas condiciones, el intervalo de confianza a nivel $1-\alpha$ para $g(\theta)$, basado en el estimador de máxima verosimilitud $\hat{\theta}$ sería:

$$\left[g\left(\hat{\theta}\right) \pm z_{\alpha/2}\sigma_g\left(\hat{\theta}\right)\right]$$

Veamos, a modo de ejemplo, como aplicar estos teoremas para estimar la probabilidad de que la altura de ola supere los 4 metros. Bajo el supuesto de que la altura de ola sigue una distribución $W(\kappa, \lambda)$, la probabilidad de que una ola supere una altura arbitraria h es:

$$g(h) = P(X > h) = \exp\left(-\left(h/\eta\right)^{\kappa}\right) = g_h(\kappa, \eta)$$

1. Implementamos esta función en R , considerando $\theta = (\kappa, \eta)$

```
> g = function(theta, altura) {
      exp(-(altura/theta[2])^theta[1])
  }
```

2. Obtenemos $g\left(\hat{\theta}\right)$ evaluando esta función para $altura = 4$ metros, y utilizando el estimador $\hat{\theta} = (\hat{\kappa}, \hat{\eta}) = (2{,}622, 3{,}427)$ obtenido anteriormente:

```
> gt = g(theta = EMV$par, altura = 4)
> gt

[1] 0.2230288
```

3. Calculamos el gradiente $\Delta g\left(\hat{\theta}\right)$ utilizando la función *grad()* que se encuentra en la librería *numDeriv*:

```
> library(numDeriv)
> Deltag = grad(g, EMV$par, altura = 4)
> Deltag

[1] -0.05178627  0.25608118
```

4. Calculamos $\sigma_g\left(\hat{\theta}\right) = \sqrt{\Delta g\left(\hat{\theta}\right)\left\{-H\left(\hat{\theta}\right)\right\}^{-1}\Delta g\left(\hat{\theta}\right)^t}$:

```
> sg = sqrt(t(Deltag) %*% (-Hinv) %*% Deltag)
> sg

           [,1]
[1,] 0.06111265
```

5. Por último construimos el intervalo de confianza para $g(\theta)$:

```
> conf = 0.95
> z = qnorm(1 - (1 - conf)/2)
> gt + c(-1, 1) * z * sg

[1] 0.1032502 0.3428074
```

De esta forma estimamos que la probabilidad de que una ola supere los 4 metros de altura es 0.223; y además con un 95 % de confianza podemos afirmar que dicha probabilidad se encuentra en el intervalo $[0{,}1033, 0{,}3428]$. Dicho de otra manera, podemos esperar que el 22.3 % de las olas supere los 4 metros, si bien dada la incertidumbre del muestreo, con un 95 % de confianza este porcentaje podría encontrarse en realidad entre el 10.33 % y el 34.28 %.

5.16. Tamaño de la muestra.

Los intervalos de confianza nos permiten determinar el tamaño de muestra necesario para estimar un parámetro con una precisión predeterminada. Para ello, el procedimiento general consiste en fijar el error máximo ε que estamos dispuestos a cometer en la estimación, y el nivel de confianza $1-\alpha$ de la misma. A continuación, utilizando el intervalo de confianza más adecuado para el parámetro que se desea estimar, se iguala el margen de error de dicho intervalo al valor de ε y se despeja el valor de n, que será entonces el tamaño de muestra buscado.

En caso de que el parámetro a estimar dependa de dos muestras de tamaños respectivos n_1 y n_2 (por ejemplo en la estimación de la diferencia de medias, la diferencia de proporciones o el cociente de varianzas), consideraremos que $n_1 = n_2 = n$ y utilizaremos el mismo tamaño muestral para ambas muestras.

