Notas de Clase: Estadística (Parte 2)

Probabilidad, Variables Aleatorias, Funciones Especiales, Estimación, Pruebas de Hipótesis, Regresión Lineal y otras herramientas del análisis de datos

Benjamín Oliva[1]

Draft Agosto 2020

[1] benjov@ciencias.unam.mx y https://github.com/benjov

Documento siempre proceso de mejora.
Comentarios, siempre serán bienvenidos...
Este documento servirá para la segunda parte del curso de Estadística de la Maestría en Economía.

Índice general

Índice de figuras

Índice de cuadros

Introducción, motivación y alcance del documento

Este documento servirá como guía para impartir el curso de Estadística en la Maestría en Economía. No obstante, con el objeto de reforzar conocimientos iniciaremos con un repaso de algunos temas previos como probabilidad y funciones de probabilidad especiales.

En este documento buscamos implementar un enfoque moderno en el cual la estadística, la mineria de datos y el aprendizaje de máquina (o aprendizaje estadístico) sean todas herramientas para colectar y analizar datos. Durante mucho tiempo se ha considerado a la estadística de forma separada de las técnicas de aprendizaje de máquinna y mineria de datos, ya que estás últimas solían impartirse en los departamentos de Ciencias de la Computación.

Un enfoque más moderno considera las tres técnicas como parte integral de la aplicación de la estadística. Además, con este buscamos reconocer la contribución de que hacen los departamentos de Estadística y de Ciencias de la Computación para la construcción de un enfoque moderno del análisis estadístico.

En la Figura 1 mostramos las relación que garda la estadística y la probabilidad.

En el curso queremos discutir los documentos de la bibliografia como: Miller y Miller (2014) I. Miller y M. Miller 2014; Larsen y Marx (2018) Larsen y Marx 2012; Wasserman (2004) Wasserman 2004; Mood, Graybill y Boes (1973) Mood, Graybill y Boes 1973, Hastie, Tibshirani y Friedman (2017) Hastie, Tibshirani y Friedman 2017, y Wackerly, Mendenhall y Scheaffer (2017) Wackerly, Mendenhall y Scheaffer 2016.

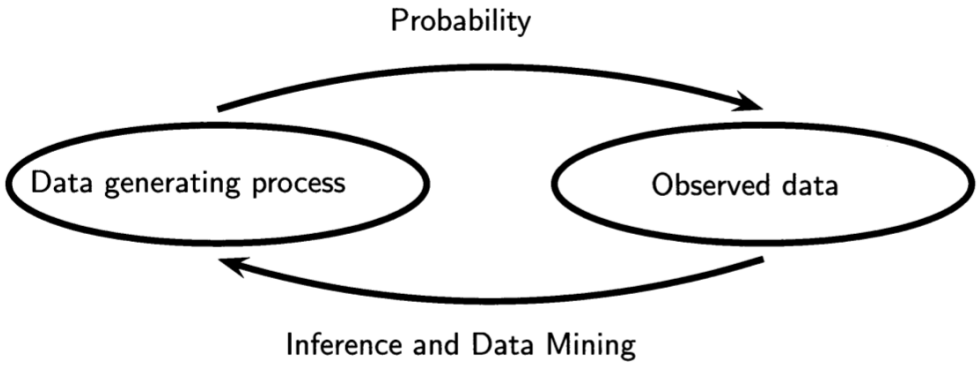

Figura 1: Esquema de la relación entre la Probabilidad y la Mineria de Datos y la Inferencia Estadística, retomado de Wasserman (2004, p. ix) Wasserman 2004

Conocimientos previos

Cálculo Diferencial e Integral y Álgebra Lineal.

Recursos en línea y otros materiales

Materiales recomendados durante el curso y autoaprendizaje.

1

Probabilidad

1.1. Espacios muestrales y eventos

La probabilidad es un lenguaje matemático para cuantificar la incertidumbre.

Definición 1.1 ***Espacio Muestral****. El espacio muestral Ω es el conjunto de los posibles resultados de un experimento.*

Definición 1.2 ***Elementos del Espacio Muestral y Eventos****. Los elemenntos $\omega \in \Omega$ se denominan resultados, realizaciones o elementos del espacio muestral, Ω. Los subconjuntos de Ω son también son llamados eventos.*

Ejemplo. Algunos ejemplos son el espacio muestral generado por el lanzamiento de dos monedas: $\Omega = \{SS, AA, AS, SA\}$. Un evento podría estar definido por obtener al menos un Aguila en cada el lanzamiento de las dos monedas: $E = \{AA, AS, SA\}$.

Ejemplo. Sea ω cualquier número real, entonces el espacio muestral podría estar definido por: $\Omega = \mathbb{R} = (-\infty, \infty)$. Un evento es considerar un intervalo como: $A = (10, 23]$.

Ejemplo. Ahora pensemos en el lanzamiento de una moneda por infinitas veces. En este caso el espacio muestral estará dado por: $\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \ldots)\} : \omega_i \in \{A, S\}$. Cada uno de los elementos del espacio muestral son los $\omega_i \in \{A, S\}$. Podríamos plantear un evento como aquel caso en el que el Aguila aparece hasta el tercer lanzamiento de la moneda: $E = \{(\omega_1, \omega_2, \omega_3, \ldots) : \omega_1 = S, \omega_2 = S, \omega_3 = A, \omega_i \in \{A, S\}, \forall i > 3\}$.

Definición 1.3 ***Complemento***. *Sea un evento A, sea $A^c = \{\omega \in \Omega : \omega \notin A\}$ al cual denominaremos como el complemento de A. Informalmente también le llamaremos como $\neg A$.*

Una observación para la definición anterior es que el complemento de Ω es el conjunto vacío $\varnothing$.

Definición 1.4 ***Unión de Eventos***. *La unión de eventos A y B está definida como:*

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ o } \omega \in B \text{ o } \omega \text{ está en ambos } \}$$

De manera informal, le llamaremos A o B. Podemos externder esta definición a casos en los que observamos una secuencia infinita de eventos $A_1, A_2, A_3, \ldots$ y obtener:

$$\cup_{i=1}^{\infty} A_i = \{\omega_i \in \Omega : \omega \in A_i \text{ para algún } i\}$$

Definición 1.5 ***Intersección de Eventos***. *La intersección de eventos A y B está definida como:*

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ y } \omega \in B\}$$

De manera informal, le llamaremos A y B. Podemos externder esta definición a casos en los que observamos una secuencia infinita de eventos $A_1, A_2, A_3, \ldots$ y obtener:

$$\cap_{i=1}^{\infty} A_i = \{\omega_i \in \Omega : \omega \in A_i, \forall i\}$$

