Probabilidad y Estadística

PRÁCTICA:

Ejercicio 1:

- En una población se quiere estimar la proporción de quienes trabajan más de 8hs diarias. Se toma una muestra de 40 personas y se halla que hay 15 que trabajan más de 8hs. Con un nivel de confianza del 95% establezca un intervalo para la proporción de las personas de la población que trabajan más de 8hs.

- $\alpha = 0{,}05$ - $1 - \alpha = 0{,}95$ - $n = 40$

- Sea X una variable aleatoria con distribución de probabilidad binomial de parámetros n y π, con π desconocido.
- Sea p la proporción muestral.
- El estadístico Z estará definido por:

$$- Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}}$$

- De esta forma:

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

- Distribuyendo y despejando π:

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

- Siendo $p = \dfrac{15}{40} = 0{,}375$

- Siendo $z_{(1-\alpha/2)} = z_{(1-(0{,}05/2))} = z_{(1-0{,}025)} = z_{(0{,}975)} = 1{,}96$

$$- L_i = p - z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}} = (0{,}375) - (1{,}96)\sqrt{\frac{(0{,}375)(1-(0{,}375))}{100}}$$

$$- L_i = (0{,}375) - (1{,}96)\sqrt{\frac{(0{,}375)(0{,}625)}{100}} \simeq 0{,}28$$

$$- L_s = p + z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}} = (0{,}375) + (1{,}96)\sqrt{\frac{(0{,}375)(1-(0{,}375))}{100}}$$

$$- L_s = (0{,}375) + (1{,}96)\sqrt{\frac{(0{,}375)(0{,}625)}{100}} \simeq 0{,}47$$

$$- P\left[0{,}28 \leq \pi \leq 0{,}47\right] = 0{,}95 = 95\,\%$$

- Este ejercicio no ha sido controlado por un profesor de la cátedra pero viendo que $p = \dfrac{15}{40} = 0{,}375$ se encuentra dentro del intervalo, es factible que el resultado sea correcto.

Ejercicio 2:

- Se quieren comparar los tiempos de respuesta de dos dispositivos electrónicos. Se toma una muestra de 50 dispositivos del Tipo I y se halla que en promedio el tiempo es de 57 microsegundos. Para una muestra de 40 dispositivos Tipo II se encontró un promedio de 54 microsegundos. Los tiempos siguen la distribución Normal, con desviación estándar de 6 y 8 microsegundos respectivamente. Compruebe con un nivel de significación del 5% si el dispositivo II es más rápido que el I.

a) **Cuál es el parámetro cuestionado?**

b) **Qué estimador utilizaría para el test de hipótesis y que distribución tiene?**

c) **Establezca la Hipótesis Nula y la Alternativa y la Regla de decisión.**

d) **Cuál es, en palabras, la conclusión de la prueba?**

e) **Calcule el valor p y explique qué indica.**

f) **Qué es la potencia de la prueba? Qué dato haría falta para calcularla?**

g) **Si las muestras fueran chicas. Qué distribución debería usarse para el test?**

Dispositivos	***n***	**Media Muestral**	**Desviación Estándar**
Tipo I	50	57	6
Tipo II	40	54	8

- $n_1 = 50$ - $\overline{x}_1 = 57$ - $\sigma_1 = 6$
- $n_2 = 40$ - $\overline{x}_2 = 54$ - $\sigma_2 = 8$

- $\alpha = 0{,}05$ - $1 - \alpha = 0{,}95$

a) **Cuál es el parámetro cuestionado?**

- Si las varianzas de las poblaciones son desconocidas y, además, distintas y tamaños muestrales grandes, se parte de una variable aleatoria Z con fin de obtener el intervalo de confianza para la diferencia de medias. La interpretación del intervalo de confianza resultante permitirá determinar si las medias poblacionales de las dos distribuciones pueden suponerse iguales o no.

