Estimación por intervalos

1.1 Estimación

En general, inferencia estadística consiste en hacer **conclusiones basado en la teoría de probabilidad sobre cantidades desconocidas (parametros).**

El objetivo de un investigador que usa estadística es aprender y obtener resultados sobre los parámetros de una población **después de observar los datos (muestra)**. Desde la perspectiva del investigador, los datos contienen información relevante para verificar algunas hipótesis en relación a algunos parametros de la población.

A fin de obtener resultados y conclusiones de uno o más parametros de la población se utiliza inferencia estadística.

Definición 1.1 (Inferencia Estadística)
Una inferencia estadística, es un procedimiento que proporciona conclusiones probabilísticas en relación a uno o más parámetros de una población objetivo.

El término **conclusiones probabilísticas** significa que para realizar conclusiones en relación a uno o mas parámetros de la población se utiliza la teoría de probabilidad.

A fin de obtener conclusiones probabilísticas es necesário utilizar un estimador para un parámetro θ

Definición 1.2 (Estimador)
Sea $X_1, X_2, \ldots, X_n$ una m.a, un estimador del parámetro θ es una función basada en la m.a

$$\widehat{\theta}(X_1, X_2, \ldots, X_n)$$

Si la muestra aleatoria es observada,

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n,$$

entonces

$$\hat{\theta}(x_1, x_2, \ldots, x_n)$$

es denominado estimación (estimativa) de θ.

Los estimadores se usan en dos formas diferentes:

1. **Estimación puntual** Basado en los datos, se calcula el valor observado del estimador $\hat{\theta}$ del parámetro θ, así se asocia

   $$\theta = \hat{\theta}.$$

   El valor observado $\hat{\theta}$ se le denomina estimación puntual de θ.

   Problema: Un estimador puntual no proporciona la precisión de la estimación.

2. **Estimación por intervalo** Tiene por finalidad estimar un conjunto de posibles valores de θ consistentes con los datos.

Ejemplo 1.1
Para los siguientes parámetros se presenta los estimadores respectivos

	parámetro	estimador	estimación
media	μ	$\overline{X}$	$\bar{x}$
varianza	σ^2	$\widehat{S}^2$	$\widehat{s}^2$
desviación estándar	σ	S	s
proporción	p	$\widehat{P}$	$\widehat{p}$
diferencia de medias	$\mu_1 - \mu_2$	$\overline{X}_1 - \overline{X}_2$	$\bar{x}_1 - \bar{x}_2$
diferencia de proporciones	$p_1 - p_2$	$\widehat{P}_2 - \widehat{P}_2$	$\widehat{p}_1 - \widehat{p}_2$
coefiente de variación	σ/μ	$S/\overline{X}$	$s/\bar{x}$

Ejemplo 1.2
Muchos de los medicamentos empleados en el tratamiento del cáncer son costosos. BusinessWeek informó de los costos de los tratamientos con Herceptin, un medicamento para tratar el cáncer de mama (BusinessWeek, 30 de enero de 2006). Los siguientes son los costos de tratamientos con Herceptin en una muestra aleatoria de 10 pacientes.

4376	5578	2717	4920	4495
4798	6446	4119	4237	3814

- Calcule una estimación puntual del costo medio y la desviacĩon estándar de un tratamiento con Herceptin
- Calcule una estimación puntual del coeficiente de variación en los costos de los tratamientos con Herceptin.

Ejemplo 1.3
Una muestra de 20 estudiantes que recientemente tomaron un curso de estadística elemental arrojó la siguiente información sobre la marca de calculadora que poseían. (T=Texas Instruments, H=Hewlett Packard, C=Casio, S=Sharp):

T	T	H	T	C	T	T	S	C	H
S	S	T	H	C	T	T	T	H	T

Estime la proporción verdadera de los estudiantes que poseen una calculadora Texas Instruments, Hewlett Packard y Casio.

El estimador $\widehat{\theta}(X_1, X_2, \ldots, X_n)$ es una variable aleatoria y su distribución de probabilidad es denominada **distribución muestral**.