Definición 1.6 ***Diferencia de Eventos***. *La diferencia de los eventos A y B está definida como:*

$$A - B = \{\omega \in \Omega : \omega \in A, \omega \notin B\}$$

Definición 1.7 ***Contención de Eventos***. *Definimos que A esta contennido en B si:*

$$A \subset B = \{\omega \in \Omega : \forall \omega_i \in A, \omega_i \in B\}$$

De forma similar podemos definir:

$$B \subset A = \{\omega \in \Omega : \forall \omega_i \in B, \omega_i \in A\}$$

Definición 1.8 ***Número de elementos de un evento****. Sea A un conjunto finito, definimos $|A|$ cómo el número de elementos de A.*

Definición 1.9 ***Eventos Disjuntos o Mutuamente Excluyentes****. Decimos que serie de conjuntos o eventos $A_1, A_2, A_3, \ldots$ son disjuntos o mutuamente excluyentes si:*

$$A_i \cap A_j = \varnothing, \; \forall i \neq j$$

Definición 1.10 ***Partición de*** Ω*. Sea una serie de conjuntos o eventos disjuntos dados por $A_1, A_2, A_3, \ldots$, definimos a una partición de Ω como:*

$$\cup_{i=1}^{\infty} A_i = \Omega$$

Definición 1.11 ***Función Indicadora de*** A*. Definimos a la función indicadora de A por:*

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 & si \; \omega \in A \\ 0 & si \; \omega \notin A \end{cases}$$

Definición 1.12 ***Función monótona creciente****. Una serie de conjuntos o eventos $A_1, A_2, A_3, \ldots$ es monótona creciennte si $A_1 \subset A_2 \subset A_3 \subset \ldots$ y si:*

$$\lim_{i=1}^{\infty} A_n = \cup_{i=1}^{\infty} A_i$$

Función monótona decreciente*. Una serie de conjuntos o eventos $A_1, A_2, A_3, \ldots$ es monótona decreciente si $A_1 \supset A_2 \supset A_3 \supset \ldots$ y si:*

$$\lim_{i=1}^{\infty} A_n = \cap_{i=1}^{\infty} A_i$$

En cualquiera de los casos decimos que $A_n \longrightarrow A$.

Ejemplo. Sea $\Omega = \mathbb{R}$ y sea $A_i = [0, 1/i)$ para $i = 1, 2, 3, \ldots$. Entonces ¿cuál es $A_n \longrightarrow A$ considerando las dos definiciones annteriores?

$$\cup_{i=1}^{\infty} A_i = [0, 1)$$
$$\cap_{i=1}^{\infty} A_i = \{0\}$$

¿Cómo cambia la conclusión si $A_i = (0, 1/i)$ para $i = 1, 2, 3, \ldots$?

$$\cup_{i=1}^{\infty} A_i = (0, 1)$$
$$\cap_{i=1}^{\infty} A_i = \varnothing$$

1.2. Definición de Probabilidad

2

Tema

2.1. Subtema

3

Tema

3.1. Subtema

4

Tema

4.1. Subtema

5

Modelos, Inferencia y Aprendizaje Estadísticos

5.1. Introducción

La inferencia estadística o aprendizaje estadístico (como se le conoce en ciencias de la computación) es el proceso de usar datos para inferir la distribución que generó los datos originalmente. Un planteamiento típico de la inferencia estadística es: dada una muestra $X_1, X_2, \ldots, X_n \sim F$, ¿cómo podemos inferir a F? No obstante, en algunos casos queremos inferir sólo algunas características de F tales como su media.

En una visión más moderna de la evolución en el uso de la estadística es el campo de la inferencia causal o, simplemente, causalidad. La causalidad es una serie de herramientas, o nueva ciencia para algunos, pretende responder a la pregunta ¿por qué?. Mediante el uso de herramientas estadísticas existentes y la construcción de contrafactuales, trata de explicar si un evento es causa de algún cambio o evento en particular. La aplicación más reciente es el uso de la causalidad en la inteligencia artificial con el objeto de replicar el comportamiento humano.

Sin embargo, en este curso nos enfocaremos únicamente en la inferencia estadística y al final del curso introduciremos algunas ideas del concepto de causalidad.

5.2. Modelos paramétricos y No paramétricos

Definición 5.1 ***Modelo estadístico***. *Un modelo estadístico $\mathfrak{F}$ es un conjunto de distribuciones, densidades o funciones de regresión.*

Definición 5.2 ***Modelo paramétrico***. *Un modelo paramétrico es un conjunto $\mathfrak{F}$ que puede ser parametrizado por un número finito de paramétros. En general, un modelo paramétrico toma la forma:*

$$\mathfrak{F} = \{f(x : \theta) : \theta \in \Theta\}$$

Donde θ es un paramétro desconocido que puede tomar valores en el espacio de paramétros Θ.

Definición 5.3 ***Modelo no paramétrico***. *Un modelo no paramétrico es un conjunto $\mathfrak{F}$ que no puede ser parametrizado por un número finito de paramétros.*

A continuación, veamos unos ejemplos:

Ejemplo. Sean $X_1, X_2, \ldots, X_n \sim F$ variables aleatorias, y asumamos que la función de densidad $f \in \mathfrak{F}$, donde $\mathfrak{F}$ esta dado por una normal, $N(\mu, \sigma)$, entonces el objetivo es estimar los paramétros μ y σ a partir de los datos observados.

Ejemplo. Sea $X_1, X_2, \ldots, X_n$ observaciones independientes de una función F. El problema es estimar a F asumiendo que $F \in \mathfrak{F}_{all}$.

5.3. Inferencia paramétrica

Partamos de que los modelos paramétricos son de la forma $\mathfrak{F} = \{f(x; \theta) : \theta \in \Theta\}$, donde $\Theta \subset \mathbb{R}^k$ es el espacio de paramétros θ, o $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ – para facilitar la notación utilizaremos indistintamente θ o $\boldsymbol{\theta}$. Así, el problema de inferencia se reduce al problema de estimar el paramétro θ.

A partir de muestras de datos el proceso de estimación de paramétros puede, a su vez, dividirse en dos problemas:

1. Estimación puntual: es un proceso mediante el cual determinamos un número que es la mejor adivinanza para los valores del paramétro.

2. Estimación por intervalo: es una estimación de un intervalo al rededor de la estimación puntual que creemos contiene los valores del paramétro.

Otra forma de velo es la siguiente. El análisis de inferencia estadística paramétrica está dividido en dos grupos.

1. En el primero se ubican los problemas de estimación puntual y por intervalo. Los cuales consisten en estimar el valor o el intervalo en el cual se ubica un parámetro o un conjunto de parámetros.

2. En el segundo, se ecuentran los procedimientos de pruebas de hipótesis. Los cuales consisten en establecer reglas para aceptar o rechazar enunciados, afirmaciones o hipótesis respecto del valor de un parámetro o conjunto de parámetros.

En este capítulo y los dos siguientes analizaré el primer tipo de problemas, en capítulos posteriores discuto el segundo tipo.