- $Z = \dfrac{\Delta\overline{X} - \Delta\mu}{\sigma_{\Delta\overline{X}}}$ - $\sigma_{\Delta\overline{X}} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

- $\Delta\overline{X} = \overline{x}_1 - \overline{x}_2$ - $\Delta\mu = \mu_1 - \mu_2$

- Por lo que el intervalo de confianza que se busca será:

$$\left[\Delta\overline{X} - Z_{(1-\alpha/2)} \ . \ \sigma_{\Delta\overline{X}} \ ; \ \Delta\overline{X} + Z_{(1-\alpha/2)} \ . \ \sigma_{\Delta\overline{X}}\right]$$

b) **Qué estimador utilizaría para el test de hipótesis y que distribución tiene?**

- Estimador con Distribución Normal:

$$\text{-}\ Z_p = \frac{\Delta\overline{X} - \overbrace{(\mu_a - \mu_b)}^{0}}{\sigma_{\Delta\overline{X}}}$$

c) **Establezca la Hipótesis Nula y la Alternativa y la Regla de decisión.**

- $\alpha = 0{,}05$

- $Z_{(1-\alpha/2)} = Z_{(1-(0{,}05/2))} = Z_{(1-0{,}025)} = Z_{(0{,}975)} = 1{,}96$

- $\Delta\overline{X} = \bar{x}_1 - \bar{x}_2 \quad \Rightarrow \quad \Delta\overline{X} = 57 - 54 \quad \Rightarrow \quad \Delta\overline{X} = 3$

- $\sigma_{\Delta\overline{X}} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} \quad \Rightarrow \quad \sigma_{\Delta\overline{X}} = \sqrt{\dfrac{6^2}{50} + \dfrac{8^2}{40}} \quad \Rightarrow \quad \sigma_{\Delta\overline{X}} = 1{,}523$

- $H_0 : \mu_1 = \mu_2 \quad \rightarrow$ Hipótesis Nula

- $H_0 : \mu_1 \neq \mu_2 \quad \rightarrow$ Hipótesis Alternativa

- Regla:
  - Si $|Z_p| \leq Z_{c(1{,}96)} \quad \rightarrow$ Acepto H_0
  - Si $|Z_p| > Z_{c(1{,}96)} \quad \rightarrow$ Rechazo H_0

$$\text{-}\ Z_p = \frac{\Delta\overline{X} - \overbrace{(\mu_a - \mu_b)}^{0}}{\sigma_{\Delta\overline{X}}} \quad \Rightarrow \quad Z_p = \frac{3 - 0}{1{,}523} \quad \Rightarrow \quad Z_p \simeq 1{,}97$$

d) **Cuál es, en palabras, la conclusión de la prueba?**

- Como $|Z_p| > Z_{c(1{,}96)}$, es decir, $1{,}97 > 1{,}96$, se rechaza H_0.
- Se puede considerar que los promedios poblacionales μ_1 y μ_2 no son iguales.

- Utilizando el intervalo de confianza descrito en el inciso a) se obtiene la misma conclusión:

$$- P\left[\Delta\overline{X} - Z_{(1-\alpha/2)}\,.\,\sigma_{\Delta\overline{X}} \leq \Delta\mu \leq \Delta\overline{X} + Z_{(1-\alpha/2)}\,.\,\sigma_{\Delta\overline{X}}\right] = 1 - \alpha = 0{,}95$$

$$- L_i = \Delta\overline{X} - Z_{(1-\alpha/2)}\,.\,\sigma_{\Delta\overline{X}} = (3) - (1{,}96)(1{,}523) \simeq 0{,}015$$

$$- L_s = \Delta\overline{X} + Z_{(1-\alpha/2)}\,.\,\sigma_{\Delta\overline{X}} = (3) + (1{,}96)(1{,}523) \simeq 5{,}985$$

$$- P\left[0{,}015 \leq \Delta\mu \leq 5{,}985\right] = 1 - \alpha = 0{,}95$$

- El intervalo [0,015; 5,985] no incluye al cero, por lo que se puede considerar que el tiempo promedio de los dispositivos Tipo I es mayor que el tiempo promedio de los dispositivos Tipo II. De esta forma, $\mu_1 > \mu_2$.
- Si un intervalo incluye el cero, entonces se considera $\mu_1 = \mu_2$.
- Si los dos límites del intervalo [0,015; 5,985] fueran negativos se podría suponer que el primer promedio poblacional es menor que el segundo. De esta forma, $\mu_1 < \mu_2$.

