Notas del Curso de Estadística

Carlos E. Martínez Rodríguez

December 25, 2019

Contents

1 Pruebas de Hipótesis

1.1 Tipos de errores

- Una hipótesis estadística es una afirmación acerca de la distribución de probabilidad de una variable aleatoria, a menudo involucran uno o más parámetros de la distribución.
- Las hipótesis son afirmaciones respecto a la población o distribución bajo estudio, no en torno a la muestra.
- La mayoría de las veces, la prueba de hipótesis consiste en determinar si la situaci ón experimental ha cambiado
- el interés principal es decidir sobre la veracidad o falsedad de una hipótesis, a este procedimiento se le llama *prueba de hipótesis*.
- Si la información es consistente con la hipótesis, se concluye que esta es verdadera, de lo contrario que con base en la información, es falsa.

Una prueba de hipótesis está formada por cinco partes

- La hipótesis nula, denotada por H_0.
- La hipótesis alterativa, denorada por H_1.
- El estadśitico de prueba y su valor p.
- La región de rechazo.
- La conclusión.

Definición 1.1 *Las dos hipótesis en competencias son la* ***hipótesis alternativa*** H_1*, usualmente la que se desea apoyar, y la* ***hipótesis nula*** H_0*, opuesta a* H_1*.*

En general, es más fácil presentar evidencia de que H_1 es cierta, que demostrar que H_0 es falsa, es por eso que por lo regular se comienza suponiendo que H_0 es cierta, luego se utilizan los datos de la muestra para decidir si existe evidencia a favor de H_1, más que a favor de H_0, así se tienen dos conclusiones

- Rechazar H_0 y concluir que H_1 es verdadera.
- Aceptar, no rechazar, H_0 como verdadera.

Ejemplo 1.1 *Se desea demostrar que el salario promedio por hora en cierto lugar es distinto de* $19usd$*, que es el promedio nacional. Entonces* $H_1 : \mu \neq 19$*, y* $H_0 : \mu = 19$*.*

A esta se le denomina **Prueba de hipótesis de dos colas**.

Ejemplo 1.2 *Un determinado proceso produce un promedio de* 5% *de piezas defectuosas. Se está interesado en demostrar que un simple ajuste en una máquina reducirá* p*, la proporción de piezas defectuosas producidas en este proceso. Entonces se tiene* $H_0 : p < 0.3$ *y* $H_1 : p = 0.03$*. Si se puede rechazar* H_0*, se concluye que el proceso ajustado produce menos del* 5% *de piezas defectuosas.*

A esta se le denomina **Prueba de hipótesis de una cola**.

La decisión de rechazar o aceptar la hipótesis nula está basada en la información contenida en una muestra proveniente de la población de interés. Esta información tiene estas formas

- **Estadśitico de prueba:** un sólo número calculado a partir de la muestra.
- p**-value:** probabilidad calculada a partir del estadśtico de prueba.

Definición 1.2 *El* p*-value es la probabilidad de observar un estadístico de prueba tanto o más alejado del valor obervado, si en realidad* H_0 *es verdadera. Valores grandes del estadśtica de prueba y valores pequeños de* p *significan que se ha observado un evento muy poco probable, si* H_0 *en realidad es verdadera.*

Todo el conjunto de valores que puede tomar el estadístico de prueba se divide en dos regiones. Un conjunto, formado de valores que apoyan la hipótesis alternativa y llevan a rechazar H_0, se denomina **región de rechazo**. El otro, conformado por los valores que sustentatn la hipótesis nula, se le denomina **región de aceptación**.

Cuando la región de rechazo está en la cola izquierda de la distribución, la prueba se denomina **prueba lateral izquierda**. Una prueba con región de rechazo en la cola derecha se le llama **prueba lateral derecha**.

Si el estadśtico de prueba cae en la región de rechazo, entonces se rechaza H_0. Si el estadśtico de prueba cae en la región de aceptación, entonces la hipótesis nula se acepta o la prueba se juzga como no concluyente.

Dependiendo del nivel de confianza que se desea agregar a las conclusiones de la prueba, y el **nivel de significancia** α, el riesgo que está dispuesto a correr si se toma una decisión incorrecta.

Definición 1.3 *Un* ***error de tipo I*** *para una prueba estadística es el error que se tiene al rechazar la hipótesis nula cuando es verdadera. El* ***nivel de significancia*** *para una prueba estadística de hipótesis es*

$$
\begin{aligned}
\alpha &= P\left\{\textit{error tipo I}\right\} = P\left\{\textit{rechazar equivocadamente } H_0\right\} \\
&= P\left\{\textit{rechazar } H_0 \textit{ cuando } H_0 \textit{ es verdadera}\right\}
\end{aligned}
$$

Este valor α representa el valor máximo de riesgo tolerable de rechazar incorrectamente H_0. Una vez establecido el nivel de significancia, la región de rechazo se define para poder determinar si se rechaza H_0 con un cierto nivel de confianza.

2 Muestras grandes: una media poblacional

2.1 Cálculo de valor p

Definición 2.1 *El* ***valor de*** p *(*p***-value****) o nivel de significancia observado de un estadístico de prueba es el valor más pequeño de* α *para el cual* H_0 *se puede rechazar. El riesgo de cometer un error tipo* I*, si* H_0 *es rechazada con base en la información que proporciona la muestra.*

Nota 2.1 *Valores pequeños de* p *indican que el valor observado del estadśtico de prueba se encuentra alejado del valor hipotético de* μ*, es decir se tiene evidencia de que* H_0 *es falsa y por tanto debe de rechazarse.*

Nota 2.2 *Valores grandes de* p *indican que el estadístico de prueba observado no está alejado de la medi hipotética y no apoya el rechazo de* H_0*.*

Definición 2.2 *Si el valor de* p *es menor o igual que el nivel de significancia* α*, determinado previamente, entonces* H_0 *es rechazada y se puede concluir que los resultados son estadísticamente significativos con un nivel de confianza del* $100\,(1-\alpha)\,\%$*.*

Es usual utilizar la siguiente clasificación de resultados

p	H_0	Significativa
$p < 0.01$	Rechazar	Altamente
$0.01 \leq p < 0.05$	Rechazar	Estadísticamente
$0.05 \leq p < 0.1$	No rechazar	Tendencia estadística
$0.01 \leq p$	No rechazar	No son estadísticamente

Nota 2.3 *Para determinar el valor de* p*, encontrar el área en la cola después del estadístico de prueba. Si la prueba es de una cola, este es el valor de* p*. Si es de dos colas, éste valor encontrado es la mitad del valor de* p*. Rechazar* H_0 *cuando el valor de* $p < \alpha$*.*

Hay dos tipos de errores al realizar una prueba de hipótesis

	H_0 es Verdadera	H_0 es Falsa
Rechazar H_0	Error tipo I	$\surd$
Aceptar H_0	$\surd$	Error tipo II

Definición 2.3 *La probabilidad de cometer el error tipo II se define por* β *donde*

$$\begin{aligned}\beta &= P\left\{\textit{error tipo II}\right\} = P\left\{\textit{Aceptar equivocadamente } H_0\right\} \\ &= P\left\{\textit{Aceptar } H_0 \textit{ cuando } H_0 \textit{ es falsa}\right\}\end{aligned}$$

Nota 2.4 *Cuando* H_0 *es falsa y* H_1 *es verdadera, no siempre es posible especificar un valor exacto de* μ*, sino más bien un rango de posibles valores.En lugar de arriesgarse a tomar una decisión incorrecta, es mejor conlcuir que no hay evidencia suficiente para rechazar* H_0*, es decir en lugar de aceptar* H_0*, no rechazar* H_0*.*

La bondad de una prueba estadística se mide por el tamaño de α y β, ambas deben de ser pequeñas. Una manera muy efectiva de medir la potencia de la prueba es calculando el complemento del error tipo II:

$$\begin{aligned}1-\beta &= P\{\text{Rechazar } H_0 \text{ cuando } H_0 \text{ es falsa}\} \\ &= P\{\text{Rechazar } H_0 \text{ cuando } H_1 \text{ es verdadera}\}\end{aligned}$$

Definición 2.4 *La* ***potencia de la prueba****,* $1-\beta$*, mide la capacidad de que la prueba funciona como se necesita.*

Ejemplo 2.1 *La producción diariade una planta química local ha promediado 880 toneladas en los últimos años. A la gerente de control de calidad le gustaría saber si este promedio ha cambiado en meses recientes. Ella selecciona al azar 50 días de la base de datos computarizada y calcula el promedio y la desviación estándar de las* $n = 50$ *producciones como* $\overline{x} = 871$ *toneladas y* $s = 21$ *toneladas, respectivamente. Pruebe la hipótesis apropiada usando* $\alpha = 0.05$*.*

Solución 2.1 *La hipótesis nula apropiada es:*

$$\begin{aligned}H_0 &: \mu = 880 \\ &\quad \textit{y la hipótesis alternativa } H_1 \textit{ es} \\ H_1 &: \mu \neq 880\end{aligned}$$

el estimador puntual para μ *es* $\overline{x}$*, entonces el estadístico de prueba es*

$$\begin{aligned}z &= \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \\ &= \frac{871-880}{21/\sqrt{50}} = -3.03\end{aligned}$$

Solución 2.2 *Para esta prueba de dos colas, hay que determinar los dos valores de* $z_{\alpha/2}$*, es decir,* $z_{\alpha/2} = \pm 1.96$

3 Pruebas de Hipótesis

3.1 Tipos de errores

- Una hipótesis estadística es una afirmación acerca de la distribución de probabilidad de una variable aleatoria, a menudo involucran uno o más parámetros de la distribución.
- Las hipótesis son afirmaciones respecto a la población o distribución bajo estudio, no en torno a la muestra.

- La mayoría de las veces, la prueba de hipótesis consiste en determinar si la situaci ón experimental ha cambiado
- el interés principal es decidir sobre la veracidad o falsedad de una hipótesis, a este procedimiento se le llama *prueba de hipótesis.*
- Si la información es consistente con la hipótesis, se concluye que esta es verdadera, de lo contrario que con base en la información, es falsa.
- La descisión de aceptar o rechazar la hipótesis nula se basa en un estadístico calculado a partir de la muestra. Esto necesariamente implica la existencia de un error.

4 Pruebas de Hipótesis

4.1 Tipos de errores

- Una hipótesis estadística es una afirmación acerca de la distribución de probabilidad de una variable aleatoria, a menudo involucran uno o más parámetros de la distribución.
- Las hipótesis son afirmaciones respecto a la población o distribución bajo estudio, no en torno a la muestra.
- La mayoría de las veces, la prueba de hipótesis consiste en determinar si la situaci ón experimental ha cambiado
- el interés principal es decidir sobre la veracidad o falsedad de una hipótesis, a este procedimiento se le llama *prueba de hipótesis.*
- Si la información es consistente con la hipótesis, se concluye que esta es verdadera, de lo contrario que con base en la información, es falsa.

Una prueba de hipótesis está formada por cinco partes

- La hipótesis nula, denotada por H_0.
- La hipótesis alterativa, denorada por H_1.
- El estadśitico de prueba y su valor p.
- La región de rechazo.
- La conclusión.

Definición 4.1 *Las dos hipótesis en competencias son la* ***hipótesis alternativa*** H_1*, usualmente la que se desea apoyar, y la* ***hipótesis nula*** H_0*, opuesta a* H_1*.*

En general, es más fácil presentar evidencia de que H_1 es cierta, que demostrar que H_0 es falsa, es por eso que por lo regular se comienza suponiendo que H_0 es cierta, luego se utilizan los datos de la muestra para decidir si existe evidencia a favor de H_1, más que a favor de H_0, así se tienen dos conclusiones

- Rechazar H_0 y concluir que H_1 es verdadera.
- Aceptar, no rechazar, H_0 como verdadera.

Ejemplo 4.1 *Se desea demostrar que el salario promedio por hora en cierto lugar es distinto de* $19usd$*, que es el promedio nacional. Entonces* $H_1 : \mu \neq 19$*, y* $H_0 : \mu = 19$.

A esta se le denomina **Prueba de hipótesis de dos colas**.

Ejemplo 4.2 *Un determinado proceso produce un promedio de* 5% *de piezas defectuosas. Se está interesado en demostrar que un simple ajuste en una máquina reducirá p, la proporción de piezas defectuosas producidas en este proceso. Entonces se tiene* $H_0 : p < 0.3$ *y* $H_1 : p = 0.03$*. Si se puede rechazar* H_0*, se concluye que el proceso ajustado produce menos del* 5% *de piezas defectuosas.*

A esta se le denomina **Prueba de hipótesis de una cola**.

La decisión de rechazar o aceptar la hipótesis nula está basada en la información contenida en una muestra proveniente de la población de interés. Esta información tiene estas formas

- **Estadśitico de prueba:** un sólo número calculado a partir de la muestra.
- p**-value:** probabilidad calculada a partir del estadśtico de prueba.

  Definición 4.2 *El p-value es la probabilidad de observar un estadístico de prueba tanto o más alejado del valor obervado, si en realidad* H_0 *es verdadera. Valores grandes del estadśtica de prueba y valores pequeños de p significan que se ha observado un evento muy poco probable, si* H_0 *en realidad es verdadera.*

Todo el conjunto de valores que puede tomar el estadístico de prueba se divide en dos regiones. Un conjunto, formado de valores que apoyan la hipótesis alternativa y llevan a rechazar H_0, se denomina **región de rechazo**. El otro, conformado por los valores que sustentatn la hipótesis nula, se le denomina **región de aceptación**.

Cuando la región de rechazo está en la cola izquierda de la distribución, la prueba se denomina **prueba lateral izquierda**. Una prueba con región de rechazo en la cola derecha se le llama **prueba lateral derecha**.

Si el estadśtico de prueba cae en la región de rechazo, entonces se rechaza H_0. Si el estadśtico de prueba cae en la región de aceptación, entonces la hipótesis nula se acepta o la prueba se juzga como no concluyente.

Dependiendo del nivel de confianza que se desea agregar a las conclusiones de la prueba, y el **nivel de significancia** α, el riesgo que está dispuesto a correr si se toma una decisión incorrecta.

Definición 4.3 *Un* ***error de tipo I*** *para una prueba estadística es el error que se tiene al rechazar la hipótesis nula cuando es verdadera. El* ***nivel de significancia*** *para una prueba estadística de hipótesis es*

$$\begin{aligned} \alpha &= P\left\{\textit{error tipo I}\right\} = P\left\{\textit{rechazar equivocadamente } H_0\right\} \\ &= P\left\{\textit{rechazar } H_0 \textit{ cuando } H_0 \textit{ es verdadera}\right\} \end{aligned}$$

Este valor α representa el valor máximo de riesgo tolerable de rechazar incorrectamente H_0. Una vez establecido el nivel de significancia, la región de rechazo se define para poder determinar si se rechaza H_0 con un cierto nivel de confianza.

5 Muestras grandes: una media poblacional

5.1 Cálculo de valor p

Definición 5.1 *El* ***valor de*** *p (p****-value****) o nivel de significancia observado de un estadístico de prueba es el valor más pequeño de α para el cual H_0 se puede rechazar. El riesgo de cometer un error tipo I, si H_0 es rechazada con base en la información que proporciona la muestra.*

Nota 5.1 *Valores pequeños de p indican que el valor observado del estadśtico de prueba se encuentra alejado del valor hipotético de μ, es decir se tiene evidencia de que H_0 es falsa y por tanto debe de rechazarse.*

Nota 5.2 *Valores grandes de p indican que el estadístico de prueba observado no está alejado de la medi hipotética y no apoya el rechazo de H_0.*

Definición 5.2 *Si el valor de p es menor o igual que el nivel de significancia α, determinado previamente, entonces H_0 es rechazada y se puede concluir que los resultados son estadísticamente significativos con un nivel de confianza del* $100\,(1-\alpha)\,\%$.