6

Estimación Puntual

6.1. Definiciones y ejemplos

La estimación puntual se refiere a la mejor adivinanza del valor del paramétro de interés, el cual podría ser de un modelo paramétrico, de una función de densidad de probabilidad, de una función de regresión o la predicción para el valor futuro de alguna variable aleatoria.

Partamos de las siguientes definiciones:

Definición 6.1 ***Estadística***. *Una estadística (coloquialmente mencionada como 'estadístico') es una función de las variables aleatorias observables en una muestra y de constantes conocidas.*

Definición 6.2 ***Estimación Puntual***. *Es el proceso de estimación del valor de un parámetro poblacional a partir de los valores de una muestra aleatoria. A este valor lo denominaremos como estimador y a la función que nos permite estimar el valor del parámetro la denominamos como una estadística puesto que es una función:*

$$f : \mathbb{R}^K \longrightarrow \mathbb{R}$$

Por convención denotaremos a un estimador puntual de θ a $\hat{\theta}$ o $\hat{\theta}_n$. Notemos que el estimador $\hat{\theta}$ depende de los datos incluidos en la muestra de variables aleatorias (v.a.s), por lo que $\hat{\theta}$ es en sí misma una variable aleatoria (v.a.).

Definición 6.3 ***Estimador***. *Cualquier estadística (función conocida de un conjunto de variables aleatorias que es en sí misma una variable aleatoria) para la cual sus valores son usados pra estimar a θ.*

Más formalmente, decimos que un estimador de θ es:

$$\hat{\theta}_n = g(X_1, X_2, \ldots, X_n) \tag{6.1}$$

Notemos que bajo esta definición un estimador siempre es una estadística la cual es también a la vez una variable aleatoria y una función.

Como resultado de la estimación tendremos un estimador que es en sí mismo una variable aleatoria (v.a.) que tiene algunas propiedades estadísticas muestrales. Estas propiedades pueden servir para determinar, en su caso, cuál estimador es el más adecuado dada la situación que enfrentamos. Algunas de esas propuedades son: insesgamiento, varianza mínima, eficiencia, consistencia, suficiencia y robustes. A lo largo de este capítulo discutiremos cada una de ellas.

Para motivar el significado del proceso de estimación partamos de un ejemplo. Supongamos el caso del lanzamiento de una moneda, en el cual la probabilidad p de que salga aguila es desconocida. Ahora supongamos que hacemos 3 lanzamientos de la moneda y a partir del resultado trataremos de determinar el valor de p.

Imaginemos que los resultados fueron: $\{A, A, S\}$, donde A denota aguila y S denota sol. Basados en esos resultados, ¿cuál es el valor que considera aporpiado para p? Probablemente la solución obvia es que $p = 2/3$. Pero, ¿cuál es el principio general que siguió para determinar ese valor de p?, ¿cómo podríamos generalizarlo?

Partamos de definir una variable aleatoria X, la cual describe cuando obtenemos aguila en el lanzamiento de una moneda, es decir:

$$X = k = \begin{cases} 1 & \text{si el resultado del lanzamiento es } A \\ 0 & \text{si el resultado del lanzamiento es } S \end{cases}$$

Así, el modelo general puede ser escrito como:

$$p_X(k) = p^k(1-p)^{1-k} = \begin{cases} p & \text{si } k = 1 \\ (1-p) & \text{si } k = 0 \end{cases} \tag{6.2}$$

Ahora, supongamos una sucesión de variables aleatorias independientes $\{X_i : i = 1, 2, \ldots, n\}$ de forma que podemos describir el fenómeno del lanza-

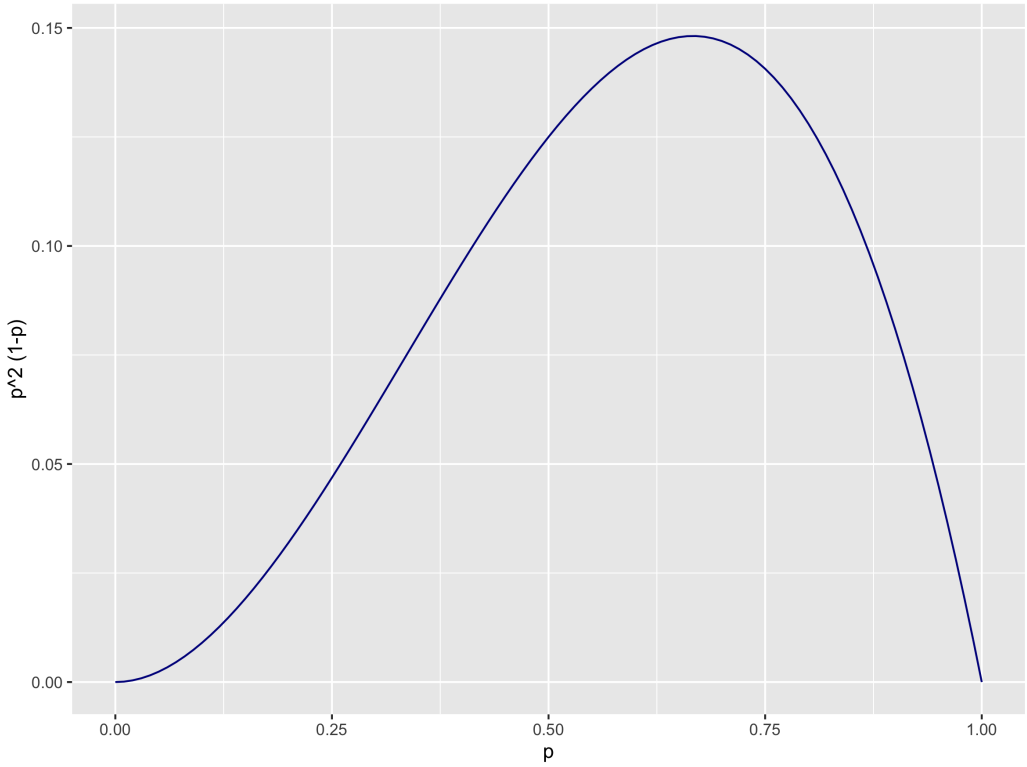

Figura 6.1: Gráfica de $p^2(1-p)$.

miento de una moneda tres veces como:

$$X_1 = 1, X_2 = 1, X_3 = 0$$

Así, de forma analóga el lanzamiento de una moneda n veces y considerando el resultado en la ecuación (6.2) estará dado por:

$$\begin{aligned} P(X_1 = 1 \cap X_2 = 1 \cap X_3 = 0) &= P(X_1 = 1) \times P(X_2 = 1) \times P(X_3 = 0) \\ &= p^2(1-p) \end{aligned}$$

Dicho esto, ¿cuál valor de p máximiza la función de probabilidad conjunta $P(X_1 = 1 \cap X_2 = 1 \cap X_3 = 0)$?. La Figura 6.1 muestra la evolución de la probabilidad y el punto en que máximiza la probabilidad.