e) **Calcule el valor** *p* **y explique qué indica.**

- Considerando las posibles muestras de tamaño n extraídas de una población infinita y distribuida binomialmente con p y q como las probabilidades respectivas de que un miembro dado exhiba o no exhiba una propiedad determinada; entonces se tiene una **distribución muestral de proporciones** cuya media μ_P y desviación típica σ_P serán:

$$- \mu_P = \mu = p \qquad\qquad - \sigma_P = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

- Para grandes valores de n ($n \geq 30$) la distribución muestral está muy próxima a una distribución normal. De esta forma:

$$\left[0{,}015 \leq \Delta p \leq 5{,}985\right]$$

- NOTA: Desconozco si esta respuesta es la correcta. Realmente lo dudo pero no entiendo por qué pide el valor de p cuando en el ejercicio no se desea conocer ningún tipo propiedad determinada, como podría ser, la proporción de dispositivos con una respuesta superior a la media, por ejemplo.

f) **Qué es la potencia de la prueba? Qué dato haría falta para calcularla?**

- La **potencia de una prueba estadística** o el **poder estadístico** es la probabilidad de que la hipótesis nula sea rechazada cuando la hipótesis alternativa es verdadera (es decir, la probabilidad de no cometer un Error de Tipo II). La potencia es en general una función de las distribuciones posibles, a menudo determinada por un parámetro, bajo la hipótesis alternativa.

- A medida que aumenta la potencia, las posibilidades de que se presente un Error de Tipo II disminuyen. La probabilidad de que ocurra un Error de Tipo II se conoce como la tasa de falsos negativos (β). Por lo tanto la potencia es igual a $1 - \beta$, que también se conoce como la **sensibilidad**.

- La potencia estadística puede depender de un gran número de factores. Algunos de estos factores pueden ser particulares a una situación de prueba específica, pero, como mínimo, el poder depende casi siempre de los siguientes dos factores:

- El criterio de significación estadística utilizado en la prueba (α).

- El tamaño de la muestra (n) usado para detectar el efecto

- El **criterio de relevancia** o **nivel de significancia** es una declaración de lo improbable que debe ser un resultado positivo para que la hipótesis nula sea rechazada. Una forma fácil de aumentar la potencia de una prueba es utilizando un criterio de significancia mayor, por ejemplo 0.10 en lugar de 0.05. Esto aumenta la probabilidad de rechazar la hipótesis nula (es decir, la obtención de un resultado estadísticamente significativo) cuando la hipótesis nula es falsa, es decir, reduce el riesgo de un Error de Tipo II (falso negativo con respecto a si existe un efecto). Pero también aumenta el riesgo de obtener un resultado estadísticamente significativo (es decir, rechazar la hipótesis nula) cuando la hipótesis nula no es falsa, es decir, que aumenta el riesgo de un Error de Tipo I (falso positivo).

- El **tamaño de la muestra** n determina la cantidad de error de muestreo inherente a un resultado de la prueba. En igualdad de condiciones, los efectos son más difíciles de detectar en muestras más pequeñas. El aumento de tamaño de la muestra es a menudo la manera más fácil de aumentar la potencia estadística de una prueba.

- La precisión con la que se miden los datos también influye en la potencia estadística. En consecuencia, la potencia a menudo se puede mejorar mediante la reducción del error de medición en los datos. Un concepto relacionado es el de mejorar la "fiabilidad" de la medida que se está evaluando.

- Aunque no existen normas formales para la potencia(a veces referido como π), la mayoría de los investigadores evalúan el poder de sus pruebas con $\pi = 0{,}8$ como un estándar para la adecuación.

- El análisis de la potencia es apropiado cuando la preocupación es sobre el rechazo correcto, o no, de una hipótesis nula.