La varianza de la distribución muestral del estimador es dado por

$$\mathrm{Var}\left[\widehat{\theta}\right] = \mathrm{E}\left[\widehat{\theta}^2\right] - \left(\mathrm{E}\left[\widehat{\theta}\right]\right)^2.$$

La raíz cuadrada de la varianza del estimador es conocida como **error estándar** del estimador

$$\mathrm{SE}\left(\widehat{\theta}\right) = \sqrt{\mathrm{Var}\left[\widehat{\theta}\right]}$$

Definición 1.3 (Error de Estimación)
El error de estimación es definido como la diferencia entre el parámetro y la estimación del parámetro

$$\textbf{error de estimación} = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}$$

1. Son pocos los casos en que el error de estimación puede ser calculado exactamente.
2. Si el error de estimación es aproximadamente cero, entonces la estimación del parámetro es buena, caso contrario existe mucho error

1.2 Estimación por Intervalos

Las estimación puntual de un parámetro no siempre es igual al valor verdadero del parámetro. Por tanto, se prefiere determinar un intervalo dentro del cual se espera encontrar el valor del parámetro. Así, se suele calcular una estimación por intervalos **al sumar y restar al estimador puntual una cantidad llamada margen de error**

Estimación puntual $\pm$ **Margen de error**
Estimación puntual $\pm$ **Cuantil** $\times$ **error estándar**

La finalidad de la estimación por intervalos es proporcionar información de que tan próximo esta el estimador puntual del valor del parámetro poblacional.

Por ejemplo, la estimación del intervalo de confianza para μ es

$$\bar{x} \pm \textbf{Margen de error}$$

Pregunta: Como obtener el margen de error?

Definición 1.4 (Estimación por Intervalos)
Dada una m.a $X_1, X_2, \ldots, X_n$ tomada de una distribución (población) con parametro θ. Sean

$$L = L(X_1, X_2, \ldots, X_n) \quad \text{y} \quad U = U(X_1, X_2, \ldots, X_n)$$

dos estadísticas que tienen la propiedad que para todo valor de θ

$$\mathbf{P}(L < \theta < U) = 1 - \alpha,$$

en que $(1-\alpha)$ es denominada de coeficiente de confianza.
Entonces el intervalo aleatorio (L, U) **es denominado intervalo de confianza de** $(1-\alpha)100\,\%$ para θ.

El coeficiente de confianza $100(1-\alpha)\,\%$ es también denominado **nivel de confianza**.

Con los valores observados de la muestra aleatoria $x_1, x_2, \ldots, x_n$ se obtiene los valores numéricos de

$$L(x_1, x_2, \ldots, x_n) = c_1 \quad \text{y} \quad U(x_1, x_2, \ldots, x_n) = c_2$$

y así el intervalo (c_1, c_2) es denominado valor observado del intervalo de confianza.

Observación:

- Las estadísticas L y U son construidas a partir de la distribución muestral del estimador del parámetro θ
- Los valores c_1 y c_2 son usados para obtener el margen de error de la estimación puntual de los parámetros.

1.2.1 Interpretación

1. Antes de:
   - realizar el experimento y de colectar los datos
   - realizar la encuesta
   - de hacer un estudio de mercado

   osea, antes de observar los datos se puede afirmar que el intervalo aleatorio (L, U)

   $$\mathbf{P}(L < \theta < U) = 1 - \alpha$$

   contiene (incluye) al verdadero valor del parámetro con probabilidad $100(1-\alpha)\,\%$ **(nivel de confianza)**.

2. Despues de observar los datos, y se calculan los valores para

   $$L = c_1 \quad \text{y} \quad U = c_2$$

   entonces es **incorrecto** interpretar que

   $$\mathbf{P}(c_2 < \theta < c_2) = 1 - \alpha$$

   es la probabilidad que el parámetro θ esté en el intervalo (c_1, c_2) con probabilidad $1-\alpha$.

   La interpretación correcta esta basada en la idea de probabilidad a largo plazo (frecuencia relativa) esto es considerando repeticiones sucesivas.

Ejemplo 1.4
Sea una m.a $X_1, X_2, \ldots, X_n$ que proviene de una distribución normal con media μ y varianza σ^2. Un IC para μ con un nivel de confianza $1-\alpha = 0.95$

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Interpretación:

I) El intervalo $\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ contiene el verdadero valor de μ con probabilidad 0.95. Esto por que los extremos del intervalo son aleatorios.

