FACULTADE DE ENFERMARÍA

Estadística y metodología de la investigación
Curso 2012-2013
Pedro Faraldo, Beatriz Pateiro

# Tema 4. Estimación de parámetros



## 1 Estimación puntual

En este tema se trata el problema de la estimación de parámetros. Para ello, comenzamos recordando algunos conceptos básicos de la inferencia estadística que ya fueron introducidos en el tema anterior, y que serán necesarios para la construcción y el estudio de los estimadores:

- **Población:** conjunto homogéneo de individuos sobre los que se estudian características observables con el objetivo de extraer alguna conclusión. Por abuso de notación, en ocasiones nos referimos a la distribución que sigue la variable de interés en vez de al conjunto de individuos. Así, se dice que estamos ante una población Normal indicando que la variable que nos interesa sigue una distribución normal.
- **Parámetro:** característica de la población, como la media y la varianza (o desviación típica) en la distribución Normal o la probabilidad de éxito en la Binomial son parámetros. Si conocemos su valor (o si somos capaces de aproximarlo con suficiente precisión) podremos responder a cualquier pregunta sobre la distribución.
- **Estadístico:** cualquier función de la muestra. Por ejemplo, la media o la varianza muestrales son estadísticos.
- **Estimadores:** son estadísticos independientes de los parámetros de la población, y que se utilizan para aproximarlos. Si $\theta$ es el parámetro de interés, el estimador se denotará por $\hat{\theta}$. En el caso de una población Normal, podemos considerar la media muestral como estimador de la media poblacional (es decir, $\overline{X} = \hat{\mu}$) y la varianza muestral como estimador de la varianza poblacional ($s^2 = \hat{\sigma}^2$). Para una distribución $Bi(m, p)$, donde $m$ denota el número de pruebas de Bernoulli, la proporción $p$ se puede estimar a partir de la proporción poblacional (que denotaremos por $\hat{p}$). Por tanto, $\overline{X}$, $s^2$ y $\hat{p}$ son **estimadores puntuales** de $\mu$, $\sigma^2$ (en distribución Normal) y $p$ (en distribución Binomial), respectivamente.
- **Método de muestreo:** procedimiento para seleccionar una muestra. Si en una población queremos obtener una muestra de un cierto tamaño $n$ (siendo $n$ menor que el tamaño de la población), la manera de obtener esta muestra no es única. En este tema, consideraremos muestras aleatorias simples (m.a.s.).

Las estimaciones puntuales de los parámetros se obtienen a partir de una muestra aleatoria simple $X_1, \ldots, X_n$ de la variable $X$. Si calculamos el valor del estimador a partir de distintas muestras, los resultados que obtendremos serán diferentes. Es decir, los estimadores, al estar construidos a partir de muestras aleatorias, son aleatorios y en consecuencia, tienen una distribución. La distribución de los estimadores se denomina distribución en el muestreo. Describimos a continuación los estimadores para la proporción (en distribución Binomial) y para la media y la varianza (en distribución Normal) y sus respectivas distribuciones en el muestreo, que serán tenidas en cuenta a la hora de construir los intervalos de confianza.

## 1.1 Estimación de la proporción en la distribución $Bi(m, p)$

Supongamos que tenemos una variable $X \sim Bi(m, p)$, donde $m$ denota el número de pruebas de Bernoulli (conocido) y $p$ es la probabilidad de éxito (desconocida). Nótese que en el Tema 3, denotamos por $n$ el número de pruebas de Bernoulli. En este tema, $n$ es el tamaño muestral. Para estimar $p$, seleccionamos una m.a.s. $X_1, \ldots, X_n$ de variables $Bi(1, p) = Ber(p)$. Como estamos interesados en la probabilidad del éxito, consideraremos una muestra con 1 si es éxito y 0 si es fracaso. La proporción muestral viene dada por:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

La proporción muestral $\hat{p}$ es una variable aleatoria y, para $n$ suficientemente grande, su distribución es Normal, como consecuencia del Teorema Central del Límite:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

Además, se puede interpretar este resultado de la siguiente forma:

- Como $\hat{p}$ sigue una distribución Normal, y esta es una distribución simétrica, los valores de $\hat{p}$ se distribuirán con la misma probabilidad por encima y por debajo de su media.
- La media de la proporción muestral es $\mathbb{E}(\hat{p}) = p$, la proporción teórica o poblacional. Por tanto, los valores de $\hat{p}$ se distribuyen simétricamente alrededor de $p$, que es desconocido.
- En la varianza de $\hat{p}$ aparece el tamaño de la muestra $n$ dividiendo. Esto indica que, al aumentar el tamaño muestral $n$, disminuye la varianza de $\hat{p}$, por lo que la distribución de $\hat{p}$ se concentra más alrededor de su media.
- **Error típico:** el error típico (ET) de un estimador simétrico es su desviación típica. En el caso de $\hat{p}$, su error típico es:

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

Nótese que $p$ es desconocido, y en consecuencia $ET(\hat{p})$ también lo es. Si queremos aproximarlo, podemos substituir $p$ por $\hat{p}$.