- Potencia $= \pi = P(\textbf{rechazar } H_0 \,|\, H_0 \textbf{ es falsa}) = 1 - \beta = 0{,}8$

- Para obtener la Potencia, se debe conocer β la cual sigue la siguiente expresión:

$$- \beta = P\left(Z_p \le \frac{\overline{X}_c - \mu}{\sigma/\sqrt{n}} \right)$$

- Como se puede ver, se poseen todos los datos necesarios menos el parámetro μ, el cual solo se tiene un intervalo con los valores posibles de él $\left[0{,}015 \le \Delta\mu \le 5{,}985\right]$.

g) **Si las muestras fueran chicas. Qué distribución debería usarse para el test?**

- Si las muestras fuesen chicas ($n < 30$), debería utilizarse una distribución t - Student.

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}} = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

TEORÍA:

Ejercicio 1:

a) **Desarrolle el concepto de Variable aleatoria, explicando además qué es el recorrido.**

- Si a cada punto de un espacio muestral se le asignase un número, se podría definir una función en el espacio muestral. Dicha función se la llama **variable aleatoria** (o variable estocástica). También es llamada función aleatoria o función estocástica y comúnmente se la denota por una letra mayúscula como X o Y. Sin embargo, cuando se habla del valor de dichas variables aleatorias, se suelen utilizar letras minúsculas, tales como x o y.

- Se llama **rango** de una variable aleatoria X y se denota R_x, a la imagen o rango de la función, es decir, al conjunto de los valores reales que ésta puede tomar. Dicho de otro modo, el rango de un Variable Aleatoria es el recorrido de la función por la que ésta queda definida.

- Por ejemplo, si se lanzaran dos monedas al aire, el espacio muestral, es decir, el conjunto de resultados elementales posibles asociado al experimento, es:

- $S = \{cc, cs, sc, ss\}$
- Siendo c el suceso "Salió cara"
- Siendo s el suceso "Salió seca"

- Sea X la Variable Aleatoria que indica el número de caras obtenidas al tirar las 2 monedas, el recorrido de X será:

- $R_x = \{0,1,2\}$

b) **Explique la diferencia entre variable aleatoria continua y discreta.**

- Una variable aleatoria que toma un número finito o infinito contable de valores se denomina **variable aleatoria discreta** mientras que una variable que toma un número infinito no contable de valores se llama **variable aleatoria continua** o "no discreta".

c) **Qué condiciones tiene que cumplir una función de densidad de probabilidad para una variable aleatoria continua?**

- Se llama **función de densidad de probabilidad** o simplemente **función de densidad** a toda función $f(x)$ que satisfaga las siguientes propiedades:

$$1)\ f(x) \geq 0 \qquad\qquad 2)\ \int_{-\infty}^{\infty} f(x)dx = 1$$

- La segunda es una proposición matemática del hecho que una variable aleatoria de valor real debe ciertamente encontrarse entre $-\infty$ y ∞. De esta forma, la probabilidad de que X se encuentre entre dos valores a y b será:

$$P(a < X < b) = \int_{a}^{b} f(x)dx$$

- En el caso de que $f(x)$ sea continua, se supone que (a menos que se establezca otra cosa) la probabilidad de que X sea igual a cualquier valor determinado es cero. De esta forma, el signo $<$ puede intercambiarse por $\leq$ sin ningún problema:

$$P(a < X < b) = P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b)$$

d) **Explique por qué un "estimador" o "estadístico" es una variable aleatoria.**

- Un estimador es un tipo de estadístico (una función de la muestra) utilizado para estimar un parámetro desconocido de una población Por lo tanto, es una variable aleatoria cuyos valores variarán de acuerdo a la muestra concreta escogida.

Ejercicio 2:

- Cuándo dos sucesos son independientes? Cuándo son excluyentes?

- En teoría de probabilidades, se dice que dos sucesos aleatorios son **independientes** entre sí cuando la probabilidad de cada uno de ellos no está influida porque el otro suceso ocurra o no, es decir, cuando ambos sucesos no están relacionados. Un ejemplo simple sería el caso de arrojar dos monedas al aire: el suceso “salió cara” en una moneda es independiente del de la otra moneda.