Es usual utilizar la siguiente clasificación de resultados

p	H_0	Significativa
$p < 0.01$	Rechazar	Altamente
$0.01 \leq p < 0.05$	Rechazar	Estadísticamente
$0.05 \leq p < 0.1$	No rechazar	Tendencia estadística
$0.01 \leq p$	No rechazar	No son estadísticamente

Nota 5.3 *Para determinar el valor de p, encontrar el área en la cola después del estadístico de prueba. Si la prueba es de una cola, este es el valor de p. Si es de dos colas, éste valor encontrado es la mitad del valor de p. Rechazar H_0 cuando el valor de $p < \alpha$.*

Hay dos tipos de errores al realizar una prueba de hipótesis

	H_0 es Verdadera	H_0 es Falsa
Rechazar H_0	Error tipo I	$\surd$
Aceptar H_0	$\surd$	Error tipo II

Definición 5.3 *La probabilidad de cometer el error tipo II se define por β donde*

$$\begin{aligned}\beta &= P\left\{\textit{error tipo II}\right\} = P\left\{\textit{Aceptar equivocadamente } H_0\right\} \\ &= P\left\{\textit{Aceptar } H_0 \textit{ cuando } H_0 \textit{ es falsa}\right\}\end{aligned}$$

Nota 5.4 *Cuando H_0 es falsa y H_1 es verdadera, no siempre es posible especificar un valor exacto de μ, sino más bien un rango de posibles valores.En lugar de arriesgarse a tomar una decisión incorrecta, es mejor conlcuir que no hay evidencia suficiente para rechazar H_0, es decir en lugar de aceptar H_0, no rechazar H_0.*

La bondad de una prueba estadística se mide por el tamaño de α y β, ambas deben de ser pequeñas. Una manera muy efectiva de medir la potencia de la prueba es calculando el complemento del error tipo II:

$$\begin{aligned} 1-\beta &= P\left\{\text{Rechazar } H_0 \text{ cuando } H_0 \text{ es falsa}\right\} \\ &= P\left\{\text{Rechazar } H_0 \text{ cuando } H_1 \text{ es verdadera}\right\} \end{aligned}$$

Definición 5.4 *La* ***potencia de la prueba****, $1-\beta$, mide la capacidad de que la prueba funciones como se necesita.*

6 Estimación por intervalos

Para la media

Recordemos que S^2 es un estimador insesgado de σ^2

Definición 6.1 *Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados de θ, parámetro poblacional. Si $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$, decimos que $\hat{\theta}_1$ un estimador más eficaz de θ que $\hat{\theta}_2$.*

Algunas observaciones que es preciso realizar

a) Para poblaciones normales, $\overline{X}$ y $\tilde{X}$ son estimadores insesgados de μ, pero con $\sigma^2_{\overline{X}} < \sigma^2_{\tilde{X}_2}$.

b) Para las estimaciones por intervalos de θ, un intervalo de la forma $\hat{\theta}_L < \theta < \hat{\theta}_U$, $\hat{\theta}_L$ y $\hat{\theta}_U$ dependen del valor de $\hat{\theta}$.

c) Para $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$, si $n \to \infty$, entonces $\hat{\theta} \to \mu$.

d) Para $\hat{\theta}$ se determinan $\hat{\theta}_L$ y $\hat{\theta}_U$ de modo tal que

$$P\left\{\hat{\theta}_L < \hat{\theta} < \hat{\theta}_U\right\} = 1-\alpha, \tag{1}$$

con $\alpha \in (0,1)$. Es decir, $\theta \in \left(\hat{\theta}_L, \hat{\theta}_U\right)$ es un intervalo de confianza del $100\,(1-\alpha)\,\%$.

e) De acuerdo con el TLC se espera que la distribución muestral de $\overline{X}$ se distribuye aproximadamente normal con media $\mu_X = \mu$ y desviación estándar $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Para $Z_{\alpha/2}$ se tiene $P\left\{-Z_{\alpha/2} < Z < Z_{\alpha/2}\right\} = 1-\alpha$, donde $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. Entonces $P\left\{-Z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right.$ $1-\alpha$ es equivalente a $P\left\{\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$

f) Si $\overline{X}$ es la media muestral de una muestra de tamaño n de una población con varianza conocida σ^2, el intervalo de confianza de $100\,(1-\alpha)\,\%$ para μ es $\mu \in \left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$.

g) Para muestras pequeñas de poblaciones no normales, no se puede esperar que el grado de confianza sea preciso.

h) Para $n \geq 30$, con distribución de forma no muy sesgada, se pueden tener buenos resultados.

Teorema 6.1 *Si $\overline{X}$ es un estimador de μ, podemos tener $100\,(1-\alpha)\,\%$ de confianza en que el error no excederá a $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, error entre $\overline{X}$ y μ.*

Teorema 6.2 *Si $\overline{X}$ es un estimador de μ, podemos tener $100\,(1-\alpha)\,\%$ de confianza en que el error no excederá una cantidad e cuando el tamaño de la muestra es*

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

Nota 6.1 *Para intervalos unilaterales*

$$P\left\{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_\alpha\right\} = 1-\alpha$$

equivalentemente

$$P\left\{\mu < \overline{X} + Z_\alpha\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha.$$

Si $\overline{X}$ es la media de una muestra aleatoria de tamaño n a partir de una población con varianza σ^2, los límites de confianza unilaterales del $100\,(1-\alpha)\,\%$ de confianza para μ están dados por

- Límite unilateral superior: $\overline{x} + z_\alpha\frac{\sigma}{\sqrt{n}}$
- Límite unilateral inferior: $\overline{x} - z_\alpha\frac{\sigma}{\sqrt{n}}$
- Para σ desconocida recordar que $T = \frac{\overline{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, donde s es la desviación estándar de la muestra. Entonces

$$P\left\{-t_{\alpha/2} < T < t_{\alpha/2}\right\} = 1-\alpha, \text{equivalentemente}$$
$$P\left\{\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right\} = 1-\alpha.$$

- Un intervalo de confianza del $100\,(1-\alpha)\,\%$ de confianza para μ, σ^2 desconocida y población normal es $\mu \in \left(\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$, donde $t_{\alpha/2}$ es una t-student con $\nu = n-1$ grados de libertad.
- Los límites unilaterales para μ con σ desconocida son $\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}$ y $\overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$.

- Cuando la población no es normal, σ desconocida y $n \geq 30$, σ se puede reemplazar por s para obtener el intervalo de confianza para muestras grandes:

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}.$$

- El estimador de $\overline{X}$ de μ, σ desconocida, la varianza de $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$, el error estándar de $\overline{X}$ es $\sigma/\sqrt{n}$.

- Si σ es desconocida y la población es normal, $s \to \sigma$ y se incluye el error estándar $s/\sqrt{n}$, entonces

$$\overline{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}.$$

Intervalos de confianza sobre la varianza

Supongamos que X se distribuye normal (μ, σ^2), desconocidas. Sea $X_1, X_2, \ldots, X_n$ muestra aleatoria de tamaño n , s^2 la varianza muestral.

Se sabe que $X^2 = \frac{(n-1)s^2}{\sigma^2}$ se distribuye χ^2_{n-1} grados de libertad. Su intervalo de confianza es

$$\begin{array}{l} P\left\{\chi^2_{1-\frac{\alpha}{2},n-1} \leq \chi^2 \leq \chi^2_{\frac{\alpha}{2},n-1}\right\} = 1-\alpha \\ P\left\{\chi^2_{1-\frac{\alpha}{2},n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n-1}\right\} = 1-\alpha \\ P\left\{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right\} = 1-\alpha \end{array} \tag{2}$$

es decir

$$\sigma^2 \in \left[\frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right] \tag{3}$$

los intervalos unilaterales son

$$\sigma^2 \in \left[\frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \infty\right] - \tag{4}$$

$$\sigma^2 \in \left[-\infty, \frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right] \tag{5}$$

Intervalos de confianza para proporciones

Supongamos que se tienen una muestra de tamaño n de una población grande pero finita, y supongamos que X, $X \leq n$, pertenecen a la clase de interés, entonces

$$\hat{p} = \frac{\overline{X}}{n}$$

es el estimador puntual de la proporción de la población que pertenece a dicha clase.

n y p son los parámetros de la distribución binomial, entonces $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ aproximadamente si p es distinto de 0 y 1; o si n es suficientemente grande. Entonces

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1), \text{aproximadamente.}$$

entonces

$$\begin{aligned} 1-\alpha &= P\left\{-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right\} \\ &= P\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right\} \end{aligned}$$

con $\sqrt{\frac{p(1-p)}{n}}$ error estándar del estimador puntual p. Una solución para determinar el intervalo de confianza del parámetro p (desconocido) es

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

entonces los intervalos de confianza, tanto unilaterales como de dos colas son:

- $p \in \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$
- $p \in \left(-\infty, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$
- $p \in \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \infty\right)$

para minimizar el error estándar, se propone que el tamaño de la muestra sea $n = \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p)$, donde $E = \mid p - \hat{p} \mid$.

7 Intervalos de confianza para dos muestras

Varianzas conocidas

Sean X_1 y X_2 variables aleatorias independientes. X_1 con media desconocida μ_1 y varianza conocida σ_1^2; y X_2 con media desconocida μ_2 y varianza conocida σ_2^2. Se busca encontrar un intervalo de confianza de $100(1-\alpha)\%$ de la diferencia entre medias μ_1 y μ_2.

Sean $X_{11}, X_{12}, \ldots, X_{1n_1}$ muestra aleatoria de n_1 observaciones de X_1, y sean $X_{21}, X_{22}, \ldots, X_{2n_2}$ muestra aleatoria de n_2 observaciones de X_2.

Sean $\overline{X}_1$ y $\overline{X}_2$, medias muestrales, entonces el estadśitico

$$Z=\frac{\left(\overline{X}_1-\overline{X}_2\right)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N\left(0,1\right), \tag{6}$$

si X_1 y X_2 son normales o aproximadamente normales si se aplican las condiciones del Teorema de Límite Central respectivamente.

Entonces se tiene

$$\begin{aligned}
1-\alpha &= P\left\{-Z_{\alpha/2}\leq Z\leq Z_{\alpha/2}\right\}\\
&= P\left\{-Z_{\alpha/2}\leq \frac{\left(\overline{X}_1-\overline{X}_2\right)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\leq Z_{\alpha/2}\right\}\\
&= P\left\{\left(\overline{X}_1-\overline{X}_2\right)-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\leq \mu_1-\mu_2\leq\right.\\
&\qquad\left.\left(\overline{X}_1-\overline{X}_2\right)+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right\}
\end{aligned}$$

Entonces los intervalos de confianza unilaterales y de dos colas al $(1-\alpha)\,\%$ de confianza son

- $\mu_1-\mu_2\in\left[\left(\overline{X}_1-\overline{X}_2\right)-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\left(\overline{X}_1-\overline{X}_2\right)+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$
- $\mu_1-\mu_2\in\left[-\infty,\left(\overline{X}_1-\overline{X}_2\right)+Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$
- $\mu_1-\mu_2\in\left[\left(\overline{X}_1-\overline{X}_2\right)-Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\infty\right]$

Nota 7.1 *Si σ_1 y σ_2 son conocidas, o por lo menos se conoce una aproximación, y los tamaños de las muestras n_1 y n_2 son iguales, $n_1=n_2=n$, se puede determinar el tamaño de la muestra para que el error al estimar $\mu_1-\mu_2$ usando $\overline{X}_1-\overline{X}_2$ sea menor que E (valor del error deseado) al $(1-\alpha)\,\%$ de confianza. El tamaño n de la muestra requerido para cada muestra es*

$$n=\left(\frac{Z_{\alpha/2}}{E}\right)^2\left(\sigma_1^2+\sigma_2^2\right).$$

Varianzas desconocidas

- Si $n_1,n_2\geq 30$ se pueden utilizar los intervalos de la distribución normal para varianza conocida
- Si n_1,n_2 son muestras pequeñas, supongase que las poblaciones para X_1 y X_2 son normales con varianzas desconocidas y con base en el intervalo de confianza para distribuciones t-student

$\sigma_1^2 = \sigma_2^2 = \sigma$

Supongamos que X_1 es una variable aleatoria con media μ_1 y varianza σ_1^2, X_2 es una variable aleatoria con media μ_2 y varianza σ_2^2. Todos los parámetros son desconocidos. Sin embargo supóngase que es razonable considerar que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Nuevamente sean X_1 y X_2 variables aleatorias independientes. X_1 con media desconocida μ_1 y varianza muestral S_1^2; y X_2 con media desconocida μ_2 y varianza muestral S_2^2. Dado que S_1^2 y S_2^2 son estimadores de σ_1^2, se propone el estimador S de σ^2 como

$$S_p^2 = \frac{(n_1-1)\,S_1^2 + (n_2-1)\,S_2^2}{n_1+n_2-2},$$

entonces, el estadístico para $\mu_1 - \mu_2$ es

$$t_\nu = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde t_ν es una t de student con $\nu = n_1 + n_2 - 2$ grados de libertad.

Por lo tanto

$$\begin{aligned} 1-\alpha &= P\left\{-t_{\alpha/2,\nu} \le t \le t_{\alpha/2,\nu}\right\} \\ &= P\Bigg\{\left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \\ &\qquad t \le \left(\overline{X}_1 - \overline{X}_2\right) + t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\Bigg\} \end{aligned}$$

luego, los intervalos de confianza del $(1-\alpha)\,\%$ para $\mu_1 - |mu_2$ son

- $\mu_1 - \mu_2 \in \left[\left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \left(\overline{X}_1 - \overline{X}_2\right) + t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$
- $\mu_1 - \mu_2 \in \left[-\infty, \left(\overline{X}_1 - \overline{X}_2\right) + t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$
- $\mu_1 - \mu_2 \in \left[\left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty\right]$

$\sigma_1^2 \neq \sigma_2^2$

Si no se tiene certeza de que $\sigma_1^2 = \sigma_2^2$, se propone el estadístico

$$t^* = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{7}$$

que se distribuye t-student con ν grados de libertad, donde

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^2/n_1}{n_1+1} + \frac{S_2^2/n_2}{n_2+1}} - 2$$

Entonces el intervalo de confianza de aproximadamente el $100\,(1-\alpha)\,\%$ para $\mu_1 - \mu_2$ con $\sigma_1^2 \neq \sigma_2^2$ es

$$\mu_1 - \mu_2 \in \left[\left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \right.$$

$$\left. \left(\overline{X}_1 - \overline{X}_2\right) + t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

8 Intervalos de confianza para razón de Varianzas

Supongamos que se toman dos muestras aleatorias independientes de las dos poblaciones de interés.