De forma general y considerando el resultado en la ecuación (6.2), el lanzamiento de una moneda n veces estará dado por la sucesión de v.a.'s:

$$X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n$$

Donde k_i, para $i = 1, 2, \ldots, n$ es como definimos en la ecuación (6.2). Entonces, el problema general estará dado por:

$$\begin{aligned} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) &= p^{k_1}(1-p)^{1-k_1} \ldots p^{k_n}(1-p)^{1-k_n} \\ &= p^{\sum_{i=1}^n k_i}(1-p)^{n-\sum_{i=1}^n k_i} \end{aligned}$$

En este caso, para determinar el valor que máximiza la función de probabilidad debemos solucionar el problema por la forma habitual: derivar, igualar a cero y resolver para un valor de p óptimo.

$$\begin{aligned} \frac{d}{dp}\left(p^{\sum_{i=1}^n k_i}(1-p)^{n-\sum_{i=1}^n k_i}\right) &= 0 \\ \hat{p} &= \frac{\sum_{i=1}^n k_i}{n} \end{aligned}$$

En conclusión, hemos encontrado un estimador para p dado por $\hat{p} = \frac{\sum_{i=1}^n k_i}{n}$, el cual es una función de la suceción de v.a.'s dadas por $X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n$.

6.2. Propiedades de los estimadores

A continuación, discutiremos las propiedades más importantes de los estimadores. Dentro de estas, quizá las más relevantes son las propiedades de insesgamiento y eficiencia. Se suele decir que un buen estimador tiene una distribución muestral que (1) está centrado al rededor del paramétro y (2) tiene un error estándar lo más pequeño posible.

6.2.1. Insesgamiento

Existen diversas propiedades que pueden cumplir los estimadores y que vamos a analizar a lo largo del curso, tales propiedades son: insesgamiento, eficiencia, consistencia, suficiencia y mínima varaianza. La primera de ellas es la propiedad de insesgamiento o, visto desde la otra óptica, la propiedad de sesgo cero. Partamos de la siguientes definiciones.

Definición 6.4 ***Esperanza de una variable aleatoria****. Sea X una v.a.. Si X es una varialble aleatoria continua con función de densidad de probabilidad (pdf, por sus siglas en inglés) $f(x)$ que cumple con:*

$$\int_{-\infty}^{\infty} |x| \cdot f(x)dx < \infty$$

Entonces, la esperanza de X estará dada por:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Si X es una v.a. discreta con pdf $p(x)$ que cumple con:

$$\sum_{x} |x| \cdot p(x) < \infty$$

Entonces, la esperanza de X estará dada por:

$$\mathbb{E}[X] = \sum_{x} x \cdot p(x)$$

Definición 6.5 ***Estimador Insesgado (1)***. *Una estadística $\hat{\theta}$ es un estimador insesgado del parámetro θ de una función de densidad de probabilidad (pdf, por sus siglas en inglés) dada para una población si y solo si $\mathbb{E}[\hat{\theta}] = \theta$, para todos los posibles valores de θ.*

Definición 6.6 ***Estimador Insesgado (2)***. *Sean $\{X_1, X_2, \ldots, X_n\}$ una suecesión de variables aleatorias agrupadas en una muestra aleatoria de una función de densidad de probabilidad (pdf, por sus siglas en inglés), donde θ es un parámetro o conjunto de parámetros desconocidos de dicha función. Decimos que un estimador de θ denotado por $\hat{\theta}(= h(X_1, X_2, \ldots, X_n))$ es un estimador insesgado si $\mathbb{E}[\hat{\theta}] = \theta$, para todos los posibles valores de θ.*

Finalmente, introduciremos dos definiciones adicionales, una que tiene un propósito ilustrativo y otra que sirve para analizar los casos en que el estimador es sesgado para muestras pequeñas, pero insesgado para muestras grandes.

Definición 6.7 ***Estimador Insesgado (3)***. *Un estimador es insesgado si el centro de la distribución muestral está al rededor del paramétro.*

La Figurura 6.2 ilustra el panteamiento de está última definición en el caso la distribución de la media y su ubicación en la distribución poblacional. A manera de constraste, la Figura permite concluir que un estimados sesgado tenderá a subestimar un paramétro, en el promedio, o tenderá a sobrestimarlo.

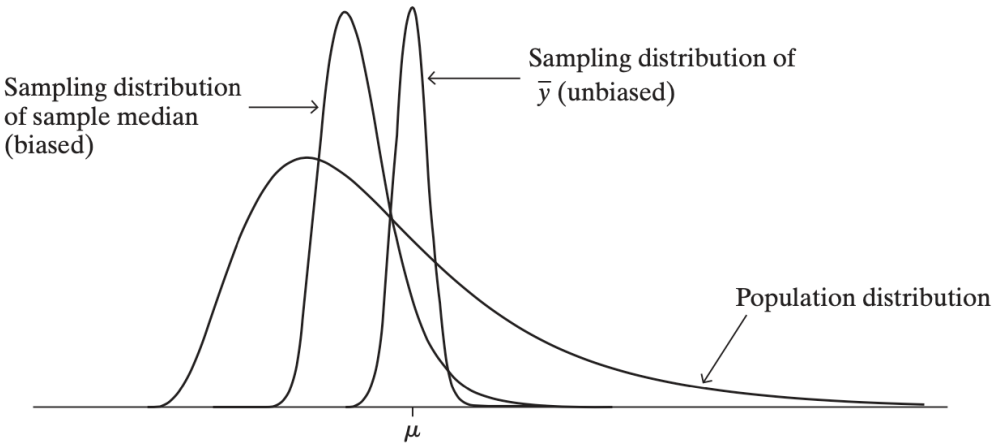

Figura 6.2: Distribución muestral de dos estimadores de la media poblacional para una distribución poblacional sesgada, retomada de Agresti, et. al. (2017; p. 104) Agresti, Franklin y Klingenberg 2017

Definición 6.8 *Sea $b_n(\theta) = \mathbb{E}[\hat{\theta}_n] - \theta$ una expresión del sesgo de un estimador $\hat{\theta}_n$ en función del tamaño de muestra n de una población con una pdf dada, decimos que $\hat{\theta}$ es un estimador asintóticamente insesgado de θ si y solo si:*

$$\lim_{n\to\infty} b_n(\theta) = 0$$

De forma relacionada, podemos decir que un estimador $\hat{\theta}_n$ es insesgado si y solo si:

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] = \theta$$

Veamos algunos ejemplos:

Ejemplo. Suponga un estimador $\hat{p}$ de un parámetro p de una distribución binomial dado por:

$$\hat{p} = \frac{X + 1/2\sqrt{n}}{n + \sqrt{n}}$$

Donde $X \sim Bin(p)$. Determine si el estimador $\hat{p}$ es un estimador insesgado de p.