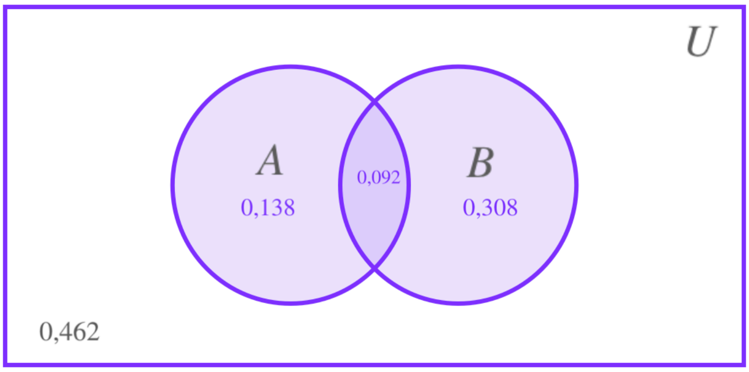
- Dos sucesos son independientes si la probabilidad de que ocurran ambos simultáneamente es igual al producto de las probabilidades de que ocurra cada uno de ellos, es decir, si $P(A \cap B) = P(A)P(B)$.

- Por otro lado, dos sucesos son **mutuamente excluyentes** si ambos no pueden ser verdaderos (o suceder simultáneamente). Un ejemplo de ello es el resultado de tirar una vez una moneda, el cual solo puede ser "cara" o "seca", pero no ambos.

- Si A y B son eventos mutuamente excluyentes, entonces la probabilidad de que A o B suceda es equivalente a la probabilidad del evento A más la probabilidad del evento B, es decir, $P(A \cup B) = P(A) + P(B)$

$a)$ **Sea** $P(A \cap B) = 0{,}092$ **y** $P(A) = 0{,}23$

- Cuánto valdría $P(B)$ si los sucesos fuesen independientes?



- Si A y B son independientes:

- $P(A \cap B) = P(A)P(B) = 0{,}092 \quad \Rightarrow \quad (0{,}23)P(B) = 0{,}092$

$\Rightarrow \quad P(B) = 0{,}4$

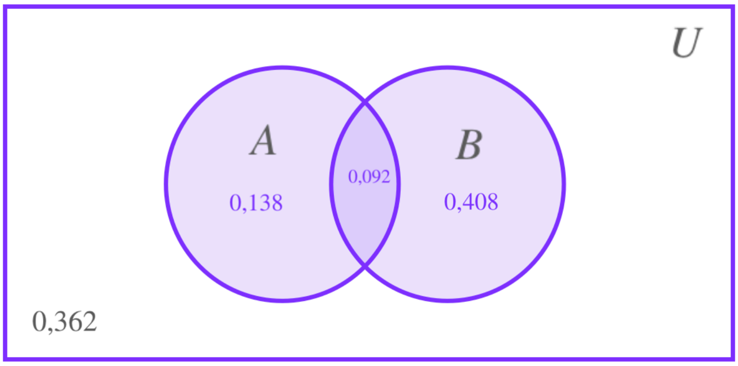
- Cuánto valdría $P(A \cup B)$?

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- $P(A \cup B) = (0{,}23) + (0{,}4) - (0{,}092)$

- $P(A \cup B) = 0{,}638$

b) **Si** $P(B) = 0{,}5$



- Cuánto vale $P(A \cup B)$?

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- $P(A \cup B) = (0{,}23) + (0{,}5) - (0{,}092)$

- $P(A \cup B) = 0{,}538$

- Cuánto vale $P(B|A)$?

- Si los sucesos A y B son dependientes:

- $P(B|A) = \dfrac{P(B \cap A)}{P(A)}$

- $P(B|A) = \dfrac{(0{,}092)}{(0{,}23)}$

- $P(B|A) = 0{,}4$

- Si los sucesos A y B son independientes:

$$- P(B|A) = P(B) = 0{,}5$$

- Ilustre ambos incisos con diagramas de Venn.

Ejercicio 3:

- Enuncie el Teorema de Límite Central y aplíquelo a la media muestral.