II) Dado $n = 31$ y $\bar{x} = 80$, entonces se obtiene el intervalo de 95 % de confianza observado (fijo) del para μ dado por $(79.3, 80.7)$. Para la interpretación se debe de considerar lo siguiente

   a) **Interpretación incorrecta:** La probabilidad que μ esté en el intervalo $(79.3, 80.7)$ es 0.95.

   b) **Interpretación correcta:** Dado que se repite el experimento, por ejemplo, 100 veces. Entonces, se espera que el 95 % de esos intervalos contengan el verdadero valor de μ.

1.3 Intervalo de confianza para la media poblacional μ

1.3.1 Varianza σ^2 es conocida

Dado que el estimador puntual de μ es la media muestral $\overline{X}$, el intervalo de confianza (IC) para la media poblacional μ cuando la varianza, σ^2, **es conocida** es dado por

$$\begin{aligned} IC(\mu): \quad & \mu \in \bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \\ & \mu \in \left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

en que $z_{\frac{\alpha}{2}}$ representa el valor del cuantil de la distribución normal. Además el error estándar del estimador $\bar{x}$ es

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

El IC para μ dada anteriormente se aplica bajo las siguientes condiciones

a) **Población normal** y cualquier tamaño de muestra.

b) **Población no normal** y muestra grande ($n \geqslant 30$)

El margen de error en la estimación por intervalo para μ es dado por

$$\textbf{margen de error} \equiv z_{\frac{\alpha}{2}} \times SE(\bar{x})$$

El ancho del intervalo de confianza es dado por

$$\begin{aligned} \textbf{ancho del intervalo} & \equiv 2 \times \textbf{margen de error} \\ & \equiv 2 \times z_{\frac{\alpha}{2}} \times SE(\bar{x}) \end{aligned}$$

El ancho del intervalo da información sobre la precisión de una estimación de intervalo.

- Si el ancho del intervalo es angosto, se tiene mayor precisión sobre el valor del parámetro
- Si el ancho del intervalo es amplio, se tiene mayor incertidumbre (menos precisión) sobre el valor del parámetro.

Los cuantiles $z_{\frac{\alpha}{2}}$ satisfacen:

$$\mathbf{P}(Z < -z_{\frac{\alpha}{2}}) = \frac{\alpha}{2} \quad \text{y} \quad \mathbf{P}(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

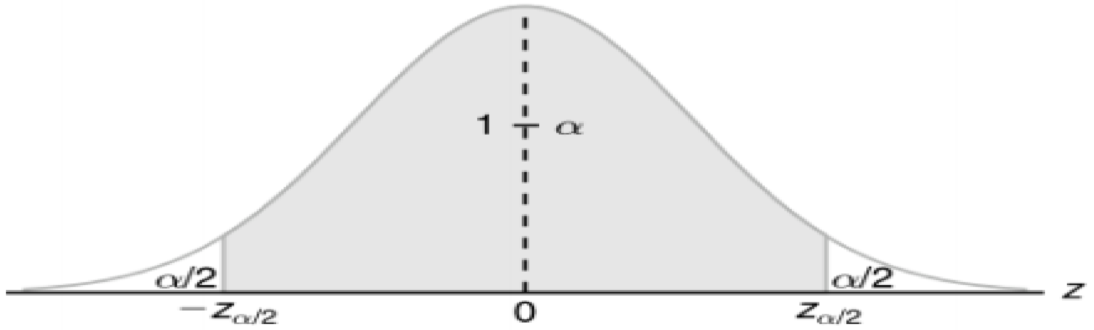

Figura 1.1: Cuantiles $-z_{\frac{\alpha}{2}}$ y $z_{\frac{\alpha}{2}}$

Ejemplo 1.5
Muchos pacientes con problemas del corazón tienen un marcapasos para controlar su ritmo cardíaco. El marcapasos tiene montado un módulo conector de plástico en la parte superior. Suponga una desviación estándar de 0.0015 pulgadas, y con base en esto calcule un intervalo de confianza del 95 % para la media de la profundidad de todos los módulos conectores fabricados por cierta empresa. Una muestra aleatoria de 75 módulos tiene una profundidad promedio de 0.310 pulgadas.

