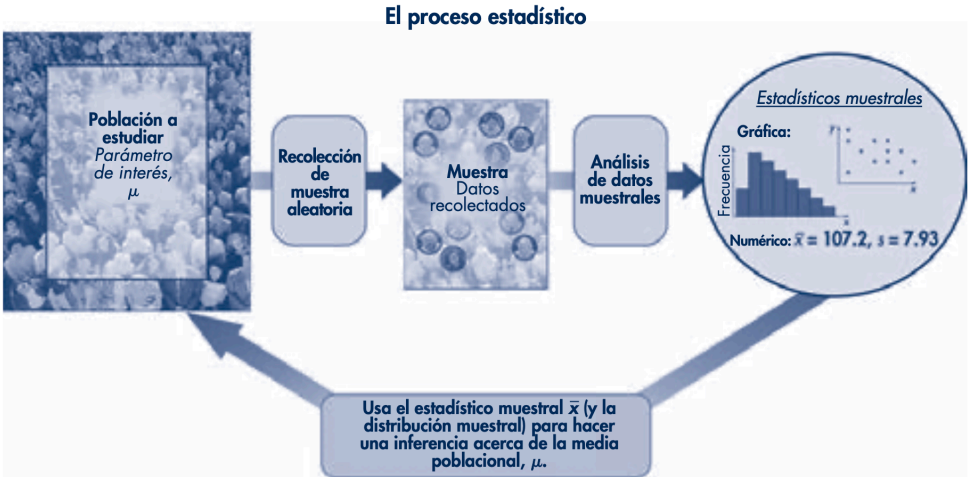
INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

El objetivo de la inferencia estadística es usar la información obtenida en los datos de una muestra para estimar parámetros de la población de la que se extrajo la muestra.



En este tema aprenderemos a responder algunos problemas de inferencia estadística, para ello nos vamos a apoyar:

1._DISTRIBUCIÓN DE LAS MEDIAS MUESTRALES. TEOREMA CENTRAL DEL LÍMITE.

Dada una población de media μ y desviación típica σ, no necesariamente normal, la distribución de las medias de la muestras de tamaño n:

- Tiene la misma media μ que la población
- Su desviación típica es $\sigma/\sqrt{n}$
- Cuando n >30 es prácticamente normal

El grado de aproximación de la distribución de medias muestras a una distribución normal depende de la población de partida y del tamaño de las muestras n.

POBLACIÓN

Media μ
Desviación típica σ

puede ser

Normal $X \sim (\mu, \sigma)$ → Normal $\bar{X} \sim (\mu, \frac{\sigma}{\sqrt{n}})$

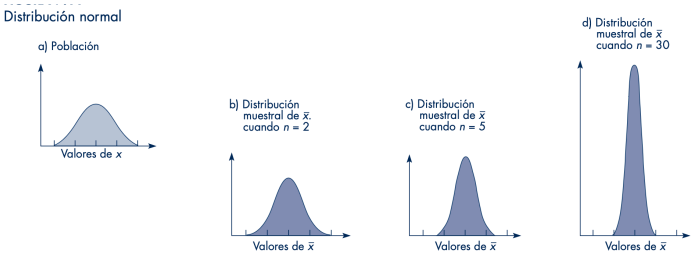
No Normal → Normal $\bar{X} \sim (\mu, \frac{\sigma}{\sqrt{n}})$

Mayor n mayor aproximación.