- El teorema del límite central o teorema central del límite indica que, en condiciones muy generales, si S_n es la suma de n variables aleatorias independientes y de varianza no nula pero finita, entonces la función de distribución de S_n "se aproxima bien" a una distribución normal (también llamada *distribución gaussiana*, *curva de Gauss* o *campana de Gauss*). Así pues, el teorema asegura que esto ocurre cuando la suma de estas variables aleatorias e independientes es lo suficientemente grande ($n > 30$).

- Sea $X_1, X_2, \ldots, X_n$ un conjunto de variables aleatorias, independientes e idénticamente distribuidas con una distribución con media μ y varianza $\sigma^2 \neq 0$.

- Si n es suficientemente grande ($n > 30$), entonces la variable aleatoria

$$- \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- tiene aproximadamente una distribución normal con $\mu_{\overline{X}} = \mu$ y $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.

- Es muy común encontrar el teorema del Limite Central con la variable estandarizada Z_n en función de la media muestral $\overline{X}_n$:

$$- Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

- Ilustrado con un ejemplo:

- Las bolsas de sal envasadas por una máquina tienen $\mu = 500g$ y $\sigma = 35g$. Las bolsas se empaquetaron en cajas de 100 unidades.

- Calcular la probabilidad de que la media de los pesos de las bolsas de un paquete sea menor que 495g.

- Como $n > 30$, se puede afirmar que la población sigue una distribución normal.

$$P(\overline{X} < 495) = P\left(Z < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) = P\left(Z < \frac{495 - 500}{35/\sqrt{100}}\right) = P\,(Z < 1{,}43)$$

- Por medio de la tabla de Distribución Z, se puede saber que $P\,(Z < 1{,}43) = 0{,}9236 = 92{,}36\,\%$. Por lo tanto:

- $P(\overline{X} < 495) = 0{,}9236 = 92{,}36\,\%$

Ejercicio 4:

- Qué condiciones tiene que cumplir una variable aleatoria para que su distribución de probabilidades sea Binomial? Exprese el recorrido y la formula con la que se calculan las probabilidades de esta variable indicando que es cada término de la fórmula.

- En cada prueba de un experimento hay una probabilidad asociada con un suceso particular. En algunos casos, la probabilidad no cambia de una prueba a la siguiente. A estas pruebas se las llama independientes y se las conoce como *pruebas de Bernoulli*.

- Sea p la probabilidad de que un suceso ocurra en una sola prueba de Bernoulli (llamada probabilidad de éxito). Entonces $q = 1 - p$ es la probabilidad de que el suceso no ocurra en una sola prueba (llamada probabilidad de fracaso). La probabilidad de que el suceso ocurra x veces en n pruebas, es decir, que ocurran x éxitos y $n - x$ fracasos, esta dada por la función de probabilidad:

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

- Donde la variable X denota el número de éxitos en n pruebas y $x = 0,1,\ldots,n$.

Ejercicio 5:

a) **Qué es un estimador puntual?**

- La estimación puntual consiste en atribuir un valor (el estimador puntual) al parámetro poblacional. Si la muestra es representativa de la población, podemos esperar que los estadísticos calculados en las muestras tengan valores semejantes a los parámetros poblacionales.

- De esta forma, la estimación consiste en asignar los valores de los estadísticos muestrales a los parámetros poblacionales. Los estadísticos con que obtenemos las estimaciones se denominan estimadores.

b) **Qué es un intervalo de confianza?**

- Se llama intervalo de confianza a un par o varios pares de números entre los cuales se estima que estará cierto valor desconocido con una determinada probabilidad de acierto. Formalmente, estos números determinan un intervalo, que se calcula a partir de datos de una muestra, y el valor desconocido es un parámetro poblacional.

- La probabilidad de éxito en la estimación se representa con $1 - \alpha$ y se denomina *nivel de confianza*. En estas circunstancias, α es el llamado *error aleatorio* o *nivel de significación*, es decir, una medida de las posibilidades de fallar en la estimación mediante tal intervalo.

- El nivel de confianza y la amplitud del intervalo varían conjuntamente, de forma que un intervalo más amplio tendrá más probabilidad de acierto (mayor nivel de confianza), mientras que para un intervalo más pequeño, que ofrece una estimación más precisa, aumenta su probabilidad de error.