Solución:
La muestra es grande, $n = 75$, la **población es no normal** y además la varianza es conocida entonces si se puede aplicar el IC para la media con varianza con conocida.

$$\begin{aligned} \sigma &= 0.0015 \\ n &= 75 \\ \bar{x} &= 0.310 \\ 1-\alpha &= 0.95 \\ z_{\frac{\alpha}{2}} &= 1.96 \end{aligned}$$

$$\begin{aligned} IC(\mu):\quad & \mu \in \bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \\ & \mu \in 0.310 \pm 1.96 \times \frac{0.0015}{\sqrt{75}} \\ & \mu \in 0.310 \pm 0.0003395 \\ & \mu \in (0.310 - 0.0003395\ ;\ 0.310 + 0.0003395) \\ & \mu \in (0.3096605\ ;\ 0.3103395) \end{aligned}$$

Ejemplo 1.6
Un científico interesado en vigilar contaminantes químicos en alimentos y, por lo tanto, la acumulación de contaminantes en la dieta humana, seleccionó una muestra aleatoria de $n = 20$ adultos hombres. Se encontró que el promedio de ingesta diaria de productos lácteos fue de $\bar{x} = 756$ gramos por día. Se tiene información que la desviación estándar poblacional es $\sigma = 35$ gramos por día. Use esta información muestral para construir un intervalo de confianza de 90 % para la ingesta diaria media de productos lácteos para hombres. Suponga que la población es aproximadamente normal.

Solución:
La muestra es pequeña, $n = 20$, la **población es normal** y además la varianza es conocida entonces si se puede aplicar el IC para la media con varianza conocida.

$$\begin{aligned} \sigma &= 35 \\ n &= 20 \\ \bar{x} &= 756 \\ 1-\alpha &= 0.90 \\ z_{\frac{\alpha}{2}} &= 1.645 \end{aligned}$$

$$\begin{aligned} IC(\mu): \quad & \mu \in \bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \\ & \mu \in 756 \pm 1.645 \times \frac{35}{\sqrt{20}} \\ & \mu \in 756 \pm 12.8742 \\ & \mu \in (756 - 12.8742 \ ; \ 756 + 12.8742) \\ & \mu \in (743.1258 \ ; \ 768.8742) \end{aligned}$$

1.3.2 Varianza σ^2 desconocida

Se tiene 2 casos cuando la muestra es grande o pequeña. En esta caso, el error estándar $SE(\bar{x})$ no puede ser obtenido porque la varianza es desconocida. En este sentido, se utiliza el error estándar estimado (SEE) dado por

$$SEE(\bar{x}) = \frac{s}{\sqrt{n}}$$

Muestra Grande

Cuando la **muestra es grande**, $n \geqslant 30$. Se considera la aproximación

$$\sigma^2 \approx s^2$$

en que s^2 es la varianza muestral. Luego, el IC para μ es dado por

$$\begin{aligned} IC(\mu): \quad & \mu \in \bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \\ & \mu \in \left(\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \quad ; \quad \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right) \end{aligned}$$

y este intervalo se calcula tanto si la población es: **normal** y **no normal**.

Muestra Pequeña

Cuando la **muestra es pequeña** ($n < 30$). En este caso, la muestra aleatoria debe seguir una (aproximadamente) distribución normal.
Luego, se considera la variable aleatoria,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_v,$$

en que $v = n - 1$ son los grados de libertad de la distribución t-student.
El IC para μ cuando σ^2 es desconocida para muestras pequeñas es dado por

$$\begin{aligned} IC(\mu): \quad & \mu \in \bar{x} \pm t_{v,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \\ & \mu \in \left(\bar{x} - t_{v,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \ ; \ \bar{x} + t_{v,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right) \end{aligned}$$

Nótese que $t_{v,\frac{\alpha}{2}}$ representa el cuantil de la distribución t-student con $v = n - 1$ grados de libertad.
El cuantil $t_{v,\frac{\alpha}{2}}$ satisface:

$$\mathbf{P}(T < -t_{v,\frac{\alpha}{2}}) = \frac{\alpha}{2} \quad \text{y} \quad \mathbf{P}(T > t_{v,\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