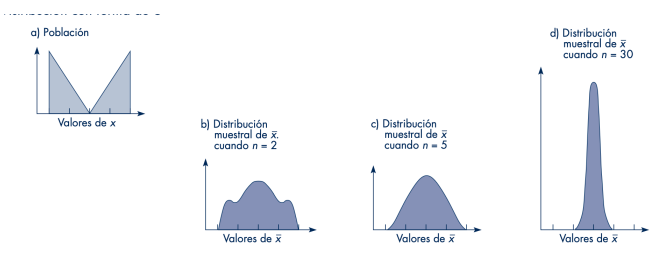
DISTRIBUCIÓN DE MEDIAS MUESTRALES

Media μ
Desviación típica $\frac{\sigma}{\sqrt{n}}$

Ejemplo: Población con distribución normal

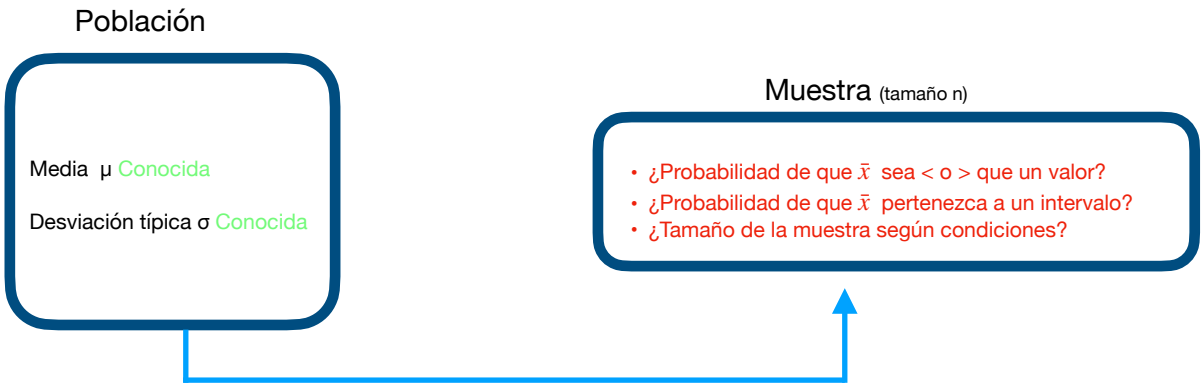


Ejemplo: Población con distribución no normal



Este teorema ya nos va a permitir responder a problemas de este tipo:

Dada una población de la que conocemos su media μ y desviación típica σ, podremos determinar la probabilidad de que la media de una muestra concreta $\bar{x}$ esté en cierto intervalo o sea mayor o menor que un valor concreto.

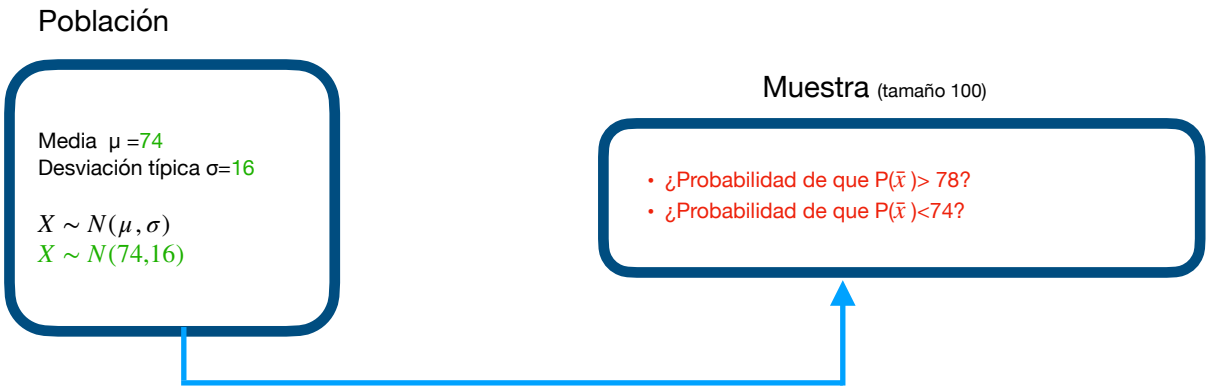


A partir de los datos de la población podremos determinar para la muestra:

Ejemplo: ABAU 2019

4. Logo de anos de utilizalo sábese que a puntuación dun test de uso habitual en certa rama industrial segue unha distribución normal de media 74 e desviación típica 16. Nunha empresa decídese realizalo a 100 dos seus empregados. **a)** Cal é a probabilidade de que se obteña unha media muestral superior a 78 puntos, de seguirse a pauta xeral? **b)** E a probabilidade de que a media muestral sexa inferior a 74 puntos?