Por ejemplo, si tenemos una variable $X \sim Bi(15, p)$ y queremos estimar el valor de $p$, tomamos 500 muestras de tamaño 100 ($X_1, \ldots, X_{100}$) y calculamos la proporción muestral en cada una de ellas, obteniendo 500 valores para $\hat{p}$. Si los representamos (se muestran en la Figura 1), podemos ver que los valores se distribuyen simétricamente alrededor de 0.7. También se puede ver que la curva Normal correspondiente (media 0.7 y varianza $0.7 * 0.3/100$) se ajusta a la gráfica del histograma.

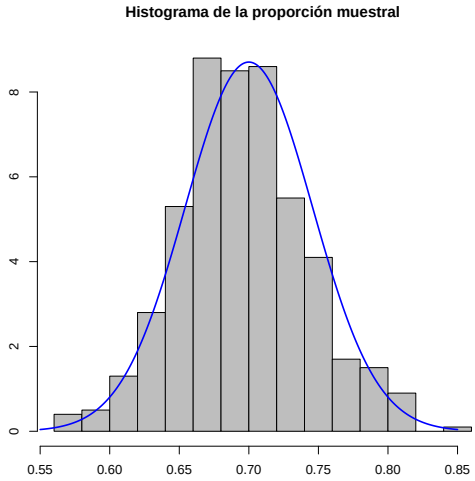


Figura 1: Distribución de la proporción muestral $\hat{p}$, a partir de 500 muestras de tamaño $n = 100$. Distribución normal de media $p = 0.7$ y varianza $p(1-p)/n$.

## 1.2 Estimación en poblaciones Normales $N(\mu, \sigma^2)$

Una v.a. $X \sim N(\mu, \sigma^2)$ queda caracterizada por dos parámetros: la media $\mu$ y la varianza $\sigma^2$ (o la desviación típica $\sigma$). A continuación, introduciremos los estimadores para estos parámetros y sus distribuciones en el muestreo. Es importante resaltar que tanto para la estimación de $\mu$ como de $\sigma^2$, debemos tener en cuenta el efecto del tamaño muestral y además, al estimar la media, también debemos ver si la varianza poblacional es conocida o desconocida.

### 1.2.1 Estimación de la media $\mu$.

Supongamos que disponemos de $X_1, \ldots, X_n$ una m.a.s. de $X \sim N(\mu, \sigma^2)$. La media poblacional $\mu$ se puede estimar con la media muestral $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$, cuya distribución en el muestreo también es Normal:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Además, dado que tenemos una Normal, podríamos tipificarla y obtener una $N(0,1)$:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{1}$$

La distribución es consecuencia de que la suma de variables Normales es también una variable Normal. Este resultado es válido **si la varianza poblacional $\sigma^2$ es conocida**. Esta distribución se puede interpretar de la siguiente forma:

- $\overline{X}$ se distribuye simétricamente (ya que su distribución es Normal) alrededor de su media, que es $\mathbb{E}(\overline{X}) = \mu$ la media poblacional o teórica.
- El tamaño muestral aparece dividiendo en la varianza, con lo que, al aumentar $n$, la distribución de $\overline{X}$ se concentra más alrededor de $\mu$, como se puede observar en la Figura 2. Los histogramas y las correspondientes densidades normales, están centrados en la media real de la población de la población, pero se puede apreciar que la concentración alrededor de este valor aumenta con el tamaño muestral.

- **Error típico:** la media muestral $\overline{X}$ es un estimador simétrico para $\mu$, por lo que podemos calcular su error típico, que viene dado por:

$$ET(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

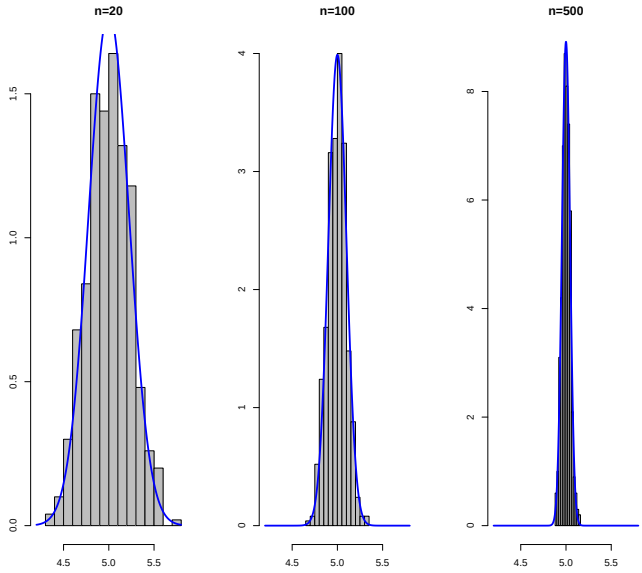


Figura 2: Distribución de la media muestral $\overline{X}$, a partir de 500 muestras de tamaño $n = 20$, $n = 100$, $n = 500$. Distribución normal de media $\mu = 5$ y varianza $1/n$.

**Si la varianza $\sigma^2$ es desconocida** no podemos utilizar la distribución obtenida en (1), y debemos substituir $\sigma^2$ por un estimador. La varianza $\sigma^2$ puede ser estimada por la **varianza muestral**:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2}$$

o por la **cuasivarianza**:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3}$$

Estos estimadores se verán con más detalle en la siguiente sección. Entonces, si queremos estimar la media $\mu$ a partir de una m.a.s. $X_1, \ldots, X_n$ y no conocemos la varianza, en la expresión (1) substituimos $\sigma^2$ (equivalentemente, $\sigma$) por un estimador de la siguiente manera:

$$\frac{\overline{X} - \mu}{s/\sqrt{n-1}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim \begin{cases} t_{n-1} & \text{si } n \leq 30, \\ N(0,1) & \text{si } n > 30, \end{cases}$$

donde $t_{n-1}$ denota una distribución T-Student, con $(n-1)$ grados de libertad. Esta distribución es simétrica y se aproxima a la $N(0,1)$ para $n$ suficientemente grande (véase Figura 3).

Al igual que en el caso anterior (con varianza conocida), seguimos teniendo un estimador simétrico, pero el **error típico** vendrá ahora dado por:

$$ET(\overline{X}) = \frac{s}{\sqrt{n-1}} = \frac{S}{\sqrt{n}}$$

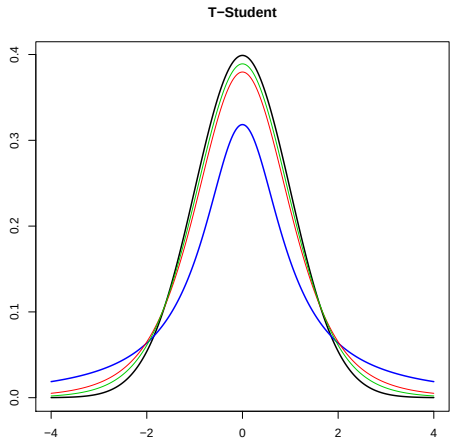


Figura 3: Distribución $t$ de Student con distintos grados de libertad. Azul: $n = 1$ (Cauchy); roja: $n = 5$; verde: $n = 10$; negra: $N(0, 1)$.

---

En resumen, cuando queremos estimar la media $\mu$ en una población Normal, debemos distinguir los siguientes casos:

1. Si la varianza $\sigma^2$ es conocida, entonces: $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$
2. Si la varianza $\sigma^2$ es desconocida y $n > 30$: $\dfrac{\overline{X} - \mu}{s/\sqrt{n-1}} = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim N(0, 1)$
3. Si la varianza $\sigma^2$ es desconocida y $n \leq 30$: $\dfrac{\overline{X} - \mu}{s/\sqrt{n-1}} = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

---

### 1.2.2 Estimación de la varianza $\sigma^2$

En la estimación de la media se hace necesario utilizar un estimador de la varianza $\sigma^2$, en caso de que esta no sea conocida. Para ello podemos utilizar la varianza muestral $s^2$ o la cuasivarianza muestral $S^2$, que vienen dadas por (2) y (3), respectivamente. Es fácil ver la relación entre ellas, ya que:

$$s^2 = \frac{n-1}{n} S^2, \quad \text{o bien} \quad S^2 = \frac{n}{n-1} s^2.$$

Estos dos estimadores sólo se distinguen en su denominador, y para $n$ grande, no hay diferencias importantes entre ellos. Como la varianza muestral o la cuasivarianza proporcionarán valores (aleatorios) positivos, su distribución tendrá como soporte $[0, \infty)$. Esta distribución será la distribución Chi-cuadrado $\chi^2$ (distribución *ji-cuadrado*).
Si $X_1, \ldots, X_n$ es una m.a.s. de variables normales con varianza $\sigma^2$, entonces:

$$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \text{o bien} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

donde $\chi^2_{n-1}$ es una distribución Chi-cuadrado con $(n-1)$ grados de libertad. Esta distribución es asimétrica y con soporte la semirrecta real positiva, como puede verse en la Figura 4.
Esta distribución es necesaria cuando el tamaño de la muestra es pequeño. Para $n$ suficientemente grande, podemos aproximar una distribución $\chi^2_n$ (Chi-cuadrado con $n$ grados de libertad) por una $N(n, 2n)$.