Previo a determinar si el estimador propuesto es insesgado, requerimos algunos resultados. Utilizando la definición (6.4) diremos que la varianza de

una v.a. estará dado por $\sigma^2 = \mathbb{E}[X - \mathbb{E}[X]]^2$. Veamos la aplicación de estos dos conceptos a el caso de una función Bernoulli y, en consecuencia, una Binomial.

Sea X una v.a. que se distribuye como una Bernoulli, entonces:

$$p(x) = \begin{cases} p^k(1-p)^{1-k} & \text{para } k = 0, 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

El valor esperado o meida de una v.a. con pdf Bernoulli estará dada por:

$$\begin{aligned} \mu = \mathbb{E}[X] &= \sum_0^1 k p^k (1-p)^{1-k} \\ &= 0 \cdot p^0(1-p)^1 + 1 \cdot p^1(1-p)^0 \\ &= p \end{aligned}$$

La varianza estará dada por:

$$\begin{aligned} \sigma^2 = \mathbb{E}[(X-\mu)^2] &= \sum_0^1 (k-p)^2 p^k (1-p)^{1-k} \\ &= (0-p)^2 p^0 (1-p)^1 + (1-p)^2 p^1 (1-p)^0 \\ &= p^2(1-p) + (1-p)^2 p \\ &= p^2 - p^3 + p - 2p^2 + p^3 \\ &= -p^2 + p \\ &= p(1-p) \end{aligned}$$

Ahora para el caso de una función Binomial utilizaremos la función generadora de momentos:

Definición 6.9 *Sea X una v.a. tal que para un valor $h > 0$, y suponiendo que el valor esperado de la función e^{tx} existe en un intervalo para t: $-h < t < h$. La función generadora de momentos de X estará definida por la función:*

$$M(t) = \mathbb{E}[e^{tx}] \text{ para } -h < t < h$$

Utilizando la función generadora de momentos podemos determinar los momentos de una v.a. X, mediante la derivada de la función evaluada en el $t = 0$. Por ejmeplo, la media estará dada por $M'(0)$ y el segundo momento por $M''(0)$, continuando de la misma forma para los demás momentos.

Utilizaremos la función generadora de momentos para determinar la esperanza y varianza de una v.a. Binomial dada por:

$$p(X=k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{para } k = 0, 1, \ldots, n \\ 0 & \text{en cualquier otro caso} \end{cases}$$

La función generadora de momentos que nos permite determinar el valor esperado estará dada por:

$$\begin{aligned} M(t) &= \sum_k e^{tk} p(k) \\ &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (1-p)^{n-k} (pe^t)^k \\ &= ((1-p) + pe^t)^n \end{aligned}$$

Para éste último paso utilizamos un resultado del álgebra para los binomios que indica:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} b^k a^{n-k}$$

Así, para determinar la media requerimos de derivar la función generadora de momentos respecto de t y evaluarla en $t = 0$:

$$\begin{aligned} \mu = M'(0) &= n((1-p) + pe^t)^{n-1} pe^t \Big|_{t=0} \\ &= np \end{aligned}$$

Por su parte, la varianza será determinada por la ecuación:

$$\sigma^2 = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2X\mu + \mu^2] = \mathbb{E}[X^2] - \mu^2$$

Es decir, la varianza de una v.a. es igual a su segundo momento de la función de densidad menos su valor esperado al cuadrado. Así, en el caso que nos ocupa estará dada por:

$$\begin{aligned} \sigma^2 &= \mathbb{E}[X^2] - \mu^2 \\ &= M''(0) - \mu^2 \\ &= n((1-p) + pe^t)^{n-1} pe^t + n(n-1)((1-p) + pe^t)^{n-2} - (np)^2 \Big|_{t=0} \\ &= np + n(n-1)p^2 - (np)^2 \\ &= np(1-p) \end{aligned}$$

Comparativamente, los resultados hasta ahora obtenidos son que la media y variaza de una v.a. con distribución Binomial ($\mu = np$ y $\sigma^2 = np(1-p)$) es n veces la media y la varianza de una v.a. con distribución Bernoulli ($\mu = p$ y $\sigma^2 = p(1-p)$).

Retomando nuestro problema panteado, buscamos el valor esperado dado por:

$$\begin{aligned}\mathbb{E}[\hat{p}] &= \mathbb{E}\left[\frac{X + 1/2\sqrt{n}}{n + \sqrt{n}}\right] \\ &= \frac{1}{n + \sqrt{n}}(\mathbb{E}[X] + 1/2\sqrt{n}) \\ &= \frac{np + 1/2\sqrt{n}}{n + \sqrt{n}} \neq p\end{aligned}$$

Por lo tanto, la propuesta de estimado es sesgado, ¿cambia en algo si analizamos si el estimador es asintóticamente insesgado?, es decir:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[\hat{p}] &= \lim_{n\to\infty}\left(\frac{np + 1/2\sqrt{n}}{n + \sqrt{n}}\right) \\ &= p\end{aligned}$$

Veamos otro ejemplo.

Ejemplo. Sea $X \sim Bin(p)$ y un estimado de p dado por:

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

¿Es $\hat{p}$ un estimador insesgado de p?

Tomando su valor esperado:

$$\begin{aligned}\mathbb{E}[\hat{p}] &= \mathbb{E}\left[\frac{X}{n}\right] \\ &= \frac{1}{n}\mathbb{E}[X] \\ &= \frac{1}{n}(np) = p\end{aligned}$$

Por lo tanto, esta propuesta de estimandor es insesgado. Ahora veamos algunos casos de funciones continuas.

Ejemplo. Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con pdf dada por:

$$f(x) = \begin{cases} e^{-(x-\delta)} & \text{para } x > \delta \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Supongamos un estimador de δ dado por $\hat{\delta} = \overline{X}$. Determinemos si el estimador es sesgado.

Previo a ello encontremos el valor esperado de las v.a.'s con esa pdf.

$$\begin{aligned} \mathbb{E}[X] &= \int_{\delta}^{\infty} x \cdot f(x) dx \\ &= \int_{\delta}^{\infty} x e^{-(x-\delta)} dx \\ &= -x e^{-(x-\delta)} + \int_{\delta}^{\infty} e^{-(x-\delta)} dx \\ &= -x e^{-(x-\delta)} - e^{-(x-\delta)}\Big|_{\delta}^{\infty} \\ &= \lim_{x \to \infty} \left(-x e^{-(x-\delta)} - e^{-(x-\delta)}\right) - \left(-\delta e^{0} - e^{0}\right) \\ &= 1 + \delta \end{aligned}$$

La forma de solucionar la integral anterior implicó utilizar una integración por partes:

$$\int u \cdot dv = uv - \int v \cdot du$$

Donde en nuestro caso particular:

$$\begin{aligned} du &= dx \\ u &= x \\ dv &= e^{-(x-\delta)} \\ v &= -e^{-(x-\delta)} \end{aligned}$$

Retomemos el planteamiento del ejemplo:

$$\mathbb{E}[\hat{\delta}] = \mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{\sum_{i=1}^{n} \mathbb{E}[X_i]}{n} = 1 + \delta \neq \delta$$

Ejemplo. Sea una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una pdf dada por la expresión que describimos más abajo. Supongamos un par de estimadores $\hat{\theta} = 3/2 \cdot \overline{X}$ y $\hat{\theta} = X_{max}$ de un parámetro θ de una pdf descrita por:

$$f(x) = \begin{cases} \frac{2x}{\theta^2} & \text{para } 0 \leq x \leq \theta \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos si los estimadores propuestos son insesgados.