Tomando los datos del problema, el esquema anterior quedaría así



Sabemos que las medias muestrales procedentes de una distribución normal, o no normal si la muestra tiene tamaño n>30, siguen una distribución normal de probabilidad $\bar{X} \sim (\mu, \frac{\sigma}{\sqrt{n}})$, para este caso $\bar{X} \sim (74, \frac{16}{\sqrt{100}})$

El problema se reduce al cálculo de probabilidades de una distribución normal, como hemos hecho anteriormente, tomando $\bar{X} \sim (74, \frac{16}{\sqrt{100}})$ $\quad \bar{X} \sim (74\ ,\ 1{,}6)$

Apartado a) $P(\bar{x} > 78) = P(Z > \frac{78-74}{1{,}6}) = P(Z > 2{,}5) = 1 - P(Z < 2{,}5) =$ (tabla)$= 1 - 0{,}9938 = 0{,}0062$

$P(\bar{x} > 78) = 0.0062$

Apartado b) $P(\bar{x} < 74) = 0{,}5$ (directamente porque es el valor de la media o mediante mismo proceso de arriba)

Ahora vamos a aprender a estimar parámetros de la de la población, a través de los datos obtenidos en una determinada muestra de tamaño n tomada de la población. Para ello también tenemos que saber:

2._ ESTIMACIÓN PUNTUAL Y ESTIMACIÓN POR INTERVALOS

Los parámetros de la población se pueden estimar a partir de los de la muestra de dos formas:

Estimación puntual: Deseamos conocer algo sobre la población, por ejemplo, la media… y para ello se selecciona de forma aleatoria una muestra. En ella podemos calcular esa media… A ese valor lo denominamos estimador o estimador puntual, y al hecho de hacerlo, una estimación puntual.
Con dicha estimación podremos inferir esa media sobre la población. Ya sabemos que no se puede asegurar que la población tenga esa media, sino que la tiene con una cierta probabilidad. Pero al hacerlo así se dice que hemos hecho una estimación puntual.
Sabemos que la media de la población será un valor cercano al de la media de la muestra, pero no sabemos, por sí solo, cuánto se aproxima. Por lo tanto este valor aislado no sirve de mucho.

Estimación por intervalos: a partir de una muestra de tamaño n, podemos estimar el valor de un parámetro de la población del siguiente modo:

- Dando un intervalo en el que confiamos que esté dicho parámetro. Se llama **intervalo de confianza**,
- Hallando la probabilidad de tal cosa ocurra, esta probabilidad se llama **nivel de confianza**.

La eficacia de esta estimación se manifiesta de dos formas:

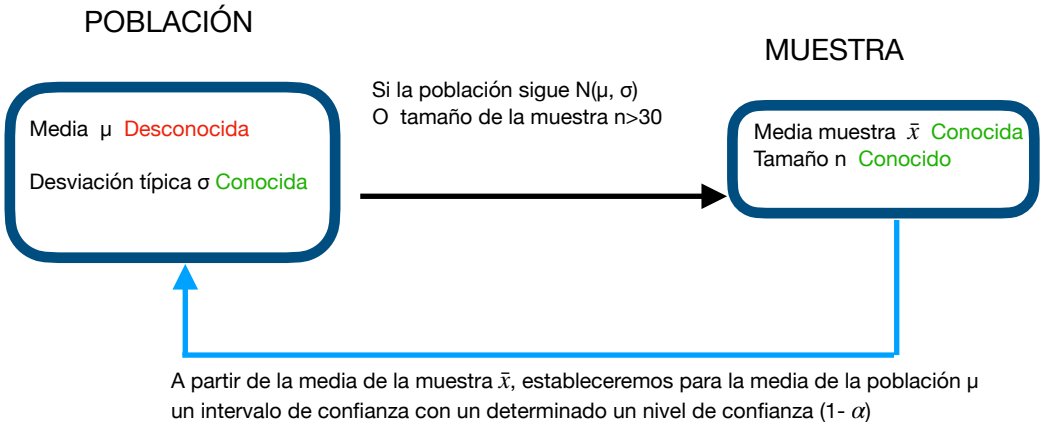
- El tamaño de intervalo (Cuanto más pequeño, mayor precisión)
- El nivel de confianza (mayor nivel de confianza, mayor seguridad en la estimación)