Asimismo, en caso de que el intervalo de confianza dependa de alguna cantidad que no se conoce antes de llevar a efecto el muestreo (caso de la varianza muestral o la proporción muestral), podemos recurrir a varias alternativas:

- Tomar una muestra piloto (usualmente una muestra de tamaño reducido que sea posible tomar de forma rápida y con un coste de tiempo y recursos dentro de lo razonable y/o disponible) que nos proporcione un valor aproximado de dicha cantidad.
- Buscar en la literatura referente al problema que nos ocupa valores que puedan resultar razonables en nuestro caso para esa cantidad desconocida.
- Utilizar como valor de n el que resultaría del intervalo más grande posible. Por ejemplo, al estimar una proporción, la longitud del intervalo depende del valor de$\hat{p}$; dicho valor no se conoce antes de tomar la muestra, pero el intervalo más grande (el peor de los posibles) se obtiene cuando $\hat{p} = 1/2$. Este valor es el que se utilizará para despejar n.
- Determinar el tamaño de muestra no para un error absoluto, sino para un error relativo.

5.16.1. Tamaño de muestra para la estimación de la media de una población normal

En este caso, el intervalo de confianza para μ es

$$\left(\bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

Por tanto, si queremos estimar μ con un error máximo ε igualamos:

$$t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} = \varepsilon$$

y despejamos n:

$$n = \left(t_{n-1,\alpha/2}\frac{S}{\varepsilon}\right)^2$$

Obviamente, como no se conoce n, no puede calcularse el valor de $t_{n-1,\alpha/2}$. Ahora bien, teniendo en cuenta que para valores grandes de n, la t de Student se aproxima a la normal (y grande en este contexto puede ser del orden de 30), en la ecuación anterior se sustituye el valor $t_{n-1,\alpha/2}$ por $z_{\alpha/2}$ y por tanto el tamaño de la muestra es:

$$n = \left(z_{\alpha/2}\frac{S}{\varepsilon}\right)^2$$

donde el valor de S (desviación típica) habrá de obtenerse por alguno de los métodos señalados anteriormente (muestra piloto o información publicada en la literatura).

Otra alternativa que puede emplearse para resolver este problema es tener en cuenta que:

$$\mu \in \left(\bar{X} \pm t_{n-1,\alpha/2}\tfrac{S}{\sqrt{n}}\right) \Leftrightarrow \mu - \bar{X} \in \left(-t_{n-1,\alpha/2}\tfrac{S}{\sqrt{n}}, t_{n-1,\alpha/2}\tfrac{S}{\sqrt{n}}\right) \Leftrightarrow$$
$$\Leftrightarrow \tfrac{\mu-\bar{X}}{S} \in \left(-\tfrac{t_{n-1,\alpha/2}}{\sqrt{n}}, \tfrac{t_{n-1,\alpha/2}}{\sqrt{n}}\right) \Leftrightarrow \left|\tfrac{\mu-\bar{X}}{S}\right| \leq \tfrac{t_{n-1,\alpha/2}}{\sqrt{n}}$$

y determinar el tamaño n de la muestra de forma que la diferencia relativa (en términos de la desviación típica) entre la media μ desconocida y su estimación muestral $\bar{X}$, sea inferior a un valor δ fijado de antemano, esto es:

$$\left|\frac{\mu - \bar{X}}{S}\right| \leq \delta$$

Para ello basta igualar:

$$\frac{t_{n-1,\alpha/2}}{\sqrt{n}} = \delta$$

y despejar n. Igual que antes, sustituimos $t_{n-1,\alpha/2}$ por $z_{\alpha/2}$, por lo que obtenemos:

$$n = \left(\frac{z_{a/2}}{\delta}\right)^2$$

5.16.2. Tamaño de muestra para la estimación de la varianza de una población normal

El intervalo de confianza a nivel $1-\alpha$ para estimar esta varianza es:

$$\sigma^2 \in \left(\frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}} \right)$$

Si queremos estimar σ^2 con un error máximo ε deberemos determinar n de forma que

$$\frac{1}{2}\left(\frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}} - \frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}} \right) = \varepsilon$$

de donde:

$$(n-1)\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} - \frac{1}{\chi^2_{n-1,\alpha/2}} \right) = \frac{2\varepsilon}{S^2}$$