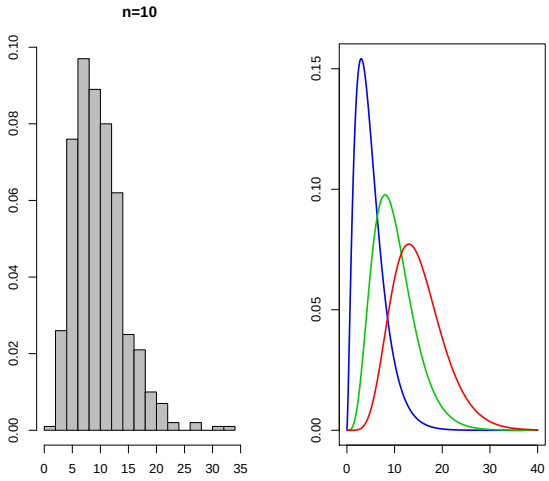


Figura 4: Distribución en el muestreo de la suma de los cuadrados de $n = 10$ variables Normales estándar (siguen una distribución $\chi^2$). Gráficas de la densidad $\chi^2_n$: línea azul: $n = 10$; línea verde: $n = 10$; línea roja: $n = 15$.

## 2 Estimación por intervalos de confianza

En algunas ocasiones, no sólo estamos interesados en dar una estimación puntual del valor del parámetro desconocido, y el objetivo se centra en obtener un rango de valores entre los que se encuentre el parámetro de la distribución con una cierta probabilidad, es decir, un **intervalo de confianza**.
Construiremos intervalos de confianza para la proporción $p$ en la distribución Binomial y para la media $\mu$ en la distribución Normal. Los estimadores que hemos introducido para la proporción y la media ($\hat{p}$ y $\overline{X}$, respectivamente) son simétricos y podemos calcular o aproximar su error típico. La fórmula general para el cálculo de intervalos de confianza será:

$$\text{Estimador} \pm \text{Cuantil} \cdot ET(\text{Estimador}) \tag{4}$$

De este modo, obtendremos intervalos de confianza **centrados en el estimador**, y cuya **amplitud** vendrá determinada por su **error típico** (donde interviene el tamaño de la muestra) y por el **cuantil de la distribución** correspondiente, que estará relacionado con la cobertura del intervalo.

### 2.1 Intervalos de confianza para la proporción $p$

Consideremos $\hat{p}$, proporción muestral, como estimador de $p$. A partir de la ecuación (4) podemos construir un intervalo de confianza de nivel (cobertura) $(1 - \alpha)$ para $p$:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{o bien} \quad \left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

En este caso, el **error típico** se aproxima por $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ y, dado que $\hat{p}$ tiene una distribución Normal, consideramos los cuantiles de una $N(0, 1)$. En concreto, para los intervalos de confianza usuales, se tiene:

- IC para $p$ al 90 %: $\hat{p} \pm 1.64\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- IC para $p$ al 95 %: $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- IC para $p$ al 99 %: $\hat{p} \pm 2.57\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

ya que para una cobertura $1-\alpha = 0.9 = 90\,\%$ ($\alpha = 0.1$), el cuantil $z_{1-\alpha/2} = 1.64$. Del mismo modo, para una cobertura del $1-\alpha = 0.95 = 95\,\%$ ($\alpha = 0.05$) el cuantil es $z_{1-\alpha/2} = 1.96$ y para un cobertura del $1-\alpha = 0.99 = 99\,\%$ ($\alpha = 0.01$) el cuantil es $z_{1-\alpha/2} = 2.57$.

## 2.2 Intervalos de confianza para la media $\mu$

Sea $X \sim N(\mu, \sigma^2)$ y consideremos $X_1, \ldots, X_n$ una m.a.s. de $X$. Si estamos interesados en obtener intervalos de confianza para la media $\mu$, tendremos que tener en cuenta las siguientes situaciones:

1. **La varianza $\sigma^2$ es conocida.** En ese caso, el IC para $\mu$ viene dado por:

$$\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

   donde $z_{1-\alpha/2}$ es el cuantil de una $N(0,1)$ que tomará valores 1.64 para cobertura del 90 %, 1.96 para cobertura del 95 % y 2.57 para cobertura del 99 % (al igual que en los intervalos para la proporción que vimos en la sección anterior).