Para ello determinemos primero el valor esperado de X:

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\theta x \cdot \frac{2x}{\theta^2} dx \\ &= \int_0^\theta \frac{2x^2}{\theta^2} dx \\ &= \left. \frac{2x^3}{3\theta^2} \right|_0^\theta \\ &= \frac{2}{3}\theta \end{aligned}$$

Empecemos con el fácil:

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}\left[\frac{3}{2}\overline{X}\right] = \frac{3}{2}\mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{3}{2}\left[\frac{2}{3}\theta\right] = \theta$$

Ahora veamos el caso más complicado. Para ello partamos de lo siguiente, la función de distribución estará dada por:

$$F(x) = \int_0^x \frac{2t}{\theta^2} dt = \left. \frac{t^2}{\theta^2} \right|_0^x = \frac{x^2}{\theta^2}$$

Entonces la función de distribución para X_{max} estará dada por:

$$F_{max}(x) = [F(x)]^n = \left(\frac{x^2}{\theta^2}\right)^n$$

Por lo tanto la función de densidad será:

$$\begin{aligned} f_{max}(x) &= \frac{dF(x)}{dx} \\ &= n\left(\frac{x^2}{\theta^2}\right)^{n-1}\frac{2x}{\theta^2} \\ &= n\frac{x^{2(n-1)}}{\theta^{2(n-1)}}\frac{2x}{\theta^2} \\ &= \begin{cases} 2n\frac{x^{2n-1}}{\theta^{2n}} & \text{para } 0 \leq x \leq \theta \\ 0 & \text{en cualquier otro caso} \end{cases} \end{aligned}$$

Ahora si podremos determinar el valor esperado de X_{max}:

$$\begin{aligned} \mathbb{E}[X_{max}] &= \int_0^\theta x \cdot 2n\frac{x^{2n-1}}{\theta^{2n}}dx \\ &= \frac{2n}{\theta^{2n}}\int_0^\theta x^{2n}dx \\ &= \frac{2n}{\theta^{2n}}\cdot\frac{x^{2n+1}}{2n+1}\Bigg|_0^\theta \\ &= \frac{2n}{\theta^{2n}}\cdot\frac{\theta^{2n+1}}{2n+1} \\ &= \frac{2n}{2n+1}\theta \neq \theta \end{aligned}$$

Pero si tomamos el siguiente límite:

$$\begin{aligned} \lim_{n\to\infty} b_n &= \lim_{n\to\infty}\{\mathbb{E}[X_{max}] - \theta\} \\ &= \lim_{n\to\infty}\left(\frac{2n}{2n+1}\theta\right) - \theta = 0 \end{aligned}$$

Ejemplo. Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria, suponga que proponemos en un estimador de $\theta = [X]$ expresado en:

$$\hat{\theta} = \sum_{i=1}^{n} a_i X_i$$

Donde los a_i's son constantes. ¿Para cuáles valores de $a_1, a_2, \ldots, a_n$, $\hat{\theta}$ es insesgado? Analicemos:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left(\sum_{i=1}^{n} a_i X_i\right) \\
&= \sum_{i=1}^{n} a_i \mathbb{E}[X_i] \\
&= \sum_{i=1}^{n} a_i \theta \\
&= \theta \sum_{i=1}^{n} a_i \\
&= \theta
\end{aligned}$$

Donde la condición es que la suma de los ponderadores sea igual a 1, es decir, $\sum_{i=1}^{n} a_i = 1$.

Ejemplo. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una pdf dada por una normal con parámetros μ y σ^2. En este sentido verifiquemos si el siguiente estimador de la varianza es insesgado:

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

Verifiquemos el valor esperado:

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\frac{\sum_{j=1}^{n}X_j}{n} + \left(\frac{\sum_{j=1}^{n}X_j}{n}\right)^2\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}\left(\mathbb{E}[X_i^2] - 2\cdot\mathbb{E}\left[X_i\frac{\sum_{j=1}^{n}X_j}{n}\right] + \mathbb{E}\left[\frac{\sum_{j=1}^{n}X_j}{n}\right]^2\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}\left(\mathbb{E}[X_i^2] - \frac{2}{n}\cdot\mathbb{E}\left[X_i\sum_{j=1}^{n}X_j\right] + \frac{1}{n^2}\cdot\mathbb{E}\left[\sum_{j=1}^{n}X_j\right]^2\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}\left((\sigma^2+\mu^2) - \frac{2}{n}\cdot(n\cdot\mu^2+\sigma^2) + \frac{1}{n^2}\cdot(n^2\cdot\mu^2 + n\cdot\sigma^2)\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}\left(\sigma^2+\mu^2 - 2\cdot\mu^2 - 2\cdot\frac{\sigma^2}{n} + \mu^2 + \frac{\sigma^2}{n}\right)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}\left(\sigma^2 - \frac{\sigma^2}{n}\right)\right] \\
&= \frac{1}{n}(n\cdot\sigma^2 - \sigma^2) \\
&= \frac{(n-1)\cdot\sigma^2}{n}
\end{aligned}$$

Donde utilizamos los siguientes resultados:

$$\begin{aligned}
\mathbb{E}\left[X_i\sum_{j=1}^{n}X_j\right] &= \mathbb{E}[X_iX_1 + \ldots + X_iX_i + \ldots + X_iX_n] \\
&= \mathbb{E}[X_iX_1] + \ldots + \mathbb{E}[X_i^2] + \ldots + \mathbb{E}[X_iX_n] \\
&= \mu^2 + \ldots + \sigma^2 + \mu^2 + \ldots + \mu^2 \\
&= n\cdot\mu^2 + \sigma^2
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}\left[\sum_{j=1}^{n} X_j\right]^2 &= \mathbb{E}[(X_1 + X_2 + \ldots + X_n)(X_1 + X_2 + \ldots + X_n)] \\
&= \mathbb{E}[X_1^2 + X_1X_2 + \ldots + X_1X_n \\
&\quad + X_2X_1 + X_2^2 + X_2X_3 + \ldots X_2X_n \\
&\quad + \ldots \\
&\quad + X_nX_1 + X_nX_2 + \ldots + X_n^2] \\
&= (n \cdot \mu^2 + \sigma^2) + (n \cdot \mu^2 + \sigma^2) + \ldots + (n \cdot \mu^2 + \sigma^2) \\
&= n \cdot (n \cdot \mu^2 + \sigma^2) \\
&= n^2 \cdot \mu^2 + n \cdot \sigma^2
\end{aligned}
$$