Sean X_1 y X_2 variables normales independientes con medias desconocidas μ_1 y μ_2 y varianzas desconocidas σ_1^2 y σ_2^2 respectivamente. Se busca un intervalo de confianza de $100\,(1-\alpha)\,\%$ para σ_1^2/σ_2^2.Supongamos n_1 y n_2 muestras aleatorias de X_1 y X_2 y sean S_1^2 y S_2^2 varianzas muestralres. Se sabe que

$$F = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2}$$

se distribuye F con $n_2 - 1$ y $n_1 - 1$ grados de libertad.

Por lo tanto

$$P\left\{F_{1-\frac{\alpha}{2},n_2-1,n_1-1} \leq F \leq F_{\frac{\alpha}{2},n_2-1,n_1-1}\right\} = 1-\alpha$$

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

por lo tanto

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

entonces

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{S_1^2}{S_2^2}F_{1-\frac{\alpha}{2},n_2-1,n_1-1}, \frac{S_1^2}{S_2^2}F_{\frac{\alpha}{2},n_2-1,n_1-1}\right]$$

donde

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

9 Intervalos de confianza para diferencia de proporciones

Sean dos proporciones de interés p_1 y p_2. Se busca un intervalo para $p_1 - p_2$ al $100\,(1-\alpha)\,\%$.

Sean dos muestras independientes de tamaño n_1 y n_2 de poblaciones infinitas de modo que X_1 y X_2 variables aleatorias binomiales independientes con parámetros (n_1, p_1) y (n_2, p_2).

X_1 y X_2 son el número de observaciones que pertenecen a la clase de interés correspondientes. Entonces $\hat{p}_1 = \frac{X_1}{n_1}$ y $\hat{p}_2 = \frac{X_2}{n_2}$ son estimadores de p_1 y p_2 respectivamente. Supongamos que se cumple la aproximación normal a la binomial, entonces

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} - \frac{p_2(1-p_2)}{n_2}}} \sim N\,(0,1)\,\text{aproximadamente}$$

entonces

$$(\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2$$
$$\leq (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} - \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}$$

10 2. Pruebas de Hipótesis

10.1 2.1 Tipos de errores

- Una hipótesis estadística es una afirmación acerca de la distribución de probabilidad de una variable aleatoria, a menudo involucran uno o más parámetros de la distribución.
- Las hipótesis son afirmaciones respecto a la población o distribución bajo estudio, no en torno a la muestra.
- La mayoría de las veces, la prueba de hipótesis consiste en determinar si la situaci ón experimental ha cambiado
- el interés principal es decidir sobre la veracidad o falsedad de una hipótesis, a este procedimiento se le llama *prueba de hipótesis*.
- Si la información es consistente con la hipótesis, se concluye que esta es verdadera, de lo contrario que con base en la información, es falsa.

Una prueba de hipótesis está formada por cinco partes

- La hipótesis nula, denotada por H_0.

- La hipótesis alterativa, denorada por H_1.
- El estadśitico de prueba y su valor p.
- La región de rechazo.
- La conclusión.

Definición 10.1 *Las dos hipótesis en competencias son la* ***hipótesis alternativa*** H_1*, usualmente la que se desea apoyar, y la* ***hipótesis nula*** H_0*, opuesta a* H_1*.*

En general, es más fácil presentar evidencia de que H_1 es cierta, que demostrar que H_0 es falsa, es por eso que por lo regular se comienza suponiendo que H_0 es cierta, luego se utilizan los datos de la muestra para decidir si existe evidencia a favor de H_1, más que a favor de H_0, así se tienen dos conclusiones:

- Rechazar H_0 y concluir que H_1 es verdadera.
- Aceptar, no rechazar, H_0 como verdadera.

Ejemplo 10.1 *Se desea demostrar que el salario promedio por hora en cierto lugar es distinto de* $19usd$*, que es el promedio nacional. Entonces* $H_1 : \mu \neq 19$*, y* $H_0 : \mu = 19$*.*

A esta se le denomina **Prueba de hipótesis de dos colas**.

Ejemplo 10.2 *Un determinado proceso produce un promedio de* 5% *de piezas defectuosas. Se está interesado en demostrar que un simple ajuste en una máquina reducirá p, la proporción de piezas defectuosas producidas en este proceso. Entonces se tiene* $H_0 : p < 0.3$ *y* $H_1 : p = 0.03$*. Si se puede rechazar* H_0*, se concluye que el proceso ajustado produce menos del* 5% *de piezas defectuosas.*

A esta se le denomina **Prueba de hipótesis de una cola**.

La decisión de rechazar o aceptar la hipótesis nula está basada en la información contenida en una muestra proveniente de la población de interés. Esta información tiene estas formas

- **Estadśitico de prueba:** un sólo número calculado a partir de la muestra.
- **p-value:** probabilidad calculada a partir del estadśtico de prueba.

Definición 10.2 *El p-value es la probabilidad de observar un estadístico de prueba tanto o más alejado del valor obervado, si en realidad* H_0 *es verdadera. Valores grandes del estadśtica de prueba y valores pequeños de p significan que se ha observado un evento muy poco probable, si* H_0 *en realidad es verdadera.*

Todo el conjunto de valores que puede tomar el estadístico de prueba se divide en dos regiones. Un conjunto, formado de valores que apoyan la hipótesis alternativa y llevan a rechazar H_0, se denomina **región de rechazo**. El otro, conformado por los valores que sustentatn la hipótesis nula, se le denomina **región de aceptación**.

Cuando la región de rechazo está en la cola izquierda de la distribución, la prueba se denomina **prueba lateral izquierda**. Una prueba con región de rechazo en la cola derecha se le llama **prueba lateral derecha**.

Si el estadśtico de prueba cae en la región de rechazo, entonces se rechaza H_0. Si el estadśtico de prueba cae en la región de aceptación, entonces la hipótesis nula se acepta o la prueba se juzga como no concluyente.

Dependiendo del nivel de confianza que se desea agregar a las conclusiones de la prueba, y el **nivel de significancia** α, el riesgo que está dispuesto a correr si se toma una decisión incorrecta.

Definición 10.3 *Un* ***error de tipo I*** *para una prueba estadística es el error que se tiene al rechazar la hipótesis nula cuando es verdadera. El* ***nivel de significancia*** *para una prueba estadística de hipótesis es*

$$\begin{aligned} \alpha &= P\left\{\textit{error tipo I}\right\} = P\left\{\textit{rechazar equivocadamente } H_0\right\} \\ &= P\left\{\textit{rechazar } H_0 \textit{ cuando } H_0 \textit{ es verdadera}\right\} \end{aligned}$$

Este valor α representa el valor máximo de riesgo tolerable de rechazar incorrectamente H_0. Una vez establecido el nivel de significancia, la región de rechazo se define para poder determinar si se rechaza H_0 con un cierto nivel de confianza.

11 2.2 Muestras grandes: una media poblacional

11.1 2.2.1 Cálculo de valor p

Definición 11.1 *El* ***valor de*** *p (p-****value****) o nivel de significancia observado de un estadístico de prueba es el valor más pequeño de α para el cual H_0 se puede rechazar. El riesgo de cometer un error tipo I, si H_0 es rechazada con base en la información que proporciona la muestra.*

Nota 11.1 *Valores pequeños de p indican que el valor observado del estadśtico de prueba se encuentra alejado del valor hipotético de μ, es decir se tiene evidencia de que H_0 es falsa y por tanto debe de rechazarse.*

Nota 11.2 *Valores grandes de p indican que el estadístico de prueba observado no está alejado de la medi hipotética y no apoya el rechazo de H_0.*

Definición 11.2 *Si el valor de p es menor o igual que el nivel de significancia α, determinado previamente, entonces H_0 es rechazada y se puede concluir que los resultados son estadísticamente significativos con un nivel de confianza del* $100\,(1-\alpha)\,\%$.

Es usual utilizar la siguiente clasificación de resultados

p	H_0	Significativa
$p < 0.01$	Rechazar	Altamente
$0.01 \leq p < 0.05$	Rechazar	Estadísticamente
$0.05 \leq p < 0.1$	No rechazar	Tendencia estadística
$0.01 \leq p$	No rechazar	No son estadísticamente

Nota 11.3 *Para determinar el valor de p, encontrar el área en la cola después del estadístico de prueba. Si la prueba es de una cola, este es el valor de p. Si es de dos colas, éste valor encontrado es la mitad del valor de p. Rechazar H_0 cuando el valor de $p < \alpha$.*

Hay dos tipos de errores al realizar una prueba de hipótesis

	H_0 es Verdadera	H_0 es Falsa
Rechazar H_0	Error tipo I	$\surd$
Aceptar H_0	$\surd$	Error tipo II

Definición 11.3 *La probabilidad de cometer el error tipo II se define por β donde*

$$\begin{aligned} \beta &= P\{\textit{error tipo II}\} = P\{\textit{Aceptar equivocadamente } H_0\} \\ &= P\{\textit{Aceptar } H_0 \textit{ cuando } H_0 \textit{ es falsa}\} \end{aligned}$$

Nota 11.4 *Cuando H_0 es falsa y H_1 es verdadera, no siempre es posible especificar un valor exacto de μ, sino más bien un rango de posibles valores.En lugar de arriesgarse a tomar una decisión incorrecta, es mejor conlcuir que no hay evidencia suficiente para rechazar H_0, es decir en lugar de aceptar H_0, no rechazar H_0.*

La bondad de una prueba estadística se mide por el tamaño de α y β, ambas deben de ser pequeñas. Una manera muy efectiva de medir la potencia de la prueba es calculando el complemento del error tipo II:

$$\begin{aligned} 1-\beta &= P\{\text{Rechazar } H_0 \text{ cuando } H_0 \text{ es falsa}\} \\ &= P\{\text{Rechazar } H_0 \text{ cuando } H_1 \text{ es verdadera}\} \end{aligned}$$

Definición 11.4 *La **potencia de la prueba**, $1-\beta$, mide la capacidad de que la prueba funciona como se necesita.*

Ejemplo 11.1 *La producción diariade una planta química local ha promediado 880 toneladas en los últimos años. A la gerente de control de calidad le gustaría saber si este promedio ha cambiado en meses recientes. Ella selecciona al azar 50 días de la base de datos computarizada y calcula el promedio y la desviación estándar de las $n = 50$ producciones como $\overline{x} = 871$ toneladas y $s = 21$ toneladas, respectivamente. Pruebe la hipótesis apropiada usando $\alpha = 0.05$.*

Solución 11.1 *La hipótesis nula apropiada es:*

$$\begin{aligned} H_0 &: \mu = 880 \\ &\quad \textit{y la hipótesis alternativa } H_1 \textit{ es} \\ H_1 &: \mu \neq 880 \end{aligned}$$

el estimador puntual para μ es $\overline{x}$, entonces el estadístico de prueba es

$$\begin{aligned} z &= \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{871 - 880}{21/\sqrt{50}} = -3.03 \end{aligned}$$

Solución 11.2 *Para esta prueba de dos colas, hay que determinar los dos valores de $z_{\alpha/2}$, es decir, $z_{\alpha/2} = \pm 1.96$, como $z > z_{\alpha/2}$, z cae en la zona de rechazo, por lo tanto la gerente puede rechazar la hipótesis nula y concluir que el promedio efectivamente ha cambiado.La probabilidad de rechazar H_0 cuando esta es verdadera es de 0.05.*

Recordemos que el valor observado del estadístico de prueba es $z = -3.03$, la región de rechazo más pequeña que puede usarse y todavía seguir rechazando H_0 es $|z| > 3.03$, entonces $p = 2(0.012) = 0.0024$, que a su vez es menor que el nivel de significancia α asignado inicialmente, y además los resultados son ***altamente significativos****.*

Finalmente determinemos la potencia de la prueba cuando μ en realidad es igual a 870 toneladas.

Recordar que la región de aceptación está entre -1.96 y 1.96, para $\mu = 880$, equivalentemente

$$874.18 < \overline{x} < 885.82$$

β es la probabilidad de aceptar H_0 cuando $\mu = 870$, calculemos los valores de z correspondientes a 874.18 y 885.82 Entonces

$$\begin{aligned} z_1 &= \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{874.18 - 870}{21/\sqrt{50}} = 1.41 \\ z_1 &= \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{885.82 - 870}{21/\sqrt{50}} = 5.33 \end{aligned}$$

por lo tanto

$$\begin{aligned} \beta &= P\{\text{aceptar } H_0 \text{ cuando } H_0 \text{ es falsa}\} \\ &= P\{874.18 < \mu < 885.82 \text{ cuando } \mu = 870\} \\ &= P\{1.41 < z < 5.33\} = P\{1.41 < z\} \\ &= 1 - 0.9207 = 0.0793 \end{aligned}$$

entonces, la potencia de la prueba es

$$1 - \beta = 1 - 0.0793 = 0.9207$$

que es la probabilidad de rechazar correctamente H_0 cuando H_0 es falsa.

Determinar la potencia de la prueba para distintos valores de H_1 y graficarlos, *curva de potencia*

H_1	$(1-\beta)$
865	
870	
872	
875	
877	
880	
883	
885	
888	
890	
895	

1. Encontrar las regiones de rechazo para el estadístico z, para una prueba de

   a) dos colas para $\alpha = 0.01, 0.05, 0.1$

   b) una cola superior para $\alpha = 0.01, 0.05, 0.1$

   c) una cola inferior para $\alpha = 0.01, 0.05, 0.1$

2. Suponga que el valor del estadístico de prueba es

   a) $z = -2.41$, sacar las conclusiones correspondientes para los incisos anteriores.

   b) $z = 2.16$, sacar las conclusiones correspondientes para los incisos anteriores.

   c) $z = 1.15$, sacar las conclusiones correspondientes para los incisos anteriores.

   d) $z = -2.78$, sacar las conclusiones correspondientes para los incisos anteriores.

   e) $z = -1.81$, sacar las conclusiones correspondientes para los incisos anteriores.

3. Encuentre el valor de p para las pruebas de hipótesis correspondientes a los valores de z del ejercicio anterior.

4. Para las pruebas dadas en el ejercicio 2, utilice el valor de p, determinado en el ejercicio 3, para determinar la significancia de los resultados.

5. Una muestra aleatoria de $n = 45$ observaciones de una población con media $\overline{x} = 2.4$, y desviación estándar $s = 0.29$. Suponga que el objetivo es demostrar que la media poblacional μ excede 2.3.

   a) Defina la hipótesis nula y alternativa para la prueba.

   b) Determine la región de rechazo para un nivel de significancia de: $\alpha = 0.1, 0.05, 0.01$.

   c) Determine el error estándar de la media muestral.

d) Calcule el valor de p para los estadísticos de prueba definidos en los incisos anteriores.

e) Utilice el valor de p pra sacar una conclusión al nivel de significancia α.

f) Determine el valor de β cuando $\mu = 2.5$

g) Graficar la curva de potencia para la prueba.

11.2 2.2.2 Prueba de hipótesis para la diferencia entre dos medias poblacionales

El estadístico que resume la información muestral respecto a la diferencia en medias poblacionales $(\mu_1 - \mu_2)$ es la diferencia de las medias muestrales $(\overline{x}_1 - \overline{x}_2)$, por tanto al probar la difencia entre las medias muestrales se verifica que la diferencia real entre las medias poblacionales difiere de un valor especificado, $(\mu_1 - \mu_2) = D_0$, se puede usar el error estándar de $(\overline{x}_1 - \overline{x}_2)$, es decir

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

cuyo estimador está dado por

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

El procedimiento para muestras grandes es:

1) **Hipótesis Nula** $H_0 : (\mu_1 - \mu_2) = D_0$,

   donde D_0 es el valor, la diferencia, específico que se desea probar. En algunos casos se querrá demostrar que no hay diferencia alguna, es decir $D_0 = 0$.