Esta ecuación no puede resolverse explícitamente, por lo que habrá que probar diversos valores de n. Del mismo modo que en el caso anterior, S^2 no se conoce antes de llevar a cabo el muestreo, por lo que su valor habrá de sustituirse por un valor calculado sobre una muestra piloto, o por un valor máximo razonable que pueda encontrarse en la bibliografía referente al problema en estudio. Otra alternativa es observar que del intervalo de confianza original se sigue que con confianza $1-\alpha$:

$$\frac{\sigma^2}{S^2} \in \left(\frac{(n-1)}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}} \right)$$

y podemos calcular un tamaño de muestra para que, en términos relativos,

$$\left| \frac{\sigma^2}{S^2} - 1 \right| \leq \delta$$

Para conseguir este objetivo bastará con elegir n de tal forma que:

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

En R podemos resolver este problema utilizando la función `uniroot()` para encontrar el valor de n tal que:

$$(n-1)\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} - \frac{1}{\chi^2_{n-1,\alpha/2}} \right) - 2\delta = 0$$

Así, por ejemplo, para $\delta = 0{,}4$ y $\alpha = 0{,}05$ el tamaño de muestra necesario puede obtenerse mediante:

```
> dif = function(n, alfa, delta) {
      (n - 1) * (1/qchisq(alfa/2, n - 1) -
          1/qchisq(1 - alfa/2, n - 1)) -
          2 * delta
 }
> n = uniroot(dif, c(2, 1000), alfa = 0.05,
      delta = 0.5)$root
> ceiling(n)

[1] 39
```

La función *ceiling()* se utiliza simplemente para redondear por exceso, ya que habitualmente el valor de n resultante del cálculo anterior no es entero.

5.16.3. Tamaño de muestra para la estimación de la diferencia de medias de poblaciones normales independientes

El intervalo de confianza para la diferencia de medias en poblaciones normales es de la forma:

$$\left((\bar{X}_1 - \bar{X}_2) \mp t_{m,\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

Si hacemos $n = n_1 = n_2$ y aproximamos $t_{m,\alpha/2} \approx z_{\alpha/2}$, el tamaño de muestra n para un error máximo ε se obtiene de:

$$\varepsilon = z_{\alpha/2} \sqrt{\frac{S_1^2 + S_2^2}{n}}$$

esto es:

$$n = \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 \left(S_1^2 + S_2^2 \right)$$

Como siempre, S_1^2 y S_2^2 habrán de obtenerse de una muestra piloto o de alguna otra fuente de información disponible.

5.16.4. Tamaño de muestra para la estimación de una proporción.

Ya hemos visto que si $np > 5$ y $n\,(1-p) > 5$, el intervalo de confianza a nivel $1 - \alpha$ para π es aproximadamente:

$$\pi \in \left(p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right)$$

Entonces, si queremos estimar π con un error inferior a un valor prefijado ε deberemos despejar n de:

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = \varepsilon \Rightarrow n = \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 p\,(1-p)$$

Obviamente, como p es desconocido, esta ecuación no resulta útil. Si se dispone de una estimación previa p (obtenida en una muestra piloto, en una revisión bibliográfica o en un problema similar) puede sustituirse dicha estimación en la fórmula anterior. Otra alternativa consiste en observar que en esta fórmula el valor más grande de n se obtiene cuando $p = 1/2$ (ya que $p(1-p)$ representa una parábola invertida con su máximo en ese valor). Por tanto, en el peor de los casos, si no se tiene información sobre p, sustituiremos el valor $p = 1/2$ en la ecuación anterior, en cuyo caso, el tamaño de muestra es:

$$n = \left(\frac{z_{\alpha/2}}{2\varepsilon}\right)^2$$

que garantiza un error de estimación inferior a ε cualquiera que sea el valor de p.