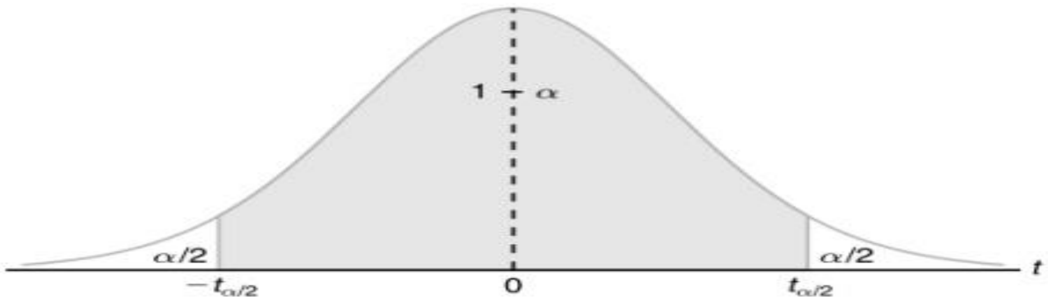

Figura 1.2: Cuantiles $-t_{\nu,\frac{\alpha}{2}}$ y $t_{\nu,\frac{\alpha}{2}}$

Ejemplo 1.7

Un proveedor vende fibras sintéticas a una compañía de manufactura. Se selecciona una muestra aleatoria simple de 81 fibras de un envío. El promedio de la fuerza de ruptura de éstas es de 29 lb y la desviación estándar de 9 lb.

a) Determine un intervalo de confianza de 95 % y 99 % para la media de la fuerza de ruptura de todas las fibras del envío.

b) ¿Cuál es el nivel de confianza del intervalo (27.5, 30.5)?

Solución:

La muestra es grande, $n = 81$ y además la varianza es desconocida entonces si se puede aplicar el IC para la media con varianza desconocida.

$$n = 81 \quad \bar{x} = 29 \quad s = 9$$

Para obtener el IC se utiliza

$$IC(\mu): \quad \mu \in \bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

a) Se tiene dos casos

- Para $1 - \alpha = 0.95 \Rightarrow z_{\frac{\alpha}{2}}$=1.96

$$\begin{aligned}
&\mu \in 29 \pm 1.96 \times \frac{9}{\sqrt{81}} \\
&\mu \in 29 \pm 1.96 \\
&\mu \in (29 - 1.96 \ ; \ 29 + 1.96) \\
&\mu \in (27.04 \ ; \ 30.96)
\end{aligned}$$

- Para $1-\alpha = 0.99 \Rightarrow z_{\frac{\alpha}{2}}$=2.576

$$\begin{aligned}
\mu &\in 29 \pm 2.576 \times \frac{9}{\sqrt{81}} \\
\mu &\in 29 \pm 2.576 \\
\mu &\in (29 - 2.576\ ;\ 29 + 2.576) \\
\mu &\in (26.424\ ;\ 31.576)
\end{aligned}$$

b) ¿Cuál es el nivel de confianza del intervalo $(27.5, 30.5)$?

$$\begin{aligned}
29 - z_{\frac{\alpha}{2}} \times \frac{9}{\sqrt{81}} &= 27.5 \\
z_{\frac{\alpha}{2}} &= 1.5
\end{aligned}$$

Luego se sabe que

$$\begin{aligned}
\mathbf{P}\left(Z \leqslant -z_{\frac{\alpha}{2}}\right) &= \frac{\alpha}{2} \\
\mathbf{P}\left(Z \leqslant -1.5\right) &= \frac{\alpha}{2} \quad \text{buscando en la tabla Z} \\
\mathbf{P}\left(Z \leqslant -1.5\right) &= \frac{\alpha}{2} = 0.0668
\end{aligned}$$

Finalmente $\alpha = 0.1336 \Rightarrow 1-\alpha = 0.8664$

Ejemplo 1.8
Se registran las siguientes mediciones del tiempo de secado, en horas, de cierta marca de pintura vinílica:

2.8	3.3	5.6	3.7	2.8
4.4	4	5.2	3	4.8
3.4	2.5	4.8	2.9	3.6

Suponga que las mediciones representan una muestra aleatoria de una población normal y con base en esto calcule el intervalo de confianza del 95 % para el tiempo medio de secado de la pintura.