2. **La varianza $\sigma^2$ es desconocida pero $n$ es grande.** Cuando la varianza no es conocida, la distribución de la media $\overline{X}$ es una T-Student, que para tamaño muestral $n \geq 30$ se puede aproximar por una $N(0,1)$. En este caso, se debe aproximar el error típico obteniendo el siguiente intervalo de confianza:

$$\overline{X} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n-1}}, \quad \text{o bien} \quad \overline{X} \pm z_{1-\alpha/2}\frac{S}{\sqrt{n}}$$

   donde nuevamente $z_{1-\alpha/2}$ es el cuantil de una $N(0,1)$.

3. **La varianza $\sigma^2$ es desconocida y $n$ es pequeño.** En este caso, debemos considerar los cuantiles de la distribución T-Student, quedando el intervalo de confianza como:

$$\overline{X} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n-1}}, \quad \text{o bien} \quad \overline{X} \pm t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}$$

   donde $t_{n-1,1-\alpha/2}$ son los correspondientes cuantiles de una distribución T-Student con $(n-1)$ grados de libertad. Estos cuantiles están tabulados.

En el caso de los intervalos de confianza para $\mu$, se puede observar que para un nivel de significación fijo, a mayor varianza, mayor longitud del intervalo. El efecto contrario se produce a medida que aumenta el tamaño muestral. En ese caso, se reduce la longitud del intervalo. Cuando no conocemos la varianza, obtenemos también intervalos más amplios que en el caso de $\sigma^2$ conocida, ya que los cuantiles de la distribución $t$ son más extremos que para la $N(0,1)$.

## 2.3 Determinación del tamaño muestral

Dado un nivel de confianza $(1-\alpha)$, nos puede interesar saber qué tamaño muestral $n$ necesitamos para alcanzarlo en un intervalo de longitud $L$, suponiendo los mismos resultados en la muestra. En los intervalos que hemos introducido para $p$ y $\mu$, sus longitudes se puede calcular fácilmente. Estas longitudes dependerán del tamaño de muestra $n$, que se puede despejar como sigue:

- Longitud de un IC de nivel $(1-\alpha)$ para $p$:

$$L = 2z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Leftrightarrow n = \frac{4z^2_{1-\alpha/2}\hat{p}(1-\hat{p})}{L^2}$$

- Longitud de un IC de nivel $(1-\alpha)$ para $\mu$, con $\sigma^2$ conocida:

$$L = 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \Leftrightarrow n = \frac{4z^2_{1-\alpha/2}\sigma^2}{L^2}$$

- Longitud de un IC de nivel $(1-\alpha)$ para $\mu$, con $\sigma^2$ desconocida y $n$ grande:

$$L = 2z_{1-\alpha/2}\frac{S}{\sqrt{n}} \Leftrightarrow n = \frac{4z^2_{1-\alpha/2}S^2}{L^2}$$

$$L = 2z_{1-\alpha/2}\frac{s}{\sqrt{n-1}} \Leftrightarrow n = \frac{4z^2_{1-\alpha/2}s^2}{L^2} + 1$$

# 3 Anexo. Intervalos de confianza para los parámetros de una población

| **Intervalos de confianza para** $X \sim N(\mu, \sigma^2)$ | **Estadístico** | **Intervalo de nivel** $(1-\alpha)$ |
|---|---|---|
| Para $\mu$, con $\sigma^2$ conocida | $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ | $\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ |
| Para $\mu$, con $\sigma^2$ desconocida y $n > 30$ | $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim N(0,1)$<br>$\frac{\overline{X}-\mu}{s/\sqrt{n-1}} \sim N(0,1)$ | $\overline{X} \pm z_{1-\alpha/2}\frac{S}{\sqrt{n}}$<br>$\overline{X} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n-1}}$ |
| Para $\mu$, con $\sigma^2$ desconocida y $n \leq 30$ | $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$<br>$\frac{\overline{X}-\mu}{s/\sqrt{n-1}} \sim t_{n-1}$ | $\overline{X} \pm t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}$<br>$\overline{X} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n-1}}$ |
| **Intervalos de confianza para** $X \sim Bi(m,p)$ | **Estadístico** | **Intervalo de nivel** $(1-\alpha)$ |
| Para $p$, con $m$ conocido | $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$ | $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ |

Cuadro 1: Intervalos de confianza para los parámetros de una población.