2) **Hipótesis Alternativa**

Prueba de una Cola	**Prueba de dos colas**
$H_1 : (\mu_1 - \mu_2) > D_0$	$H_1 : (\mu_1 - \mu_2) \neq D_0$
$H_1 : (\mu_1 - \mu_2) < D_0$	

3) Estadístico de prueba:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	**Prueba de dos col**
$z > z_0$	
$z < -z_\alpha$ cuando $H_1 : (\mu_1 - \mu_2) < D_0$	$z > z_{\alpha/2}$ o $z < -z_{\alpha/}$
cuando $p < \alpha$	

Ejemplo 11.2 *Para determinar si ser propietario de un automóvil afecta el rendimiento académico de un estudiante, se tomaron dos muestras aleatorias de 100 estudiantes varones. El promedio de calificaciones para los $n_1 = 100$ no propietarios de un auto tuvieron un promedio y varianza de $\overline{x}_1 = 2.7$ y $s_1^2 = 0.36$, respectivamente, mientras que para para la*

segunda muestra con $n_2 = 100$ propietarios de un auto, se tiene $\overline{x}_2 = 2.54$ y $s_2^2 = 0.4$. Los datos presentan suficiente evidencia para indicar una diferencia en la media en el rendimiento académico entre propietarios y no propietarios de un automóvil? Hacer pruebas para $\alpha = 0.01, 0.05$ y $\alpha = 0.1$.

Solución 11.3

- *Solución utilizando la técnica de regiones de rechazo:realizando las operaciones $z = 1.84$, determinar si excede los valores de $z_{\alpha/2}$.*
- *Solución utilizando el p-value:Calcular el valor de p, la probabilidad de que z sea mayor que $z = 1.84$ o menor que $z = -1.84$, se tiene que $p = 0.0658$. Concluir.*
- Si el intervalo de confianza que se construye contiene el valor del parámetro especificado por H_0, entonces ese valor es uno de los posibles valores del parámetro y H_0 no debe ser rechazada.
- Si el valor hipotético se encuentra fuera de los límites de confianza, la hipótesis nula es rechazada al nivel de significancia α.

1. Del libro Mendenhall resolver los ejercicios 9.18, 9.19 y 9.20(Mendenhall).
2. Del libro Mendenhall resolver los ejercicios: 9.23, 9.26 y 9.28.

11.3 2.2.3 Prueba de Hipótesis para una Proporción Binomial

Para una muestra aleatoria de n intentos idénticos, de una población binomial, la proporción muesrtal $\hat{p}$ tiene una distribución aproximadamente normal cuando n es grande, con media p y error estándar

$$SE = \sqrt{\frac{pq}{n}}.$$

La prueba de hipótesis de la forma

$$\begin{aligned} H_0 &: p = p_0 \\ H_1 &: p > p_0, \text{ o } p < p_0 \text{ o } p \neq p_0 \end{aligned}$$

El estadístico de prueba se construye con el mejor estimador de la proporción verdadera, $\hat{p}$, con el estadístico de prueba z, que se distribuye normal estándar.

El procedimiento es

1) Hipótesis nula: $H_0 : p = p_0$

2) Hipótesis alternativa

Prueba de una Cola	Prueba de dos colas
$H_1 : p > p_0$	$p \neq p_0$
$H_1 : p < p_0$	

3) Estadístico de prueba:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{pq}{n}}}, \hat{p} = \frac{x}{n}$$

donde x es el número de éxitos en n intentos binomiales.

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$z > z_0$ $z < -z_\alpha$ cuando $H_1 : p < p_0$ cuando $p < \alpha$	$z > z_{\alpha/2}$ o $z < -z_{\alpha/2}$

Ejemplo 11.3 *A cualquier edad, alrededor del 20% de los adultos de cierto país realiza actividades de acondicionamiento físico al menos dos veces por semana. En una encuesta local de $n = 100$ adultos de más de 40 a nos, un total de 15 personas indicaron que realizaron actividad física al menos dos veces por semana. Estos datos indican que el porcentaje de participación para adultos de más de 40 a nos de edad es considerablemente menor a la cifra del 20%? Calcule el valor de p y úselo para sacar las conclusiones apropiadas.*

1. Resolver los ejercicios: 9.30, 9.32, 9.33, 9.35 y 9.39.

11.4 2.2.4 Prueba de Hipótesis diferencia entre dos Proporciones Binomiales

Nota 11.5 *Cuando se tienen dos muestras aleatorias independientes de dos poblaciones binomiales, el objetivo del experimento puede ser la diferencia $(p_1 - p_2)$ en las proporciones de individuos u objetos que poseen una característica especifica en las dos poblaciones. En este caso se pueden utilizar los estimadores de las dos proporciones $(\hat{p}_1 - \hat{p}_2)$ con error estándar dado por*

$$SE = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

considerando el estadístico z con un nivel de significancia $(1 - \alpha)\, 100\%$

Nota 11.6 *La hipótesis nula a probarse es de la forma*

H_0: $p_1 = p_2$ o equivalentemente $(p_1 - p_2) = 0$, contra una hipótesis alternativa H_1 de una o dos colas.

Nota 11.7 *Para estimar el error estándar del estadístico z, se debe de utilizar el hecho de que suponiendo que H_0 es verdadera, las dos proporciones son iguales a algún valor común, p. Para obtener el mejor estimador de p es*

$$p = \frac{\textit{número total de éxitos}}{\textit{Número total de pruebas}} = \frac{x_1 + x_2}{n_1 + n_2}$$

1) **Hipótesis Nula:** $H_0 : (p_1 - p_2) = 0$

2) **Hipótesis Alternativa:** H_1 :

Prueba de una Cola	Prueba de dos colas
$H_1 : (p_1 - p_2) > 0$	$H_1 : (p_1 - p_2) \neq 0$
$H_1 : (p_1 - p_2) < 0$	

3) Estadístico de prueba:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}$$

donde $\hat{p_1} = x_1/n_1$ y $\hat{p_2} = x_2/n_2$, dado que el valor común para p_1 y p_2 es p, entonces $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ y por tanto el estadístico de prueba es

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$z > z_\alpha$	
$z < -z_\alpha$ cuando $H_1 : p < p_0$	$z > z_{\alpha/2}$ o $z < -z_{\alpha/2}$
cuando $p < \alpha$	

Ejemplo 11.4 *Los registros de un hospital, indican que 52 hombres de una muestra de 1000 contra 23 mujeres de una muestra de 1000 fueron ingresados por enfermedad del corazón. Estos datos presentan suficiente evidencia para indicar un porcentaje más alto de enfermedades del corazón entre hombres ingresados al hospital?, utilizar distintos niveles de confianza de* α.

1. Resolver los ejercicios 9.42

2. Resolver los ejercicios: 9.45, 9.48, 9.50

12 2.3 Muestras Pequeñas

12.1 2.3.1 Una media poblacional

1) **Hipótesis Nula:** $H_0 : \mu = \mu_0$

2) **Hipótesis Alternativa:** H_1 :

Prueba de una Cola	Prueba de dos colas
$H_1 : \mu > \mu_0$	$H_1 : \mu \neq \mu_0$
$H_1 : \mu < \mu 0$	

3) Estadístico de prueba:

$$t = \frac{\overline{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$t > t_\alpha$	
$t < -t_\alpha$ cuando $H_1 : \mu < mu_0$	$t > t_{\alpha/2}$ o $t < -t_{\alpha/2}$
cuando $p < \alpha$	

Ejemplo 12.1 *Las etiquetas en latas de un gal'on de pintura por lo general indican el tiempo de secado y el área puede cubrir una capa. Casi todas las marcas de pintura indican que, en una capa, un galón cubrirá entre 250 y 500 pies cuadrados, dependiento de la textura de la superficie a pintarse, un fabricante, sin embargo afirma que un galón de su pintura cubrirá 400 pies cuadrados de área superficial. Para probar su afirmación, una muestra aleatoria de 10 latas de un galón de pintura blanca se empleó para pintar 10 áreas idénticas usando la misma clase de equipo. Las áreas reales en pies cuadrados cubiertas por estos 10 galones de pintura se dan a continuacón:*

310	*311*	*412*	*368*	*447*
376	*303*	*410*	*365*	*350*

Ejemplo 12.2 *Los datos presentan suficiente evidencia para indicar que el promedio de la cobertura difiere de 400 pies cuadrados? encuentre el valor de p para la prueba y úselo para evaluar la significancia de los resultados.*

1. Resolver los ejercicios: 10.2, 10.3,10.5, 10.7, 10.9, 10.13 y 10.16

12.2 2.3.2 Diferencia entre dos medias poblacionales: M.A.I.

Nota 12.1 *Cuando los tama nos de muestra son peque nos, no se puede asegurar que las medias muestrales sean normales, pero si las poblaciones originales son normales, entonces la distribución muestral de la diferencia de las medias muestales, $(\overline{x}_1 - \overline{x}_2)$, será normal con media $(\mu_1 - \mu_2)$ y error estándar*

$$ES = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

1) **Hipótesis Nula** $H_0 : (\mu_1 - \mu_2) = D_0$,

   donde D_0 es el valor, la diferencia, específico que se desea probar. En algunos casos se querrá demostrar que no hay diferencia alguna, es decir $D_0 = 0$.

2) **Hipótesis Alternativa**

Prueba de una Cola	Prueba de dos colas
$H_1 : (\mu_1 - \mu_2) > D_0$	$H_1 : (\mu_1 - \mu_2) \neq D_0$
$H_1 : (\mu_1 - \mu_2) < D_0$	

3) Estadístico de prueba:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

donde

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

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos col
$z > z_0$	
$z < -z_\alpha$ cuando $H_1 : (\mu_1 - \mu_2) < D_0$	$z > z_{\alpha/2}$ o $z < -z_{\alpha/}$
cuando $p < \alpha$	

Los valores críticos de t, $t_{-\alpha}$ y $t_{\alpha/2}$ están basados en $(n_1 + n_2 - 2)$ grados de libertad.

12.3 2.3.3 Diferencia entre dos medias poblacionales: Diferencias Pareadas

1) **Hipótesis Nula:** $H_0 : \mu_d = 0$

2) **Hipótesis Alternativa:** $H_1 : \mu_d$

Prueba de una Cola	Prueba de dos colas
$H_1 : \mu_d > 0$	$H_1 : \mu_d \neq 0$
$H_1 : \mu_d < 0$	

3) Estadístico de prueba:

$$t = \frac{\overline{d}}{\sqrt{\frac{s_d^2}{n}}}$$

donde n es el número de diferencias pareadas, $\overline{d}$ es la media de las diferencias muestrales, y s_d es la desviación estándar de las diferencias muestrales.

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$t > t_\alpha$	
$t < -t_\alpha$ cuando $H_1 : \mu < mu_0$	$t > t_{\alpha/2}$ o $t < -t_{\alpha/2}$
cuando $p < \alpha$	

Los valores críticos de t, $t_{-\alpha}$ y $t_{\alpha/2}$ están basados en $(n_1 + n_2 - 2)$ grados de libertad.

12.4 2.3.4 Inferencias con respecto a la Varianza Poblacional

1) **Hipótesis Nula:** $H_0 : \sigma^2 = \sigma_0^2$

2) **Hipótesis Alternativa:** H_1

Prueba de una Cola	Prueba de dos colas
$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$
$H_1 : \sigma^2 < \sigma_0^2$	

3) Estadístico de prueba:

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2}$$

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$\chi^2 > \chi^2_\alpha$	
$\chi^2 < \chi^2_{(1-\alpha)}$ cuando $H_1 : \chi^2 < \chi^2_0$	$\chi^2 > \chi^2_{\alpha/2}$ o $\chi^2 < \chi^2_{(1-\alpha/2}$
cuando $p < \alpha$	

Los valores críticos de χ^2,están basados en (n_1+) grados de libertad.

12.5 2.3.5 Comparación de dos varianzas poblacionales

1) **Hipótesis Nula** $H_0 : (\sigma_1^2 - \sigma_2^2) = D_0$,

donde D_0 es el valor, la diferencia, específico que se desea probar. En algunos casos se querrá demostrar que no hay diferencia alguna, es decir $D_0 = 0$.

2) **Hipótesis Alternativa**

Prueba de una Cola	Prueba de dos colas
$H_1 : (\sigma_1^2 - \sigma_2^2) > D_0$	$H_1 : (\sigma_1^2 - \sigma_2^2) \neq D_0$
$H_1 : (\sigma_1^2 - \sigma_2^2) < D_0$	

3) Estadístico de prueba:

$$F = \frac{s_1^2}{s_2^2}$$

donde s_1^2 es la varianza muestral más grande.

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$F > F_\alpha$	$F > F_{\alpha/2}$
cuando $p < \alpha$	

13 2. Pruebas de Hipótesis

13.1 2.1 Tipos de errores

- Una hipótesis estadística es una afirmación acerca de la distribución de probabilidad de una variable aleatoria, a menudo involucran uno o más parámetros de la distribución.
- Las hipótesis son afirmaciones respecto a la población o distribución bajo estudio, no en torno a la muestra.
- La mayoría de las veces, la prueba de hipótesis consiste en determinar si la situaci ón experimental ha cambiado
- el interés principal es decidir sobre la veracidad o falsedad de una hipótesis, a este procedimiento se le llama *prueba de hipótesis*.
- Si la información es consistente con la hipótesis, se concluye que esta es verdadera, de lo contrario que con base en la información, es falsa.

Una prueba de hipótesis está formada por cinco partes

- La hipótesis nula, denotada por H_0.
- La hipótesis alterativa, denorada por H_1.
- El estadśitico de prueba y su valor p.
- La región de rechazo.
- La conclusión.

Definición 13.1 *Las dos hipótesis en competencias son la* ***hipótesis alternativa*** H_1*, usualmente la que se desea apoyar, y la* ***hipótesis nula*** H_0*, opuesta a* H_1*.*

En general, es más fácil presentar evidencia de que H_1 es cierta, que demostrar que H_0 es falsa, es por eso que por lo regular se comienza suponiendo que H_0 es cierta, luego se utilizan los datos de la muestra para decidir si existe evidencia a favor de H_1, más que a favor de H_0, así se tienen dos conclusiones:

- Rechazar H_0 y concluir que H_1 es verdadera.
- Aceptar, no rechazar, H_0 como verdadera.

Ejemplo 13.1 *Se desea demostrar que el salario promedio por hora en cierto lugar es distinto de* $19usd$*, que es el promedio nacional. Entonces* $H_1 : \mu \neq 19$*, y* $H_0 : \mu = 19$*.*

A esta se le denomina **Prueba de hipótesis de dos colas**.

Ejemplo 13.2 *Un determinado proceso produce un promedio de* 5% *de piezas defectuosas. Se está interesado en demostrar que un simple ajuste en una máquina reducirá p, la proporción de piezas defectuosas producidas en este proceso. Entonces se tiene* $H_0 : p < 0.3$ *y* $H_1 : p = 0.03$*. Si se puede rechazar* H_0*, se concluye que el proceso ajustado produce menos del* 5% *de piezas defectuosas.*

A esta se le denomina **Prueba de hipótesis de una cola**.