Solución:
La muestra es pequeña, $n = 15$, la **población es normal** y además la varianza es desconocida entonces se aplica el IC para la media con varianza desconocida para muestras pequeñas

$$\begin{aligned} s &= 0.9709 \\ n &= 15 \\ \bar{x} &= 3.7867 \\ 1-\alpha &= 0.95 \\ \nu = n-1 = 15-1 &= 14 \text{ (gl)} \\ t_{\frac{\alpha}{2}} &= 2.145 \end{aligned}$$

$$\begin{aligned} IC(\mu): \quad & \mu \in \bar{x} \pm t_{\nu,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \\ & \mu \in 3.7867 \pm 2.145 \times \frac{0.9709}{\sqrt{15}} \\ & \mu \in 3.7867 \pm 0.5377 \\ & \mu \in (3.7867 - 0.5377 \ ; \ 3.7867 + 0.5377) \\ & \mu \in (3.249 \ ; \ 4.3244) \end{aligned}$$

Ejemplo 1.9
En un estudio de National Retail Foundation se encontró que las familias estaban dispuestas a gastar en promedio \$649 durante las vacaciones decembrinas (*The Wall Street Journal*, 2 de diciembre de 2002). Suponga que en el estudio participaron 600 familias y que la desviación estándar muestral fue \$175.

a) ¿Con 90 % de confianza cuál es el margen de error?

b) ¿Cuál es el intervalo de confianza de 90 % para estimar la media poblacional?

Ejemplo 1.10
Merrill Lynch Securities y Health Care Retirement, Inc., son dos grandes empresas ubicadas en el centro de Toledo, Ohio. Contemplan ofrecer de forma conjunta servicio de guardería para sus empleados. Como parte del estudio de viabilidad del proyecto, desean calcular el costo medio semanal por el cuidado de niños de los empleados. Una muestra de 10 empleados que recurren al servicio de guardería revela las siguientes cantidades gastadas (en dolares) la semana pasada.

107	92	97	95	105	101	91	99	95	104

Asumiendo que la la muestra es tomada de una población normal. Construya un intervalo de confianza de 90 % para la media poblacional.

PRUEBA DE HIPOTESIS

Definición 2.1 (Hipótesis Estadística)
Una hipótesis estadística o simplemente hipótesis es una pretensión, conjetura o aseveración sobre el valor de uno o mas parámetros (característica de una población).
Las hipótesis pueden envolver parámetros de diferentes poblaciones

A fin de conocer la verdad o falsedad de una hipótesis estadística será evaluada por tomar una muestra aleatoria de la poblacion en estudio. Los datos colectados, sirven para proporcionar evidencia para aceptar o rechazar la hipótesis.

Ejemplo 2.1
Algunos ejemplos de hipótesis estadísticas

- el tiempo de secado de un cierto tipo de pintura en condiciones de prueba especificadas es 75 min, μ=75 min
- la proporción de personas que tienen un seguro de salud en Arequipa es mayor que 40 %, $p > 40\,\%$
- La duración media del neumático de la marca A es mayor que la del neumático B, $\mu_A > \mu_B$
- La variabilidad del línea de ensamble A es igual que la línea B, $\sigma_A^2 = \sigma_B^2$

En prueba de hipótesis, se tomará dos decisiones: aceptar o rechazar la hipótesis nula, H_0. Pero esta decisión será tomada llevando en cuenta lo siguiente:

- rechazar H_0 a favor de H_1 debido a evidencia suficiente encontrada en los datos ó
- no rechazar H_0 debido a evidencia insuficiente en los datos

Ejemplo 2.2
De las hipótesis, $H_0 : \mu=0.75$ contra la alternativa $H_1 : \mu \neq 0.75$. Se rechazará H_0 sólo si los datos muestrales sugieren fuertemente que μ es diferente de 0.75. Si los datos no muestran evidencia para ello, H_0 no deberá ser rechazada, puesto que sigue siendo bastante plausible.

En el presente curso serán considerados las siguientes hipótesis nula con sus respectivas hipótesis alternativas

hipótesis nula	$H_0 : \theta = \theta_0$	$H_0 : \theta \geqslant \theta_0$	$H_0 : \theta \leqslant \theta_0$
hipótesis alternativa	$H_1 : \theta \neq \theta_0$	$H_1 : \theta < \theta_0$	$H_1 : \theta > \theta_0$

Nótese que la H_1 no tiene posibilidad de ser una igualdad, *i.e.*, no se puede plantear H_1 como

$$H_1 : \theta = \theta_0 \quad \text{ó} \quad H_1 : \theta \leqslant \theta_0 \quad \text{ó} \quad H_1 : \theta \geqslant \theta_0$$