Además, cuanto mayor sea el tamaño de la muestra, mayor eficacia tendremos en la estimación.

Tamaño de la muestra, longitud del intervalo y nivel de confianza, son 3 variables estrechamente relacionadas.Conocidas dos de ellas obtendremos la tercera.

3._INTERVALO DE CONFIANZA PARA LA MEDIA

Vamos estimar la media poblacional μ, a través de los datos de la media $\bar{x}$ obtenidos en una muestra de tamaño n, mediante la estimación por intervalos y en el supuesto de que la desviación típica σ de la población es un dato conocido.

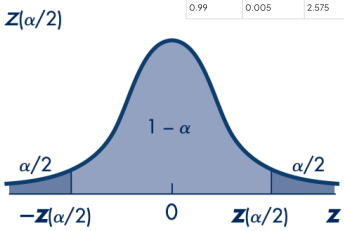


Intervalo de confianza con nivel de confianza (1-α)

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$\bar{x}$= media de la muestra
σ= desviación típica de la población
n= tamaño de la muestra
$z_{(\alpha/2)}$= valor crítico según nivel de confianza, se obtiene de la tabla de la distribución normal

$1-\alpha$	$\alpha/2$	$z_{\alpha/2}$
0.90	0.05	1.645
0.95	0.025	1.96
0.99	0.005	2.575



Vamos a resolver un problema de este tipo

Ejemplo: ABAU 2019

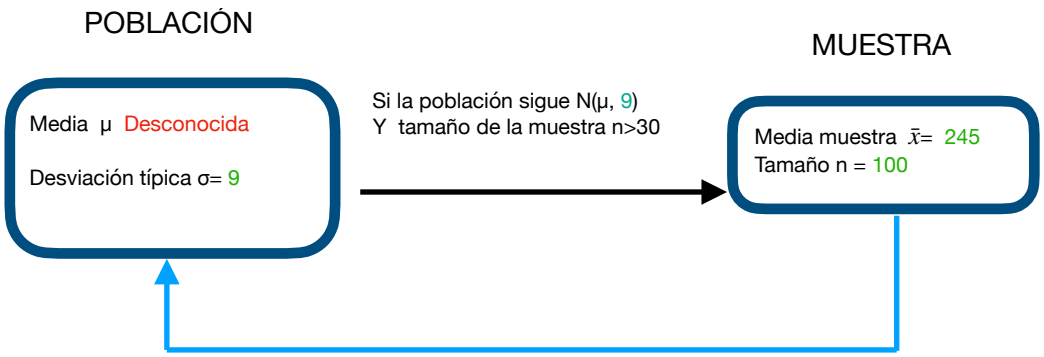
4. Un consumidor cre que o peso medio dun produto é distinto do que indica o envase. Para estudar este feito, o consumidor toma unha mostra aleatoria simple de 100 produtos nos que se observou un peso medio de 245 g. Suponse ademais que o peso do produto por envase segue unha distribución normal con desviación típica 9 g.

a) Constrúe un intervalo de confianza para o peso medio dese produto ao 95 % de confianza.

b) Cal sería o tamaño muestral mínimo necesario para estimar o verdadeiro peso medio a partir da media mostral cun erro de estimación máximo de 2 g e un nivel de confianza do 90 %?