La decisión de rechazar o aceptar la hipótesis nula está basada en la información contenida en una muestra proveniente de la población de interés. Esta información tiene estas formas

- **Estadśitico de prueba:** un sólo número calculado a partir de la muestra.
- p**-value:** probabilidad calculada a partir del estadśtico de prueba.

Definición 13.2 *El p-value es la probabilidad de observar un estadístico de prueba tanto o más alejado del valor obervado, si en realidad* H_0 *es verdadera. Valores grandes del estadśtica de prueba y valores pequeños de p significan que se ha observado un evento muy poco probable, si* H_0 *en realidad es verdadera.*

Todo el conjunto de valores que puede tomar el estadístico de prueba se divide en dos regiones. Un conjunto, formado de valores que apoyan la hipótesis alternativa y llevan a rechazar H_0, se denomina **región de rechazo**. El otro, conformado por los valores que sustentatn la hipótesis nula, se le denomina **región de aceptación**.

Cuando la región de rechazo está en la cola izquierda de la distribución, la prueba se denomina **prueba lateral izquierda**. Una prueba con región de rechazo en la cola derecha se le llama **prueba lateral derecha**.

Si el estadśtico de prueba cae en la región de rechazo, entonces se rechaza H_0. Si el estadśtico de prueba cae en la región de aceptación, entonces la hipótesis nula se acepta o la prueba se juzga como no concluyente.

Dependiendo del nivel de confianza que se desea agregar a las conclusiones de la prueba, y el **nivel de significancia** α, el riesgo que está dispuesto a correr si se toma una decisión incorrecta.

Definición 13.3 *Un **error de tipo I** para una prueba estadística es el error que se tiene al rechazar la hipótesis nula cuando es verdadera. El **nivel de significancia** para una prueba estadística de hipótesis es*

$$\begin{aligned} \alpha &= P\{\textit{error tipo I}\} = P\{\textit{rechazar equivocadamente } H_0\} \\ &= P\{\textit{rechazar } H_0 \textit{ cuando } H_0 \textit{ es verdadera}\} \end{aligned}$$

Este valor α representa el valor máximo de riesgo tolerable de rechazar incorrectamente H_0. Una vez establecido el nivel de significancia, la región de rechazo se define para poder determinar si se rechaza H_0 con un cierto nivel de confianza.

14 2.2 Muestras grandes: una media poblacional

14.1 2.2.1 Cálculo de valor p

Definición 14.1 *El **valor de** p (p-**value**) o nivel de significancia observado de un estadístico de prueba es el valor más pequeño de α para el cual H_0 se puede rechazar. El riesgo de cometer un error tipo I, si H_0 es rechazada con base en la información que proporciona la muestra.*

Nota 14.1 *Valores pequeños de p indican que el valor observado del estadśtico de prueba se encuentra alejado del valor hipotético de μ, es decir se tiene evidencia de que H_0 es falsa y por tanto debe de rechazarse.*

Nota 14.2 *Valores grandes de p indican que el estadístico de prueba observado no está alejado de la medi hipotética y no apoya el rechazo de H_0.*

Definición 14.2 *Si el valor de p es menor o igual que el nivel de significancia α, determinado previamente, entonces H_0 es rechazada y se puede concluir que los resultados son estadísticamente significativos con un nivel de confianza del $100(1-\alpha)\%$.*

Es usual utilizar la siguiente clasificación de resultados

p	H_0	Significativa
$p < 0.01$	Rechazar	Altamente
$0.01 \leq p < 0.05$	Rechazar	Estadísticamente
$0.05 \leq p < 0.1$	No rechazar	Tendencia estadística
$0.01 \leq p$	No rechazar	No son estadísticamente

Nota 14.3 *Para determinar el valor de p, encontrar el área en la cola después del estadístico de prueba. Si la prueba es de una cola, este es el valor de p. Si es de dos colas, éste valor encontrado es la mitad del valor de p. Rechazar H_0 cuando el valor de $p < \alpha$.*

Hay dos tipos de errores al realizar una prueba de hipótesis

	H_0 es Verdadera	H_0 es Falsa
Rechazar H_0	Error tipo I	$\surd$
Aceptar H_0	$\surd$	Error tipo II

Definición 14.3 *La probabilidad de cometer el error tipo II se define por β donde*

$$\begin{aligned} \beta &= P\{\textit{error tipo II}\} = P\{\textit{Aceptar equivocadamente } H_0\} \\ &= P\{\textit{Aceptar } H_0 \textit{ cuando } H_0 \textit{ es falsa}\} \end{aligned}$$

Nota 14.4 *Cuando H_0 es falsa y H_1 es verdadera, no siempre es posible especificar un valor exacto de μ, sino más bien un rango de posibles valores.En lugar de arriesgarse a tomar una decisión incorrecta, es mejor conlcuir que no hay evidencia suficiente para rechazar H_0, es decir en lugar de aceptar H_0, no rechazar H_0.*

La bondad de una prueba estadística se mide por el tamaño de α y β, ambas deben de ser pequeñas. Una manera muy efectiva de medir la potencia de la prueba es calculando el complemento del error tipo II:

$$\begin{aligned} 1-\beta &= P\{\text{Rechazar } H_0 \text{ cuando } H_0 \text{ es falsa}\} \\ &= P\{\text{Rechazar } H_0 \text{ cuando } H_1 \text{ es verdadera}\} \end{aligned}$$

Definición 14.4 *La* ***potencia de la prueba***, $1-\beta$, *mide la capacidad de que la prueba funciona como se necesita.*

Ejemplo 14.1 *La producción diariade una planta química local ha promediado 880 toneladas en los últimos años. A la gerente de control de calidad le gustaría saber si este promedio ha cambiado en meses recientes. Ella selecciona al azar 50 días de la base de datos computarizada y calcula el promedio y la desviación estándar de las $n = 50$ producciones como $\overline{x} = 871$ toneladas y $s = 21$ toneladas, respectivamente. Pruebe la hipótesis apropiada usando $\alpha = 0.05$.*

Solución 14.1 *La hipótesis nula apropiada es:*

$$\begin{aligned} H_0 &: \mu = 880 \\ & \quad \textit{y la hipótesis alternativa } H_1 \textit{ es} \\ H_1 &: \mu \neq 880 \end{aligned}$$

el estimador puntual para μ es $\overline{x}$, entonces el estadístico de prueba es

$$\begin{aligned} z &= \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{871 - 880}{21/\sqrt{50}} = -3.03 \end{aligned}$$

Solución 14.2 *Para esta prueba de dos colas, hay que determinar los dos valores de $z_{\alpha/2}$, es decir, $z_{\alpha/2} = \pm 1.96$, como $z > z_{\alpha/2}$, z cae en la zona de rechazo, por lo tanto la gerente puede rechazar la hipótesis nula y concluir que el promedio efectivamente ha cambiado.La probabilidad de rechazar H_0 cuando esta es verdadera es de 0.05.*

*Recordemos que el valor observado del estadístico de prueba es $z = -3.03$, la región de rechazo más pequeña que puede usarse y todavía seguir rechazando H_0 es $|z| > 3.03$, entonces $p = 2(0.012) = 0.0024$, que a su vez es menor que el nivel de significancia α asignado inicialmente, y además los resultados son **altamente significativos**.*

Finalmente determinemos la potencia de la prueba cuando μ en realidad es igual a 870 toneladas.

Recordar que la región de aceptación está entre -1.96 y 1.96, para $\mu = 880$, equivalentemente

$$874.18 < \overline{x} < 885.82$$

β es la probabilidad de aceptar H_0 cuando $\mu = 870$, calculemos los valores de z correspondientes a 874.18 y 885.82 Entonces

$$\begin{aligned} z_1 &= \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{874.18 - 870}{21/\sqrt{50}} = 1.41 \\ z_1 &= \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{885.82 - 870}{21/\sqrt{50}} = 5.33 \end{aligned}$$

por lo tanto

$$\begin{aligned} \beta &= P\{\text{aceptar } H_0 \text{ cuando } H_0 \text{ es falsa}\} \\ &= P\{874.18 < \mu < 885.82 \text{ cuando } \mu = 870\} \\ &= P\{1.41 < z < 5.33\} = P\{1.41 < z\} \\ &= 1 - 0.9207 = 0.0793 \end{aligned}$$

entonces, la potencia de la prueba es

$$1 - \beta = 1 - 0.0793 = 0.9207$$

que es la probabilidad de rechazar correctamente H_0 cuando H_0 es falsa.

Determinar la potencia de la prueba para distintos valores de H_1 y graficarlos, *curva de potencia*

H_1	$(1-\beta)$
865	
870	
872	
875	
877	
880	
883	
885	
888	
890	
895	

1. Encontrar las regiones de rechazo para el estadístico z, para una prueba de
   a) dos colas para $\alpha = 0.01, 0.05, 0.1$
   b) una cola superior para $\alpha = 0.01, 0.05, 0.1$
   c) una cola inferior para $\alpha = 0.01, 0.05, 0.1$
2. Suponga que el valor del estadístico de prueba es
   a) $z = -2.41$, sacar las conclusiones correspondientes para los incisos anteriores.
   b) $z = 2.16$, sacar las conclusiones correspondientes para los incisos anteriores.
   c) $z = 1.15$, sacar las conclusiones correspondientes para los incisos anteriores.
   d) $z = -2.78$, sacar las conclusiones correspondientes para los incisos anteriores.
   e) $z = -1.81$, sacar las conclusiones correspondientes para los incisos anteriores.
3. Encuentre el valor de p para las pruebas de hipótesis correspondientes a los valores de z del ejercicio anterior.
4. Para las pruebas dadas en el ejercicio 2, utilice el valor de p, determinado en el ejercicio 3, para determinar la significancia de los resultados.
5. Una muestra aleatoria de $n = 45$ observaciones de una población con media $\overline{x} = 2.4$, y desviación estándar $s = 0.29$. Suponga que el objetivo es demostrar que la media poblacional μ excede 2.3.
   a) Defina la hipótesis nula y alternativa para la prueba.
   b) Determine la región de rechazo para un nivel de significancia de: $\alpha = 0.1, 0.05, 0.01$.
   c) Determine el error estándar de la media muestral.

d) Calcule el valor de p para los estadísticos de prueba definidos en los incisos anteriores.

e) Utilice el valor de p pra sacar una conclusión al nivel de significancia α.

f) Determine el valor de β cuando $\mu = 2.5$

g) Graficar la curva de potencia para la prueba.

14.2 2.2.2 Prueba de hipótesis para la diferencia entre dos medias poblacionales

El estadístico que resume la información muestral respecto a la diferencia en medias poblacionales $(\mu_1 - \mu_2)$ es la diferencia de las medias muestrales $(\overline{x}_1 - \overline{x}_2)$, por tanto al probar la difencia entre las medias muestrales se verifica que la diferencia real entre las medias poblacionales difiere de un valor especificado, $(\mu_1 - \mu_2) = D_0$, se puede usar el error estándar de $(\overline{x}_1 - \overline{x}_2)$, es decir

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

cuyo estimador está dado por

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

El procedimiento para muestras grandes es:

1) **Hipótesis Nula** $H_0 : (\mu_1 - \mu_2) = D_0$,

   donde D_0 es el valor, la diferencia, específico que se desea probar. En algunos casos se querrá demostrar que no hay diferencia alguna, es decir $D_0 = 0$.

2) **Hipótesis Alternativa**

Prueba de una Cola	**Prueba de dos colas**
$H_1 : (\mu_1 - \mu_2) > D_0$	$H_1 : (\mu_1 - \mu_2) \neq D_0$
$H_1 : (\mu_1 - \mu_2) < D_0$	

3) Estadístico de prueba:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	**Prueba de dos col**
$z > z_0$	
$z < -z_\alpha$ cuando $H_1 : (\mu_1 - \mu_2) < D_0$	$z > z_{\alpha/2}$ o $z < -z_{\alpha/}$
cuando $p < \alpha$	

Ejemplo 14.2 *Para determinar si ser propietario de un automóvil afecta el rendimiento académico de un estudiante, se tomaron dos muestras aleatorias de 100 estudiantes varones. El promedio de calificaciones para los* $n_1 = 100$ *no propietarios de un auto tuvieron un promedio y varianza de* $\overline{x}_1 = 2.7$ *y* $s_1^2 = 0.36$*, respectivamente, mientras que para para la*

segunda muestra con $n_2 = 100$ propietarios de un auto, se tiene $\overline{x}_2 = 2.54$ y $s_2^2 = 0.4$. Los datos presentan suficiente evidencia para indicar una diferencia en la media en el rendimiento académico entre propietarios y no propietarios de un automóvil? Hacer pruebas para $\alpha = 0.01, 0.05$ y $\alpha = 0.1$.

Solución 14.3

- *Solución utilizando la técnica de regiones de rechazo:realizando las operaciones $z = 1.84$, determinar si excede los valores de $z_{\alpha/2}$.*
- *Solución utilizando el p-value:Calcular el valor de p, la probabilidad de que z sea mayor que $z = 1.84$ o menor que $z = -1.84$, se tiene que $p = 0.0658$. Concluir.*
- Si el intervalo de confianza que se construye contiene el valor del parámetro especificado por H_0, entonces ese valor es uno de los posibles valores del parámetro y H_0 no debe ser rechazada.
- Si el valor hipotético se encuentra fuera de los límites de confianza, la hipótesis nula es rechazada al nivel de significancia α.

1. Del libro Mendenhall resolver los ejercicios 9.18, 9.19 y 9.20(Mendenhall).
2. Del libro Mendenhall resolver los ejercicios: 9.23, 9.26 y 9.28.

14.3 2.2.3 Prueba de Hipótesis para una Proporción Binomial

Para una muestra aleatoria de n intentos idénticos, de una población binomial, la proporción muesrtal $\hat{p}$ tiene una distribución aproximadamente normal cuando n es grande, con media p y error estándar

$$SE = \sqrt{\frac{pq}{n}}.$$

La prueba de hipótesis de la forma

$$\begin{aligned} H_0 &: p = p_0 \\ H_1 &: p > p_0, \text{ o } p < p_0 \text{ o } p \neq p_0 \end{aligned}$$

El estadístico de prueba se construye con el mejor estimador de la proporción verdadera, $\hat{p}$, con el estadístico de prueba z, que se distribuye normal estándar.

El procedimiento es

1) Hipótesis nula: $H_0 : p = p_0$

2) Hipótesis alternativa

Prueba de una Cola	Prueba de dos colas
$H_1 : p > p_0$	$p \neq p_0$
$H_1 : p < p_0$	

3) Estadístico de prueba:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{pq}{n}}}, \hat{p} = \frac{x}{n}$$

donde x es el número de éxitos en n intentos binomiales.