Para llegar al resultado anterior utilizamos que si X_i y X_j son v.a.s con pdf normal con parámetros μ y σ^2, entonces:

$$
\begin{aligned}
\sigma^2 &= \mathbb{E}[X_i^2] - \mu^2 \\
\mathbb{E}[X_i^2] &= \sigma^2 + \mu^2 \\
Cov[X_i, X_j] &= \mathbb{E}[X_iX_j] - \mu^2 = 0 \\
\mathbb{E}[X_iX_j] &= \mu^2
\end{aligned}
$$

Dicho lo anterior, entonces podemos concluir que el estimador $\hat{\sigma}^2$ es sesgado, pero podríamos revisar si dicho sesgo se sostiene en un escenario asintótico:

$$
\lim_{n\to\infty} \mathbb{E}[\hat{\sigma}^2] = \lim_{n\to\infty} \left[\frac{(n-1)\cdot\sigma^2}{n}\right] = \sigma^2
$$

Pero en el fondo quisieramos que el estimador fuera insesgado sin importar el tamaño de la muestra, por ello podríamos ajustar el estimador $\hat{\sigma}^2$ de forma tal que sea:

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

De esta forma:

$$
\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] \\
&= \frac{1}{n-1}(n-1)\cdot\sigma^2 \\
&= \sigma^2
\end{aligned}
$$

En general este nuevo estimador se conoce como el estimador insesgado de la varianza.

6.2.2. Eficiencia

Una vez que hemos estudiado y ejemplificado la propiedad de insesgamiento, ahora discutiremos qué hacer cuando tenemos una familia o grupo de estimadores, todos insesgados, ¿cómo seleccionar alguno de entre el grupo?, ¿qué criterio utilizar para seleccionarlo? De eso se trata el análisis de eficiencia de los estimadores.

Empecemos enunciando algunas definiciones.

Definición 6.10 ***Estimador eficiente****. Decimoes que un estimador es eficiente si tiene un error estándar que es el más pequeño de aquellos estimadores posibles.*

Otra forma de decirlo es que un estimador eficiente tiende a 'caer' más cercano al paramétro que tratemos de estimar. En suma, un 'buen' estimador de un paramétro es insesgado, o es cercano a serlo, y eficiente.

Definición 6.11 ***Estimador insesgado de mínima varianza****. El estimdor $\hat{\theta}$ del parámetro θ de una pdf dada que tiene la varianza mínima de entre la familia de estimadores insesgados Θ se dice que es el mejor estimador insesgado.*

Definición 6.12 *Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados del parámetro θ si:*

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2) \tag{6.3}$$

Decimos que $\hat{\theta}_1$ es un estimador más eficiente que $\hat{\theta}_2$.

Algunos teoremas relevantes.

Teorema 6.1 *Sea $\hat{\theta}$ un estimador insesgado de un parámetro θ de una pdf dada que cumple con las condiciones de regularidad, si dicho estimador cumple con:*

$$Var(\hat{\theta}) = \left[n \cdot \mathbb{E} \left[\left(\frac{\partial ln(f(x))}{\partial \theta} \right)^2 \right] \right]^{-1} \tag{6.4}$$

Entonces, decimos que $\hat{\theta}$ es el estimador insesgado de mínima varianza de θ.

Definición 6.13 ***Condiciones de regularidad***. *Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una función de densidad $f(\cdot;\theta)$ donde $\theta \in \Theta$. Suponga-mos que Θ es un subconjunto de la recta real y seea $\hat{\theta}$ un estimador insesgado de θ. Dicho esto, diremos que las siguientes son las condicioness de regula-ridad para $f(\cdot;\theta)$:*

$$i)\ \frac{\partial}{\partial\theta} ln(f(x;\theta))\ existe\ \forall x, \forall\theta$$

$$ii)\ \frac{\partial}{\partial\theta}\int\cdots\int \Pi_{i=1}^{n} f(x_i;\theta)dx_1\cdots dx_n = \int\cdots\int \frac{\partial}{\partial\theta}\Pi_{i=1}^{n} f(x_i;\theta)dx_1\cdots dx_n$$

$$\begin{aligned} iii)\ &\frac{\partial}{\partial\theta}\int\cdots\int h(x_1,\ldots,x_n)\Pi_{i=1}^{n} f(x_i;\theta)dx_1\cdots dx_n \\ &= \int\cdots\int h(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}\Pi_{i=1}^{n} f(x_i;\theta)dx_1\cdots dx_n \end{aligned}$$

$$iv)\ 0 < \mathbb{E}\left[\left[\frac{\partial}{\partial\theta} ln(f(x_i;\theta))\right]^2\right] < \infty\ ,\ \forall\theta\in\Theta$$

Ahora uno de los teoremas más relevantes para el análisis estadístico.

Teorema 6.2 ***Desigualdad de Cramér-Rao***. *Sea f(x) una pdf, suponga-mos que tenemos un conjunto de observaviones de x para las que $f(x) \neq 0$, que no depende de θ.*

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población de una $f(x)$ y sea $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ un estimador insesgado de θ. Entonces,

$$\begin{aligned} Var(\hat{\theta}) &\geq \left[n\cdot\mathbb{E}\left[\left(\frac{\partial ln(f(x))}{\partial\theta}\right)^2\right]\right]^{-1} && (6.5) \\ &= \left[-n\cdot\mathbb{E}\left[\left(\frac{\partial^2 ln(f(x))}{\partial\theta^2}\right)\right]\right]^{-1} && (6.6) \end{aligned}$$

A la expresión contenida en el términno de la esperanza en ambos teore-mas se le conoce como el término que contiene la información sobre θ que es aportada por la muestra. Así, si este término contiene toda la información disponible en la muestra, entonces la varianza será mínima.

En los siguientes ejemplos ilustraremos el uso de estas definiciones y teoremas. particularmente mostraremos cómo es cierta la igualdad en el teorema de la Desigualdad de Cramér-Rao. Partamos del siguiente ejemplo.

Ejemplo. Sean X_1, X_2, X_3 una muestra aleatoria de una población con distribución normal con parámetros conocidos μ y σ^2. ¿Cuál de los siguientes estimadores es un estimador más eficiente para μ?