Tomando los datos del problema, el esquema anterior quedaría así

Apartado a)



A partir de la media de la muestra $\bar{x}$=245 estableceremos para la media de la población μ un intervalo de confianza con un nivel de confianza 95%

Con un nivel de confianza 95% hallamos el valor de $z_{(\alpha/2)}$:

(1-α) =0,95 α= 0,05 $\alpha/2$= 0,025. $z_{(\alpha/2)}$= (tabla) 1,96

$$\left(\bar{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$ Sustituyendo: $$\left(245 - 1{,}96\cdot\frac{9}{\sqrt{100}}, 245 + 1{,}96\cdot\frac{9}{\sqrt{100}}\right)=$$

= (245-1,764 , 245+1,764) =(243,236 , 246,764)

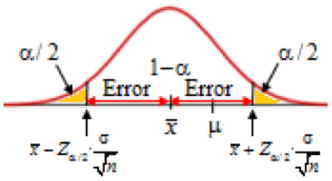
El intervalo de confianza del peso medio del producto con un nivel de confianza de 95% es $(243{,}236\ ,\ 246{,}764)$

Para resolver el apartado b y problemas en los que se relacione tamaño de la muestra, error y nivel de confianza Tenemos que saber lo siguiente:

4._RELACIÓN ENTRE EL NIVEL DE CONFIANZA, ERROR ADMISIBLE Y TAMAÑO DE LA MUESTRA

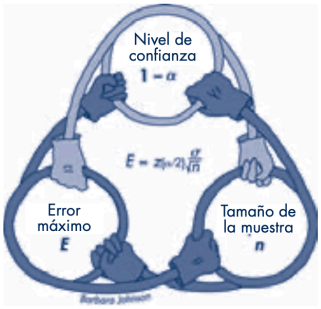
El valor $E = z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$ se llama **error máximo admisible**

Es el valor que se le suma o resta a la media para obtener el intervalo de confianza para un determinado nivel de confianza y tamaño de muestra.



$$\left(\bar{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

El error admisible depende del nivel de confianza y tamaño de la muestra de esta forma:

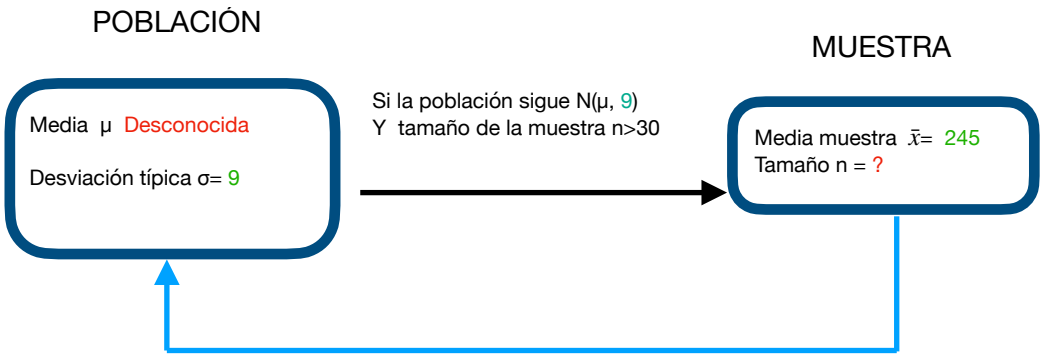


- Cuanto mayor es el tamaño n de la muestra menor es el error E (El intervalo será más estrecho y la estimación más afinada)
- Cuanto mayor sea el nivel de confianza (1-α) mayor es el error E (es decir cuanto más seguros querremos estar de la estimación más amplio será el intervalo, ya que a mayor (1-α) mayor valor de $z_{\alpha/2}$

El error E, el nivel de confianza (1-α) y el tamaño de la muestra son tres variables estrechamente relacionadas. Conocidas dos de ellas obtendremos la tercera despejando de la fórmula $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Ahora vamos a resolver el resto del problema

Apartado b) Calculamos el tamaño de la muestra para un error máximo de 2 gr con nivel de confianza 90%



Estableceremos el tamaño de la muestra para poder estimar la media de la población μ con un nivel de confianza 90% y error máximo E=2 gr.