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$z > z_0$	
$z < -z_\alpha$ cuando $H_1 : p < p_0$	$z > z_{\alpha/2}$ o $z < -z_{\alpha/2}$
cuando $p < \alpha$	

Ejemplo 14.3 *A cualquier edad, alrededor del* 20% *de los adultos de cierto país realiza actividades de acondicionamiento físico al menos dos veces por semana. En una encuesta local de* $n = 100$ *adultos de más de* 40 *a nos, un total de 15 personas indicaron que realizaron actividad física al menos dos veces por semana. Estos datos indican que el porcentaje de participación para adultos de más de 40 a nos de edad es considerablemente menor a la cifra del* 20%*? Calcule el valor de p y úselo para sacar las conclusiones apropiadas.*

1. Resolver los ejercicios: 9.30, 9.32, 9.33, 9.35 y 9.39.

14.4 2.2.4 Prueba de Hipótesis diferencia entre dos Proporciones Binomiales

Nota 14.5 *Cuando se tienen dos muestras aleatorias independientes de dos poblaciones binomiales, el objetivo del experimento puede ser la diferencia* $(p_1 - p_2)$ *en las proporciones de individuos u objetos que poseen una característica especifica en las dos poblaciones. En este caso se pueden utilizar los estimadores de las dos proporciones* $(\hat{p}_1 - \hat{p}_2)$ *con error estándar dado por*

$$SE = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

considerando el estadístico z con un nivel de significancia $(1 - \alpha)\,100\%$

Nota 14.6 *La hipótesis nula a probarse es de la forma*

H_0: $p_1 = p_2$ *o equivalentemente* $(p_1 - p_2) = 0$*, contra una hipótesis alternativa* H_1 *de una o dos colas.*

Nota 14.7 *Para estimar el error estándar del estadístico z, se debe de utilizar el hecho de que suponiendo que* H_0 *es verdadera, las dos proporciones son iguales a algún valor común, p. Para obtener el mejor estimador de p es*

$$p = \frac{\textit{número total de éxitos}}{\textit{Número total de pruebas}} = \frac{x_1 + x_2}{n_1 + n_2}$$

1) **Hipótesis Nula:** $H_0 : (p_1 - p_2) = 0$

2) **Hipótesis Alternativa:** H_1 :

Prueba de una Cola	Prueba de dos colas
$H_1 : (p_1 - p_2) > 0$	$H_1 : (p_1 - p_2) \neq 0$
$H_1 : (p_1 - p_2) < 0$	

3) Estadístico de prueba:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}$$

donde $\hat{p_1} = x_1/n_1$ y $\hat{p_2} = x_2/n_2$, dado que el valor común para p_1 y p_2 es p, entonces $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ y por tanto el estadístico de prueba es

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$z > z_\alpha$ $z < -z_\alpha$ cuando $H_1 : p < p_0$ cuando $p < \alpha$	$z > z_{\alpha/2}$ o $z < -z_{\alpha/2}$

Ejemplo 14.4 *Los registros de un hospital, indican que 52 hombres de una muestra de 1000 contra 23 mujeres de una muestra de 1000 fueron ingresados por enfermedad del corazón. Estos datos presentan suficiente evidencia para indicar un porcentaje más alto de enfermedades del corazón entre hombres ingresados al hospital?, utilizar distintos niveles de confianza de α.*

1. Resolver los ejercicios 9.42
2. Resolver los ejercicios: 9.45, 9.48, 9.50

15 2.3 Muestras Pequeñas

15.1 2.3.1 Una media poblacional

1) **Hipótesis Nula:** $H_0 : \mu = \mu_0$

2) **Hipótesis Alternativa:** H_1 :

Prueba de una Cola	Prueba de dos colas
$H_1 : \mu > \mu_0$	$H_1 : \mu \neq \mu_0$
$H_1 : \mu < \mu 0$	

3) Estadístico de prueba:

$$t = \frac{\overline{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$t > t_\alpha$	
$t < -t_\alpha$ cuando $H_1 : \mu < mu_0$	$t > t_{\alpha/2}$ o $t < -t_{\alpha/2}$
cuando $p < \alpha$	

Ejemplo 15.1 *Las etiquetas en latas de un gal'on de pintura por lo general indican el tiempo de secado y el área puede cubrir una capa. Casi todas las marcas de pintura indican que, en una capa, un galón cubrirá entre 250 y 500 pies cuadrados, dependiento de la textura de la superficie a pintarse, un fabricante, sin embargo afirma que un galón de su pintura cubrirá 400 pies cuadrados de área superficial. Para probar su afirmación, una muestra aleatoria de 10 latas de un galón de pintura blanca se empleó para pintar 10 áreas idénticas usando la misma clase de equipo. Las áreas reales en pies cuadrados cubiertas por estos 10 galones de pintura se dan a continuacón:*

310	*311*	*412*	*368*	*447*
376	*303*	*410*	*365*	*350*

Ejemplo 15.2 *Los datos presentan suficiente evidencia para indicar que el promedio de la cobertura difiere de 400 pies cuadrados? encuentre el valor de p para la prueba y úselo para evaluar la significancia de los resultados.*

1. Resolver los ejercicios: 10.2, 10.3,10.5, 10.7, 10.9, 10.13 y 10.16

15.2 2.3.2 Diferencia entre dos medias poblacionales: M.A.I.

Nota 15.1 *Cuando los tama nos de muestra son peque nos, no se puede asegurar que las medias muestrales sean normales, pero si las poblaciones originales son normales, entonces la distribución muestral de la diferencia de las medias muestales, $(\overline{x}_1 - \overline{x}_2)$, será normal con media $(\mu_1 - \mu_2)$ y error estándar*

$$ES = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

1) **Hipótesis Nula** $H_0 : (\mu_1 - \mu_2) = D_0$,

donde D_0 es el valor, la diferencia, específico que se desea probar. En algunos casos se querrá demostrar que no hay diferencia alguna, es decir $D_0 = 0$.

2) **Hipótesis Alternativa**

Prueba de una Cola	Prueba de dos colas
$H_1 : (\mu_1 - \mu_2) > D_0$	$H_1 : (\mu_1 - \mu_2) \neq D_0$
$H_1 : (\mu_1 - \mu_2) < D_0$	

3) Estadístico de prueba:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

donde

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

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos col
$z > z_0$	
$z < -z_\alpha$ cuando $H_1 : (\mu_1 - \mu_2) < D_0$	$z > z_{\alpha/2}$ o $z < -z_{\alpha/}$
cuando $p < \alpha$	

Los valores críticos de t, $t_{-\alpha}$ y $t_{\alpha/2}$ están basados en $(n_1 + n_2 - 2)$ grados de libertad.

15.3 2.3.3 Diferencia entre dos medias poblacionales: Diferencias Pareadas

1) **Hipótesis Nula:** $H_0 : \mu_d = 0$

2) **Hipótesis Alternativa:** $H_1 : \mu_d$

Prueba de una Cola	Prueba de dos colas
$H_1 : \mu_d > 0$	$H_1 : \mu_d \neq 0$
$H_1 : \mu_d < 0$	

3) Estadístico de prueba:

$$t = \frac{\overline{d}}{\sqrt{\frac{s_d^2}{n}}}$$

donde n es el número de diferencias pareadas, $\overline{d}$ es la media de las diferencias muestrales, y s_d es la desviación estándar de las diferencias muestrales.

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$t > t_\alpha$	
$t < -t_\alpha$ cuando $H_1 : \mu < mu_0$	$t > t_{\alpha/2}$ o $t < -t_{\alpha/2}$
cuando $p < \alpha$	

Los valores críticos de t, $t_{-\alpha}$ y $t_{\alpha/2}$ están basados en $(n_1 + n_2 - 2)$ grados de libertad.

15.4 2.3.4 Inferencias con respecto a la Varianza Poblacional

1) **Hipótesis Nula:** $H_0 : \sigma^2 = \sigma_0^2$

2) **Hipótesis Alternativa:** H_1

Prueba de una Cola	Prueba de dos colas
$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$
$H_1 : \sigma^2 < \sigma_0^2$	

3) Estadístico de prueba:

$$\chi^2 = \frac{(n - 1)\, s^2}{\sigma_0^2}$$

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$\chi^2 > \chi^2_{\alpha}$ $\chi^2 < \chi^2_{(1-\alpha)}$ cuando $H_1 : \chi^2 < \chi^2_0$ cuando $p < \alpha$	$\chi^2 > \chi^2_{\alpha/2}$ o $\chi^2 < \chi^2_{(1-\alpha/2}$

Los valores críticos de χ^2,están basados en $(n_1 +)$ grados de libertad.

15.5 2.3.5 Comparación de dos varianzas poblacionales

1) **Hipótesis Nula** $H_0 : (\sigma_1^2 - \sigma_2^2) = D_0$,

donde D_0 es el valor, la diferencia, específico que se desea probar. En algunos casos se querrá demostrar que no hay diferencia alguna, es decir $D_0 = 0$.

2) **Hipótesis Alternativa**

Prueba de una Cola	Prueba de dos colas
$H_1 : (\sigma_1^2 - \sigma_2^2) > D_0$	$H_1 : (\sigma_1^2 - \sigma_2^2) \neq D_0$
$H_1 : (\sigma_1^2 - \sigma_2^2) < D_0$	

3) Estadístico de prueba:

$$F = \frac{s_1^2}{s_2^2}$$

donde s_1^2 es la varianza muestral más grande.

4) Región de rechazo: rechazar H_0 cuando

Prueba de una Cola	Prueba de dos colas
$F > F_{\alpha}$	$F > F_{\alpha/2}$
cuando $p < \alpha$	

16 Ejercicios

1) Del libro Probabililidad y Estadśitica para Ingeniería de Hines, Montgomery, Goldsman y Borror resolver los siguientes ejercicios: 10-9, 10-10,10-13,10-16 y 10-20.

2) Realizar un programa en R para cada una de las secciones y subsecciones revisadas en clase, para determinar intervalos de confianza.

3) Aplicar los programas elaborados en el ejercicio anterior a la siguiente lista: 10-39, 10-41, 10-45, 10-47, 10-48, 10-50, 10-52, 10-54, 10-56,10-57, 10-58, 10-65, 10-68, 10-72 y 10-73.

4) Elaborar una rutina en R que grafique las siguientes distribuciones, permitiendo variar los parámetros de las distribuciones: Binomial, Uniforme continua, Gamma, Beta, Exponencial, Normal y t-Student.

5) Presentar el primer capítulo del libro del curso en formato *Rnw* con su respectivo archivo *pdf* generado

17 Análisis de Regresion Lineal (RL)

Nota 17.1

- *En muchos problemas hay dos o más variables relacionadas, para medir el grado de relación se utiliza el* ***análisis de regresión****.*
- *Supongamos que se tiene una única variable dependiente, y, y varias variables independientes, $x_1, x_2, \ldots, x_n$.*
- *La variable y es una varaible aleatoria, y las variables independientes pueden ser distribuidas independiente o conjuntamente.*
- *A la relación entre estas variables se le denomina modelo regresión de y en $x_1, x_2, \ldots, x_n$, por ejemplo $y = \phi(x_1, x_2, \ldots, x_n)$, lo que se busca es una función que mejor aproxime a $\phi(\cdot)$.*

18 Análisis de Regresion Lineal (RL)

Nota 18.1

- *En muchos problemas hay dos o más variables relacionadas, para medir el grado de relación se utiliza el* ***análisis de regresión****.*
- *Supongamos que se tiene una única variable dependiente, y, y varias variables independientes, $x_1, x_2, \ldots, x_n$.*
- *La variable y es una varaible aleatoria, y las variables independientes pueden ser distribuidas independiente o conjuntamente.*
- *A la relación entre estas variables se le denomina modelo regresión de y en $x_1, x_2, \ldots, x_n$, por ejemplo $y = \phi(x_1, x_2, \ldots, x_n)$, lo que se busca es una función que mejor aproxime a $\phi(\cdot)$.*

18.1 Regresión Lineal Simple (RLS)

Supongamos que de momento solamente se tienen una variable independiente x, para la variable de respuesta y. Y supongamos que la relación que hay entre x y y es una línea recta, y que para cada observación de x, y es una variable aleatoria.

El valor esperado de y para cada valor de x es

$$E(y|x) = \beta_0 + \beta_1 x \tag{8}$$

β_0 es la ordenada al orígen y β_1 la pendiente de la recta en cuestión, ambas constantes desconocidas.

Supongamos que cada observación y se puede describir por el modelo

$$y = \beta_0 + \beta_1 x + \epsilon \tag{9}$$

donde ϵ es un error aleatorio con media cero y varianza σ^2. Para cada valor y_i se tiene ϵ_i variables aleatorias no correlacionadas, cuando se incluyen en el modelo 54, este se le llama *modelo de regresión lineal simple.*

Suponga que se tienen n pares de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, estos datos pueden utilizarse para estimar los valores de β_0 y β_1. Esta estimación es por el **métodos de mínimos cuadrados**.

Entonces la ecuación (54) se puede reescribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ para } i = 1, 2, \ldots, n. \tag{10}$$

Si consideramos la suma de los cuadrados de los errores aleatorios, es decir, el cuadrado de la diferencia entre las observaciones con la recta de regresión

$$L = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{11}$$

Para obtener los estimadores por mínimos cuadrados de β_0 y β_1, $\hat{\beta}_0$ y $\hat{\beta}_1$, es preciso calcular las derivadas parciales con respecto a β_0 y β_1, igualar a cero y resolver el sistema de ecuaciones lineales resultante:

$$\frac{\partial L}{\partial \beta_0} = 0$$
$$\frac{\partial L}{\partial \beta_1} = 0$$

evaluando en $\hat{\beta}_0$ y $\hat{\beta}_1$, se tiene

$$\begin{aligned} -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) &= 0 \\ -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i &= 0 \end{aligned}$$

simplificando

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Las ecuaciones anteriores se les denominan *ecuaciones normales de mínimos cuadrados* con solución

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{12}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{13}$$

entonces el modelo de regresión lineal simple ajustado es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{14}$$

Se intrduce la siguiente notación

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{15}$$

$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \overline{x}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \tag{16}$$

y por tanto

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{17}$$

18.2 Regresión Lineal Simple (RLS)

Supongamos que de momento solamente se tienen una variable independiente x, para la variable de respuesta y. Y supongamos que la relación que hay entre x y y es una línea recta, y que para cada observación de x, y es una variable aleatoria.

El valor esperado de y para cada valor de x es

$$E(y|x) = \beta_0 + \beta_1 x \tag{18}$$

β_0 es la ordenada al orígen y β_1 la pendiente de la recta en cuestión, ambas constantes desconocidas.

Supongamos que cada observación y se puede describir por el modelo

$$y = \beta_0 + \beta_1 x + \epsilon \tag{19}$$

donde ϵ es un error aleatorio con media cero y varianza σ^2. Para cada valor y_i se tiene ϵ_i variables aleatorias no correlacionadas, cuando se incluyen en el modelo 54, este se le llama *modelo de regresión lineal simple*.

Suponga que se tienen n pares de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, estos datos pueden utilizarse para estimar los valores de β_0 y β_1. Esta estimación es por el **métodos de mínimos cuadrados**.

Entonces la ecuación (54) se puede reescribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ para } i = 1, 2, \ldots, n. \tag{20}$$

Si consideramos la suma de los cuadrados de los errores aleatorios, es decir, el cuadrado de la diferencia entre las observaciones con la recta de regresión

$$L = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{21}$$

Para obtener los estimadores por mínimos cuadrados de β_0 y β_1, $\hat{\beta}_0$ y $\hat{\beta}_1$, es preciso calcular las derivadas parciales con respecto a β_0 y β_1, igualar a cero y resolver el sistema de ecuaciones lineales resultante:

$$\frac{\partial L}{\partial \beta_0} = 0$$
$$\frac{\partial L}{\partial \beta_1} = 0$$

evaluando en $\hat{\beta}_0$ y $\hat{\beta}_1$, se tiene

$$\begin{aligned} -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) &= 0 \\ -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i &= 0 \end{aligned}$$

simplificando

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Las ecuaciones anteriores se les denominan *ecuaciones normales de mínimos cuadrados* con solución

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{22}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{23}$$

entonces el modelo de regresión lineal simple ajustado es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{24}$$

Se intrduce la siguiente notación

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{25}$$
$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \overline{x}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \tag{26}$$

y por tanto

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{27}$$

19 3. Análisis de Regresion Lineal (RL)

Nota 19.1

- *En muchos problemas hay dos o más variables relacionadas, para medir el grado de relación se utiliza el **análisis de regresión**.*
- *Supongamos que se tiene una única variable dependiente, y, y varias variables independientes, $x_1, x_2, \ldots, x_n$.*
- *La variable y es una varaible aleatoria, y las variables independientes pueden ser distribuidas independiente o conjuntamente.*

19.1 3.1 Regresión Lineal Simple (RLS)

- A la relación entre estas variables se le denomina modelo regresión de y en $x_1, x_2, \ldots, x_n$, por ejemplo $y = \phi(x_1, x_2, \ldots, x_n)$, lo que se busca es una función que mejor aproxime a $\phi(\cdot)$.

Supongamos que de momento solamente se tienen una variable independiente x, para la variable de respuesta y. Y supongamos que la relación que hay entre x y y es una línea recta, y que para cada observación de x, y es una variable aleatoria.

El valor esperado de y para cada valor de x es

$$E(y|x) = \beta_0 + \beta_1 x \tag{28}$$

β_0 es la ordenada al orígen y β_1 la pendiente de la recta en cuestión, ambas constantes desconocidas.

19.2 3.2 Método de Mínimos Cuadrados

Supongamos que cada observación y se puede describir por el modelo

$$y = \beta_0 + \beta_1 x + \epsilon \tag{29}$$

donde ϵ es un error aleatorio con media cero y varianza σ^2. Para cada valor y_i se tiene ϵ_i variables aleatorias no correlacionadas, cuando se incluyen en el modelo 54, este se le llama *modelo de regresión lineal simple*.

Suponga que se tienen n pares de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, estos datos pueden utilizarse para estimar los valores de β_0 y β_1. Esta estimación es por el **métodos de mínimos cuadrados**.

Entonces la ecuación 54 se puede reescribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ para } i = 1, 2, \ldots, n. \tag{30}$$

Si consideramos la suma de los cuadrados de los errores aleatorios, es decir, el cuadrado de la diferencia entre las observaciones con la recta de regresión

$$L = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{31}$$

Para obtener los estimadores por mínimos cuadrados de β_0 y β_1, $\hat{\beta}_0$ y $\hat{\beta}_1$, es preciso calcular las derivadas parciales con respecto a β_0 y β_1, igualar a cero y resolver el sistema de ecuaciones lineales resultante:

$$\frac{\partial L}{\partial \beta_0} = 0$$
$$\frac{\partial L}{\partial \beta_1} = 0$$

evaluando en $\hat{\beta}_0$ y $\hat{\beta}_1$, se tiene

$$\begin{aligned} -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) &= 0 \\ -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i &= 0 \end{aligned}$$

simplificando

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Las ecuaciones anteriores se les denominan *ecuaciones normales de mínimos cuadrados* con solución

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{32}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{33}$$

entonces el modelo de regresión lineal simple ajustado es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{34}$$

Se intrduce la siguiente notación

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{35}$$
$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \overline{x}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \tag{36}$$

y por tanto

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{37}$$

19.3 3.3 Propiedades de los Estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$

Nota 19.2
- *Las propiedades estadísticas de los estimadores de mínimos cuadrados son útiles para evaluar la suficiencia del modelo.*
- *Dado que $\hat{\beta}_0$ y $\hat{\beta}_1$ son combinaciones lineales de las variables aleatorias y_i, también resultan ser variables aleatorias.*

A saber

$$\begin{aligned}
E\left(\hat{\beta}_1\right) &= E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} y_i \left(x_i - \overline{x}\right)\right) \\
&= \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i + \epsilon_i\right)\left(x_i - \overline{x}\right)\right) \\
&= \frac{1}{S_{xx}} \left[\beta_0 E\left(\sum_{k=1}^{n} \left(x_k - \overline{x}\right)\right) + E\left(\beta_1 \sum_{k=1}^{n} x_k \left(x_k - \overline{x}\right)\right)\right. \\
&+ \left. E\left(\sum_{k=1}^{n} \epsilon_k \left(x_k - \overline{x}\right)\right)\right] = \frac{1}{S_{xx}} \beta_1 S_{xx} = \beta_1
\end{aligned}$$

por lo tanto

$$E\left(\hat{\beta}_1\right) = \beta_1 \tag{38}$$

Nota 19.3 *Es decir, $\hat{\beta}_1$ es un estimador insesgado.*

Ahora calculemos la varianza:

$$\begin{aligned}
V\left(\hat{\beta}_1\right) &= V\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} V\left(\sum_{k=1}^{n} y_k \left(x_k - \overline{x}\right)\right) \\
&= \frac{1}{S_{xx}^2} \sum_{k=1}^{n} V\left(y_k \left(x_k - \overline{x}\right)\right) = \frac{1}{S_{xx}^2} \sum_{k=1}^{n} \sigma^2 \left(x_k - \overline{x}\right)^2 \\
&= \frac{\sigma^2}{S_{xx}^2} \sum_{k=1}^{n} \left(x_k - \overline{x}\right)^2 = \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

por lo tanto

$$V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \tag{39}$$

Proposición 19.1

$$
\begin{aligned}
E\left(\hat{\beta}_0\right) &= \beta_0, \\
V\left(\hat{\beta}_0\right) &= \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right), \\
Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) &= -\frac{\sigma^2\overline{x}}{S_{xx}}.
\end{aligned}
$$

Para estimar σ^2 es preciso definir la diferencia entre la observación y_k, y el valor predecido $\hat{y}_k$, es decir

$$e_k = y_k - \hat{y}_k, \text{ se le denomina } \textbf{residuo}.$$

La suma de los cuadrados de los errores de los reisduos, *suma de cuadrados del error*

$$SC_E = \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2 \tag{40}$$

sustituyendo $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ se obtiene

$$
\begin{aligned}
SC_E &= \sum_{k=1}^{n} y_k^2 - n\overline{y}^2 - \hat{\beta}_1 S_{xy} = S_{yy} - \hat{\beta}_1 S_{xy}, \\
E\left(SC_E\right) &= (n-2)\,\sigma^2, \text{ por lo tanto} \\
\hat{\sigma}^2 &= \frac{SC_E}{n-2} = MC_E \text{ es un estimador insesgado de } \sigma^2.
\end{aligned}
$$

19.4 3.4 Prueba de Hipótesis en RLS

- Para evaluar la suficiencia del modelo de regresión lineal simple, es necesario lleva a cabo una prueba de hipótesis respecto de los parámetros del modelo así como de la construcción de intervalos de confianza.
- Para poder realizar la prueba de hipótesis sobre la pendiente y la ordenada al orígen de la recta de regresión es necesario hacer el supuesto de que el error ϵ_i se distribuye normalmente, es decir $\epsilon_i \sim N\left(0,\sigma^2\right)$.

Suponga que se desea probar la hipótesis de que la pendiente es igual a una constante, $\beta_{0,1}$ las hipótesis Nula y Alternativa son:

H_0: $\beta_1 = \beta_{1,0}$,

H_1: $\beta_1 \neq \beta_{1,0}$.

donde dado que las $\epsilon_i \sim N\left(0,\sigma^2\right)$, se tiene que y_i son variables aleatorias normales $N\left(\beta_0+\beta_1 x_1,\sigma^2\right)$. De las ecuaciones (57) se desprende que $\hat{\beta}_1$ es combinación lineal de variables aleatorias normales independientes, es decir, $\hat{\beta}_1 \sim N\left(\beta_1,\sigma^2/S_{xx}\right)$, recordar las ecuaciones (63) y (64). Entonces se tiene que el estadístico de prueba apropiado es

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{MC_E/S_{xx}}} \tag{41}$$

que se distribuye t con $n-2$ grados de libertad bajo $H_0 : \beta_1 = \beta_{1,0}$. Se rechaza H_0 si

$$t_0| > t_{\alpha/2,n-2}. \tag{42}$$

Para β_0 se puede proceder de manera análoga para

$H_0 :\ \beta_0 = \beta_{0,0}$,

$H_1 :\ \beta_0 \neq \beta_{0,0}$,

con $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$, por lo tanto

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{MC_E\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}, \tag{43}$$

con el que rechazamos la hipótesis nula si

$$t_0| > t_{\alpha/2,n-2}. \tag{44}$$

- No rechazar $H_0 : \beta_1 = 0$ es equivalente a decir que no hay relación lineal entre x y y.
- Alternativamente, si $H_0 : \beta_1 = 0$ se rechaza, esto implica que x explica la variabilidad de y, es decir, podría significar que la línea recta esel modelo adecuado.

El procedimiento de prueba para $H_0 : \beta_1 = 0$ puede realizarse de la siguiente manera:

$$S_{yy} = \sum_{k=1}^{n}(y_k - \overline{y})^2 = \sum_{k=1}^{n}(\hat{y}_k - \overline{y})^2 + \sum_{k=1}^{n}(y_k - \hat{y}_k)^2$$

$$\begin{aligned}
S_{yy} &= \sum_{k=1}^{n}(y_k - \overline{y})^2 = \sum_{k=1}^{n}(y_k - \hat{y}_k + \hat{y}_k - \overline{y})^2 \\
&= \sum_{k=1}^{n}\left[(\hat{y}_k - \overline{y}) + (y_k - \hat{y}_k)\right]^2 \\
&= \sum_{k=1}^{n}\left[(\hat{y}_k - \overline{y})^2 + 2(\hat{y}_k - \overline{y})(y_k - \hat{y}_k) + (y_k - \hat{y}_k)^2\right] \\
&= \sum_{k=1}^{n}(\hat{y}_k - \overline{y})^2 + 2\sum_{k=1}^{n}(\hat{y}_k - \overline{y})(y_k - \hat{y}_k) + \sum_{k=1}^{n}(y_k - \hat{y}_k)^2
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=1}^{n}(\hat{y}_k - \overline{y})(y_k - \hat{y}_k) = \sum_{k=1}^{n}\hat{y}_k(y_k - \hat{y}_k) - \sum_{k=1}^{n}\overline{y}(y_k - \hat{y}_k) \\
= &\sum_{k=1}^{n}\hat{y}_k(y_k - \hat{y}_k) - \overline{y}\sum_{k=1}^{n}(y_k - \hat{y}_k) \\
= &\sum_{k=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_k\right)\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) - \overline{y}\sum_{k=1}^{n}\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n} \hat{\beta}_0 \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) + \sum_{k=1}^{n} \hat{\beta}_1 x_k \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) \\
&- \overline{y} \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) \\
&= \hat{\beta}_0 \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) + \hat{\beta}_1 \sum_{k=1}^{n} x_k \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) \\
&- \overline{y} \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) = 0 + 0 + 0 = 0.
\end{aligned}
$$

Por lo tanto, efectivamente se tiene

$$
S_{yy} = \sum_{k=1}^{n} \left(\hat{y}_k - \overline{y} \right)^2 + \sum_{k=1}^{n} \left(y_k - \hat{y}_k \right)^2, \tag{45}
$$

donde se hacen las definiciones

$$
SC_E = \sum_{k=1}^{n} \left(\hat{y}_k - \overline{y} \right)^2 \cdots \text{Suma de Cuadrados del Error} \tag{46}
$$

$$
SC_R = \sum_{k=1}^{n} \left(y_k - \hat{y}_k \right)^2 \cdots \text{ Suma de Regresión de Cuadrados} \tag{47}
$$

Por lo tanto la ecuación (70) se puede reescribir como

$$
S_{yy} = SC_R + SC_E \tag{48}
$$

recordemos que $SC_E = S_{yy} - \hat{\beta}_1 S_{xy}$

$$
\begin{aligned}
S_{yy} &= SC_R + \left(S_{yy} - \hat{\beta}_1 S_{xy} \right) \\
S_{xy} &= \frac{1}{\hat{\beta}_1} SC_R
\end{aligned}
$$

S_{xy} tiene $n-1$ grados de libertad y SC_R y SC_E tienen 1 y $n-2$ grados de libertad respectivamente.

Proposición 19.2

$$
E\left(SC_R\right) = \sigma^2 + \beta_1 S_{xx} \tag{49}
$$

además, SC_E y SC_R son independientes.

Recordemos que $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$. Para $H_0 : \beta_1 = 0$ verdadera,

$$
F_0 = \frac{SC_R/1}{SC_E/(n-2)} = \frac{MC_R}{MC_E}
$$

se distribuye $F_{1,n-2}$, y se rechazaría H_0 si $F_0 > F_{\alpha,1,n-2}$.

El procedimiento de prueba de hipótesis puede presentarse como la tabla de análisis de varianza siguiente

Fuente de variación	Suma de Cuadrados	Grados de Libertad	Media Cuadrática	F_0
Regresión	SC_R	1	MC_R	MC_R/MC_E
Error Residual	SC_E	$n-2$	MC_E	
Total	S_{yy}	$n-1$		

La prueba para la significación de la regresión puede desarrollarse basándose en la expresión (66), con $\hat{\beta}_{1,0} = 0$, es decir

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MC_E/S_{xx}}} \tag{50}$$

Elevando al cuadrado ambos términos:

$$t_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MC_E} = \frac{\hat{\beta}_1 S_{xy}}{MC_E} = \frac{MC_R}{MC_E}$$

Observar que $t_0^2 = F_0$, por tanto la prueba que se utiliza para t_0 es la misma que para F_0.

19.5 Estimación de Intervalos en RLS

- Además de la estimación puntual para los parámetros β_1 y β_0, es posible obtener estimaciones del intervalo de confianza de estos parámetros.
- El ancho de estos intervalos de confianza es una medida de la calidad total de la recta de regresión.

Si los ϵ_k se distribuyen normal e independientemente, entonces

$$\frac{\left(\hat{\beta}_1 - \beta_1\right)}{\sqrt{\frac{MC_E}{S_{xx}}}} \quad y \quad \frac{\left(\hat{\beta}_0 - \beta_0\right)}{\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}}$$

se distribuyen t con $n-2$ grados de libertad. Por tanto un intervalo de confianza de $100\,(1-\alpha)\,\%$ para β_1 está dado por

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}}. \tag{51}$$

De igual manera, para β_0 un intervalo de confianza al $100\,(1-\alpha)\,\%$ es

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \tag{52}$$

20 3. Análisis de Regresion Lineal (RL)

Nota 20.1
- *En muchos problemas hay dos o más variables relacionadas, para medir el grado de relación se utiliza el* ***análisis de regresión****.*
- *Supongamos que se tiene una única variable dependiente, y, y varias variables independientes, $x_1, x_2, \ldots, x_n$.*
- *La variable y es una varaible aleatoria, y las variables independientes pueden ser distribuidas independiente o conjuntamente.*

20.1 3.1 Regresión Lineal Simple (RLS)

- A la relación entre estas variables se le denomina modelo regresión de y en $x_1, x_2, \ldots, x_n$, por ejemplo $y = \phi\left(x_1, x_2, \ldots, x_n\right)$, lo que se busca es una función que mejor aproxime a $\phi\left(\cdot\right)$.

Supongamos que de momento solamente se tienen una variable independiente x, para la variable de respuesta y. Y supongamos que la relación que hay entre x y y es una línea recta, y que para cada observación de x, y es una variable aleatoria.