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq 2$$

Para un nivel de confianza 90% , (1-α)= 0,9. α= 1 $\alpha/2$= 0,5. P ($z_{(\alpha/2)}$) < 0,95 $z_{(\alpha/2)}$= (tabla) 1,645

Sustituyendo los datos:

$$1{,}645 \cdot \frac{9}{\sqrt{n}} \leq 2 \quad \sqrt{n} \geq \frac{1{,}645 \cdot 9}{2} \quad \sqrt{n} \geq \frac{1{,}645 \cdot 9}{2} \quad \sqrt{n} \geq 7{,}4025 \quad n \geq 7{,}4025^2 \quad n \geq 54{,}79$$

Se necesitaría un tamaño de muestra de 55 productos como mínimo.

Por último vamos a aprender a resolver problemas de inferencia entorno a situaciones en las que solo hay dos posibles resultados: que pase algo o que no pase, éxito o fracaso, es decir, situaciones propias de un experimento binomial.

5._DISTRIBUCIÓN DE LA PROPORCIÓN DE LAS MUESTRAS

Cuando se trata de determinar la proporción de una población que posee cierto atributo (está de acuerdo / no está de acuerdo; vota al partido A / no vota al partido A; éxito / fracaso), su estudio es equiparable al de una distribución binomial.

Así pues, en una población, la proporción de individuos que posee cierta característica es p (éxito) la de que no la posea será $q=1-p$ (fracaso). Considerando todas las posibles muestras aleatorias de tamaño n que se pueden extraer de esa población, la proporción de individuos con dicha característica en cada una de las muestras de tamaño n es $\hat{p}$ (también se puede escribir pr).

La proporción $\hat{p}$ de individuos con dicha característica en las muestras de tamaño n, sigue una distribución normal de media p y desviación típica $\sqrt{\frac{p \cdot q}{n}}$. Es decir: $\hat{p} \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$

En cada muestra se puede obtener $\hat{p} = \frac{x}{n}$

siendo $x = n^o$ *de individuos con la característica* y $n = n^o$ *individuos de la muestra.*

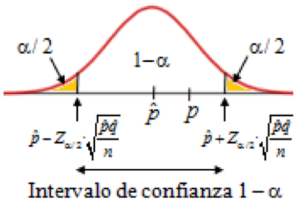
Recuerda que una distribución binomial se aproxima a una distribución normal si $n>30$ y $np>5$ $nq>5$

5.1_INTERVALO DE CONFIANZA PARA LA PROPORCIÓN

Queremos estimar la proporción p de individuos con cierta característica que hay en una población, a partir de una determinada muestra de tamaño n de la que se obtiene su proporción muestral $\hat{p}$. Para ello estableceremos que:

El intervalo de confianza para la proporción de la población, que se obtiene a partir de una muestra de tamaño *n*, con un nivel de confianza de $1-\alpha$, es:

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



1 - α	α/2	$z_{\alpha/2}$
0.90	0.05	1.645
0.95	0.025	1.96
0.99	0.005	2.575

El procedimiento para hallar los intervalos de confianza de la proporción, es análogo a lo visto anteriormente.

Recuerda de que si $\hat{p}$ es la proporción del éxito de la muestra (tener una característica), $\hat{q}$ es la proporción del fracaso (no tener la característica) y $\hat{q} = 1 - \hat{p}$ (Experimento binomial) .

Para que esta estimación sea válida se necesita y que $np>5$ $nq>5$ y que tamaño de la muestra sea $n>30$

5.2._ ERROR ADMISIBLE Y TAMAÑO DE LA MUESTRA

El error máximo admisible es $E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$ varía en relación al nivel de confianza y al tamaño de la muestra.

El Error E, el nivel de confianza (1-α) y el tamaño de la muestra n son tres variables estrechamente relacionadas. Conocidas dos de ellas obtendremos la tercera despejando de la fórmula.

Para estimar el tamaño de muestra necesario para dar un intervalo de confianza para la proporción de la población, sólo tenemos que despejar n en esa fórmula, pero puedo tener 3 posibles situaciones:

- Conozco sólo la proporción de la **muestra** $\hat{p}$, entonces despejaría directamente de esa fórmula
- Conozco por estudios previos la proporción de la **población** p, entonces usaría en la fórmula la desviación típica de la población $E = z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}}$
- No tengo datos de partida y quiero estimar el intervalo más amplio posible, entonces tomamos $p=q=0,5$ y los sustuimos en la fórmula.

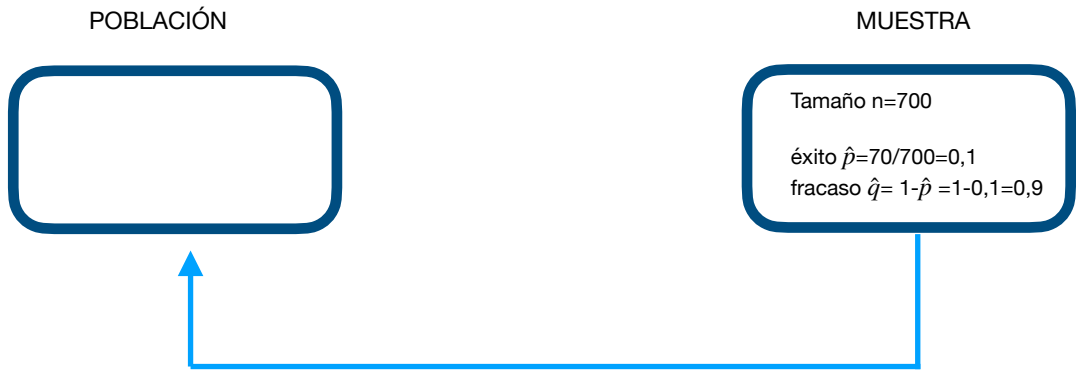
Vamos a ver un problema de este tipo.

ABAU 2018 EXTRAORDINARIA

4. Nun estanque deséxase estimar a porcentaxe de peixes dourados. Para iso, tómase unha mostra aleatoria de 700 peixes e atópase que exactamente 70 deles son dourados.

a) Acha, cun nivel de confianza do 99 %, un intervalo para estimar a proporción de peixes dourados no estanque **b)** No intervalo anterior, canto vale o erro de estimación? **c)** Considerando dita mostra, que lle ocorrería ao erro de estimación se aumentase o nivel de confianza? Xustifica a resposta.

Esquematizando el problema:



Establecer un intervalo de confianza para estimar la proporción de la población p con nivel confianza 99%

Apartado a)

El intervalo de confianza para la proporción de la población p a partir de la proporción de la muestra $\hat{p}$ tiene la forma:

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

sabemos que $\hat{p}=\frac{70}{700}$ = 0,1 Por tanto $\hat{q}$= 1-$\hat{p}$ =0,9

Calculamos $z_{(\alpha/2)}$ para un nivel de confianza 1-α=0,99 α=0,01 $\alpha/2$= 0,005. P ($z_{(\alpha/2)}$) < 0,995 $z_{(\alpha/2)}$= (tabla) 2,575

El intervalo quedaría sustituyendo $\left(0,1 - 2,575 \cdot \sqrt{\frac{0,1 \cdot 0,9}{700}}, 0,1 + 2,575 \cdot \sqrt{\frac{0,1 \cdot 0,9}{700}}\right) = (0,0708\ ,\ \ 0,1292)$

Con un nivel de confianza del 99% la proporción de peces dorados del estanque estaría entre 7,018% y 12,92%

Apartado b)

El error de estimación es: $E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$. Sustituyendo $E = 2,575 \cdot \sqrt{\frac{0,1 \cdot 0,9}{700}}$ $E = 0,292$ Error 2,92%

Apartado c)

A mayor nivel de confianza, mayor valor de $z_{\alpha/2}$ por tanto sustituyéndolo en la forma el error aumentaría.