El valor esperado de y para cada valor de x es

$$E\left(y|x\right) = \beta_0 + \beta_1 x \tag{53}$$

β_0 es la ordenada al orígen y β_1 la pendiente de la recta en cuestión, ambas constantes desconocidas.

20.2 3.2 Método de Mínimos Cuadrados

Supongamos que cada observación y se puede describir por el modelo

$$y = \beta_0 + \beta_1 x + \epsilon \tag{54}$$

donde ϵ es un error aleatorio con media cero y varianza σ^2. Para cada valor y_i se tiene ϵ_i variables aleatorias no correlacionadas, cuando se incluyen en el modelo 54, este se le llama *modelo de regresión lineal simple*.

Suponga que se tienen n pares de observaciones $\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_n, y_n\right)$, estos datos pueden utilizarse para estimar los valores de β_0 y β_1. Esta estimación es por el **métodos de mínimos cuadrados**.

Entonces la ecuación 54 se puede reescribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ para } i = 1, 2, \ldots, n. \tag{55}$$

Si consideramos la suma de los cuadrados de los errores aleatorios, es decir, el cuadrado de la diferencia entre las observaciones con la recta de regresión

$$L = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \tag{56}$$

Para obtener los estimadores por mínimos cuadrados de β_0 y β_1, $\hat{\beta}_0$ y $\hat{\beta}_1$, es preciso calcular las derivadas parciales con respecto a β_0 y β_1, igualar a cero y resolver el sistema de ecuaciones lineales resultante:

$$\frac{\partial L}{\partial \beta_0} = 0$$
$$\frac{\partial L}{\partial \beta_1} = 0$$

evaluando en $\hat{\beta}_0$ y $\hat{\beta}_1$, se tiene

$$\begin{aligned} -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) &= 0 \\ -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i &= 0 \end{aligned}$$

simplificando

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0\sum_{i=1}^{n} x_i + \hat{\beta}_1\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Las ecuaciones anteriores se les denominan *ecuaciones normales de mínimos cuadrados* con solución

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{57}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{58}$$

entonces el modelo de regresión lineal simple ajustado es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{59}$$

Se intrduce la siguiente notación

$$S_{xx} = \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{60}$$

$$S_{xy} = \sum_{i=1}^{n} y_i\left(x_i - \overline{x}\right) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \tag{61}$$

y por tanto

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{62}$$

20.3 3.3 Propiedades de los Estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$

Nota 20.2 • *Las propiedades estadísticas de los estimadores de mínimos cuadrados son útiles para evaluar la suficiencia del modelo.*

- *Dado que $\hat{\beta}_0$ y $\hat{\beta}_1$ son combinaciones lineales de las variables aleatorias y_i, también resultan ser variables aleatorias.*

A saber

$$\begin{aligned}
E\left(\hat{\beta}_1\right) &= E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} y_i \left(x_i - \overline{x}\right)\right) \\
&= \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i + \epsilon_i\right)\left(x_i - \overline{x}\right)\right) \\
&= \frac{1}{S_{xx}} \left[\beta_0 E\left(\sum_{k=1}^{n} \left(x_k - \overline{x}\right)\right) + E\left(\beta_1 \sum_{k=1}^{n} x_k \left(x_k - \overline{x}\right)\right)\right. \\
&+ \left. E\left(\sum_{k=1}^{n} \epsilon_k \left(x_k - \overline{x}\right)\right)\right] = \frac{1}{S_{xx}} \beta_1 S_{xx} = \beta_1
\end{aligned}$$

por lo tanto

$$E\left(\hat{\beta}_1\right) = \beta_1 \tag{63}$$

Nota 20.3 *Es decir, $\hat{\beta}_1$ es un estimador insesgado.*

Ahora calculemos la varianza:

$$\begin{aligned}
V\left(\hat{\beta}_1\right) &= V\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} V\left(\sum_{k=1}^{n} y_k \left(x_k - \overline{x}\right)\right) \\
&= \frac{1}{S_{xx}^2} \sum_{k=1}^{n} V\left(y_k \left(x_k - \overline{x}\right)\right) = \frac{1}{S_{xx}^2} \sum_{k=1}^{n} \sigma^2 \left(x_k - \overline{x}\right)^2 \\
&= \frac{\sigma^2}{S_{xx}^2} \sum_{k=1}^{n} \left(x_k - \overline{x}\right)^2 = \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

por lo tanto

$$V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \tag{64}$$

Proposición 20.1

$$\begin{aligned}
E\left(\hat{\beta}_0\right) &= \beta_0, \\
V\left(\hat{\beta}_0\right) &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right), \\
Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) &= -\frac{\sigma^2 \overline{x}}{S_{xx}}.
\end{aligned}$$

Para estimar σ^2 es preciso definir la diferencia entre la observación y_k, y el valor predecido $\hat{y}_k$, es decir

$$e_k = y_k - \hat{y}_k, \text{ se le denomina } \textbf{residuo}.$$

La suma de los cuadrados de los errores de los reisduos, *suma de cuadrados del error*

$$SC_E = \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \tag{65}$$

sustituyendo $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ se obtiene

$$\begin{aligned}
SC_E &= \sum_{k=1}^{n} y_k^2 - n\overline{y}^2 - \hat{\beta}_1 S_{xy} = S_{yy} - \hat{\beta}_1 S_{xy}, \\
E(SC_E) &= (n-2)\sigma^2, \text{ por lo tanto} \\
\hat{\sigma}^2 &= \frac{SC_E}{n-2} = MC_E \text{ es un estimador insesgado de } \sigma^2.
\end{aligned}$$

20.4 3.4 Prueba de Hipótesis en RLS

- Para evaluar la suficiencia del modelo de regresión lineal simple, es necesario lleva a cabo una prueba de hipótesis respecto de los parámetros del modelo así como de la construcción de intervalos de confianza.
- Para poder realizar la prueba de hipótesis sobre la pendiente y la ordenada al orígen de la recta de regresión es necesario hacer el supuesto de que el error ϵ_i se distribuye normalmente, es decir $\epsilon_i \sim N(0, \sigma^2)$.

Suponga que se desea probar la hipótesis de que la pendiente es igual a una constante, $\beta_{0,1}$ las hipótesis Nula y Alternativa son:

H_0: $\beta_1 = \beta_{1,0}$,

H_1: $\beta_1 \neq \beta_{1,0}$.

donde dado que las $\epsilon_i \sim N(0, \sigma^2)$, se tiene que y_i son variables aleatorias normales $N(\beta_0 + \beta_1 x_1, \sigma^2)$. De las ecuaciones (57) se desprende que $\hat{\beta}_1$ es combinación lineal de variables aleatorias normales independientes, es decir, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$, recordar las ecuaciones (63) y (64).

Entonces se tiene que el estadístico de prueba apropiado es

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{MC_E/S_{xx}}} \tag{66}$$

que se distribuye t con $n-2$ grados de libertad bajo $H_0 : \beta_1 = \beta_{1,0}$. Se rechaza H_0 si

$$t_0| > t_{\alpha/2, n-2}. \tag{67}$$

Para β_0 se puede proceder de manera análoga para

$H_0 :$ $\beta_0 = \beta_{0,0}$,

$H_1:\ \beta_0 \neq \beta_{0,0}$,

con $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)\right)$, por lo tanto

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{MC_E\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]}, \tag{68}$$

con el que rechazamos la hipótesis nula si

$$t_0| > t_{\alpha/2,n-2}. \tag{69}$$

- No rechazar $H_0 : \beta_1 = 0$ es equivalente a decir que no hay relación lineal entre x y y.
- Alternativamente, si $H_0 : \beta_1 = 0$ se rechaza, esto implica que x explica la variabilidad de y, es decir, podría significar que la línea recta esel modelo adecuado.

El procedimiento de prueba para $H_0 : \beta_1 = 0$ puede realizarse de la siguiente manera:

$$S_{yy} = \sum_{k=1}^{n}\left(y_k - \overline{y}\right)^2 = \sum_{k=1}^{n}\left(\hat{y}_k - \overline{y}\right)^2 + \sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2$$

$$\begin{aligned}
S_{yy} &= \sum_{k=1}^{n}\left(y_k - \overline{y}\right)^2 = \sum_{k=1}^{n}\left(y_k - \hat{y}_k + \hat{y}_k - \overline{y}\right)^2 \\
&= \sum_{k=1}^{n}\left[\left(\hat{y}_k - \overline{y}\right) + \left(y_k - \hat{y}_k\right)\right]^2 \\
&= \sum_{k=1}^{n}\left[\left(\hat{y}_k - \overline{y}\right)^2 + 2\left(\hat{y}_k - \overline{y}\right)\left(y_k - \hat{y}_k\right) + \left(y_k - \hat{y}_k\right)^2\right] \\
&= \sum_{k=1}^{n}\left(\hat{y}_k - \overline{y}\right)^2 + 2\sum_{k=1}^{n}\left(\hat{y}_k - \overline{y}\right)\left(y_k - \hat{y}_k\right) + \sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=1}^{n}\left(\hat{y}_k - \overline{y}\right)\left(y_k - \hat{y}_k\right) = \sum_{k=1}^{n}\hat{y}_k\left(y_k - \hat{y}_k\right) - \sum_{k=1}^{n}\overline{y}\left(y_k - \hat{y}_k\right) \\
= &\sum_{k=1}^{n}\hat{y}_k\left(y_k - \hat{y}_k\right) - \overline{y}\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right) \\
= &\sum_{k=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_k\right)\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) - \overline{y}\sum_{k=1}^{n}\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right)
\end{aligned}$$

$$\begin{aligned}
= &\sum_{k=1}^{n}\hat{\beta}_0\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) + \sum_{k=1}^{n}\hat{\beta}_1 x_k\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) \\
- &\overline{y}\sum_{k=1}^{n}\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) \\
= &\hat{\beta}_0\sum_{k=1}^{n}\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) + \hat{\beta}_1\sum_{k=1}^{n}x_k\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) \\
- &\overline{y}\sum_{k=1}^{n}\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) = 0 + 0 + 0 = 0.
\end{aligned}$$

Por lo tanto, efectivamente se tiene

$$S_{yy} = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 + \sum_{k=1}^{n} (y_k - \hat{y}_k)^2, \tag{70}$$

donde se hacen las definiciones

$$SC_E = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 \cdots \text{Suma de Cuadrados del Error} \tag{71}$$

$$SC_R = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \cdots \text{ Suma de Regresión de Cuadrados} \tag{72}$$

Por lo tanto la ecuación (70) se puede reescribir como

$$S_{yy} = SC_R + SC_E \tag{73}$$

recordemos que $SC_E = S_{yy} - \hat{\beta}_1 S_{xy}$

$$\begin{aligned} S_{yy} &= SC_R + \left(S_{yy} - \hat{\beta}_1 S_{xy}\right) \\ S_{xy} &= \frac{1}{\hat{\beta}_1} SC_R \end{aligned}$$

S_{xy} tiene $n-1$ grados de libertad y SC_R y SC_E tienen 1 y $n-2$ grados de libertad respectivamente.

Proposición 20.2

$$E(SC_R) = \sigma^2 + \beta_1 S_{xx} \tag{74}$$

además, SC_E y SC_R son independientes.

Recordemos que $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$. Para $H_0 : \beta_1 = 0$ verdadera,

$$F_0 = \frac{SC_R/1}{SC_E/(n-2)} = \frac{MC_R}{MC_E}$$

se distribuye $F_{1,n-2}$, y se rechazaría H_0 si $F_0 > F_{\alpha,1,n-2}$.

El procedimiento de prueba de hipótesis puede presentarse como la tabla de análisis de varianza siguiente

Fuente de variación	Suma de Cuadrados	Grados de Libertad	Media Cuadrática	F_0
Regresión	SC_R	1	MC_R	MC_R/MC_E
Error Residual	SC_E	$n-2$	MC_E	
Total	S_{yy}	$n-1$		

La prueba para la significación de la regresión puede desarrollarse basándose en la expresión (66), con $\hat{\beta}_{1,0} = 0$, es decir

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MC_E/S_{xx}}} \tag{75}$$

Elevando al cuadrado ambos términos:

$$t_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MC_E} = \frac{\hat{\beta}_1 S_{xy}}{MC_E} = \frac{MC_R}{MC_E}$$

Observar que $t_0^2 = F_0$, por tanto la prueba que se utiliza para t_0 es la misma que para F_0.

20.5 Estimación de Intervalos en RLS

- Además de la estimación puntual para los parámetros β_1 y β_0, es posible obtener estimaciones del intervalo de confianza de estos parámetros.
- El ancho de estos intervalos de confianza es una medida de la calidad total de la recta de regresión.

Si los ϵ_k se distribuyen normal e independientemente, entonces

$$\frac{\left(\hat{\beta}_1-\beta_1\right)}{\sqrt{\frac{MC_E}{S_{xx}}}} \quad y \quad \frac{\left(\hat{\beta}_0-\beta_0\right)}{\sqrt{MC_E\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)}}$$

se distribuyen t con $n-2$ grados de libertad. Por tanto un intervalo de confianza de $100\left(1-\alpha\right)\%$ para β_1 está dado por

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}}. \tag{76}$$

De igual manera, para β_0 un intervalo de confianza al $100\left(1-\alpha\right)\%$ es

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)} \tag{77}$$

20.6 Predicción

Supongamos que se tiene un valor x_0 de interés, entonces la estimación puntual de este nuevo valor

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{78}$$

Nota 20.4 *Esta nueva observación es independiente de las utilizadas para obtener el modelo de regresión, por tanto, el intervalo en torno a la recta de regresión es inapropiado, puesto que se basa únicamente en los datos empleados para ajustar el modelo de regresión.*

El intervalo de confianza en torno a la recta de regresión se refiere a la respuesta media verdadera $x = x_0$, no a observaciones futuras.

Sea y_0 la observación futura en $x = x_0$, y sea $\hat{y}_0$ dada en la ecuación anterior, el estimador de y_0. Si se define la variable aleatoria

$$w = y_0 - \hat{y}_0,$$

esta se distribuye normalmente con media cero y varianza

$$V\left(w\right) = \sigma^2\left[1+\frac{1}{n}+\frac{\left(x-x_0\right)^2}{S_{xx}}\right]$$

dado que y_0 es independiente de $\hat{y}_0$, por lo tanto el intervalo de predicción al nivel α para futuras observaciones x_0 es

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MC_E\left[1+\frac{1}{n}+\frac{(x-x_0)^2}{S_{xx}}\right]} \leq y_0$$
$$\leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MC_E\left[1+\frac{1}{n}+\frac{(x-x_0)^2}{S_{xx}}\right]}.$$

20.7 Coeficiente de Determinación

La cantidad

$$R^2 = \frac{SC_R}{S_{yy}} = 1 - \frac{SC_E}{S_{yy}} \tag{79}$$

se denomina coeficiente de determinación y se utiliza para saber si el modelo de regresión es suficiente o no. Se puede demostrar que $0 \leq R^2 \leq 1$, una manera de interpretar este valor es que si $R^2 = k$, entonces el modelo de regresión explica el $k * 100\%$ de la variabilidad en los datos. R^2

- No mide la magnitud de la pendiente de la recta de regresión
- Un valor grande de R^2 no implica una pendiente empinada.
- No mide la suficiencia del modelo.
- Valores grandes de R^2 no implican necesariamente que el modelo de regresión proporcionará predicciones precisas para futuras observaciones.