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3 \\ \hat{\mu}_2 &= \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\end{aligned}$$

Para responder a la pregunta debemos verificar si ambos son insesgados, y es más o menos claro que lo son ya que:

$$\mathbb{E}[\hat{\mu}_1] = \mathbb{E}[\hat{\mu}_2] = \mu \tag{6.7}$$

Dicho esto, debemos determinar la varianza de cada uno de ellos.

$$\begin{aligned}Var(\hat{\mu}_1) &= Var\left(\frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3\right) \\ &= Var\left(\frac{1}{4}X_1\right) + Var\left(\frac{1}{2}X_2\right) + Var\left(\frac{1}{4}X_3\right) \\ &= \frac{1}{16}Var(X_1) + \frac{1}{4}Var(X_2) + \frac{1}{16}Var(X_3) \\ &= \frac{1}{16}\sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{16}\sigma^2 \\ &= \frac{3}{8}\sigma^2\end{aligned}$$

$$\begin{aligned}Var(\hat{\mu}_1) &= Var\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) \\ &= Var\left(\frac{1}{3}X_1\right) + Var\left(\frac{1}{3}X_2\right) + Var\left(\frac{1}{3}X_3\right) \\ &= \frac{1}{9}Var(X_1) + \frac{1}{9}Var(X_2) + \frac{1}{9}Var(X_3) \\ &= \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 \\ &= \frac{3}{9}\sigma^2\end{aligned}$$

Por lo tanto el primer estimador parece ser el más eficiente, es decir, el mejor de entre los que tenemos.

Ejemplo. Sean $X_1, X_2, \ldots, X_n$ una nuestra aleatoria de una población dada y $\hat{\mu} = \overline{X}$ el estimador de la media de la pdf de donde se realizó el muestreo que es descrita por:

$$f(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} & \text{para } -\infty < x < \infty \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Entonces, ¿es $\hat{\mu}$ un estimador de mínima varianza?

Para poder responder tendríamos que atender a (1) si el estimador es insesgado y (2) determinar si alcanza la cota de Cramér-Rao. Respecto del primero es fácil observar que: $\mathbb{E}[\overline{X}] = \mu$. Respecto del segundo implica resolver:

$$\begin{aligned} Var(\hat{\mu}) &= Var(\overline{X}) \\ &= \mathbb{E}[\overline{X}^2] - \mathbb{E}[\overline{X}]^2 \\ &= \frac{1}{n^2} \cdot \mathbb{E}[(X_1 + X_2 + \ldots + X_n)^2] - \mu^2 \\ &= \frac{1}{n^2} \cdot \mathbb{E}[(X_1 + X_2 + \ldots + X_n)(X_1 + X_2 + \ldots + X_n)] - \mu^2 \\ &= \frac{1}{n^2} \cdot \mathbb{E}[X_1^1 + X_1X_2 + \ldots + X_1X_n + X_2X_1 + X_2^2 + \ldots + X_2X_n \\ &\quad + \ldots + X_nX_1 + X_nX_2 + \ldots + X_n^2] - \mu^2 \\ &= \frac{1}{n^2}(\sigma^2 + n \cdot \mu 2 + \sigma^2 + n \cdot \mu 2 + \ldots \sigma^2 + n \cdot \mu 2) - \mu 2 \\ &= \frac{1}{n^2}(n \cdot \sigma^2 + n^2 \cdot \mu 2) - \mu 2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Por otro lado, tomemos el logaritmo de la pdf antes descrita y obteniendo su primer y segunda derivadas:

$$\begin{aligned} lnf(x) &= -\ln(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \\ \frac{\partial lnf(x)}{\partial \mu} &= \frac{1}{\sigma^2}(x - \mu) \\ \frac{\partial^2 lnf(x)}{\partial \mu^2} &= -\frac{1}{\sigma^2} \end{aligned}$$

Entonces, tomando los dos teoremas anteriores:

$$\begin{aligned}
\left[n \cdot \mathbb{E}\left[\left(\frac{\partial ln f(x)}{\partial \mu}\right)^2\right]\right]^{-1} &= \left[n \cdot \mathbb{E}\left[\left(\frac{1}{\sigma^2}(X-\mu)\right)^2\right]\right]^{-1} \\
&= \left[\frac{n}{\sigma^4} \cdot \mathbb{E}\left[(X-\mu)^2\right]\right]^{-1} \\
&= \left[\frac{n}{\sigma^4} \cdot \mathbb{E}\left[X^2 - 2X\mu + \mu^2\right]\right]^{-1} \\
&= \left[\frac{n}{\sigma^4} \cdot (\sigma^2 + \mu^2 - 2\mu^2 + \mu^2)\right]^{-1} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Por su parte:

$$\begin{aligned}
\left[-n \cdot \mathbb{E}\left[\left(\frac{\partial^2 ln f(x)}{\partial \mu^2}\right)^2\right]\right]^{-1} &= \left[-n \cdot \mathbb{E}\left[\left(-\frac{1}{\sigma^2}\right)\right]\right]^{-1} \\
&= \left[n \cdot \frac{1}{\sigma^2}\right]^{-1} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Por lo tanto, el estimador $\hat{\mu}$ propuesto es el mejor estimador insesgado de μ, ya que éste alcaza la cota inferior de Cramér-Rao:

$$\begin{aligned}
Var(\hat{\mu}) &= \frac{\sigma^2}{n} \\
&= \left[n \cdot \mathbb{E}\left[\left(\frac{\partial ln f(x)}{\partial \mu}\right)^2\right]\right]^{-1} \\
&= \left[-n \cdot \mathbb{E}\left[\left(\frac{\partial^2 ln f(x)}{\partial \mu^2}\right)^2\right]\right]^{-1}
\end{aligned}$$

Ejemplo. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de experimentos de Bernoulli, es decir, toman sólo valores de 0 y 1. Entonces cada una de las v.a.'s cumple con:

$$p_{X_i}(k) = \begin{cases} p^k(1-p)^{1-k} & \text{para } k = 0, 1 \text{ y } 0 < p < 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Sea $X = X_1 + X_2 + \ldots + X_n$ el número total de éxitos, definamos a un estimador de p como $\hat{p} = \frac{X}{n}$. ¿Es este estimador eficiente?

Para probarlo requerimos de verificar que sea un estimador insesgado y determinar su varianza:

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{X}{n}\right] = \frac{1}{n}\mathbb{E}\left[X_1 + X_2 + \ldots + X_n\right] = \frac{1}{n}\mathbb{E}\left[p + p + \ldots + p\right] = p$$

Para resolver la varianza, recordemos que para cualquie X_i con distribución Bernoulli:

$$Var(X_i) = p(1-p) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \mathbb{E}[X_i^2] - p^2$$

Por lo tanto:

$$\mathbb{E}[X_i^2] = p(1-p) + p^2$$

Tambien:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] = p^2$$

Así:

$$\begin{aligned}
Var(\hat{p}) &= Var\left[\frac{X}{n}\right] \\
&= Var\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] \\
&= \frac{1}{n^2}Var(X_1 + X_2 + \ldots + X_n) \\
&= \frac{1}{n^2}(\mathbb{E}[(X_1 + X_2 + \ldots + X_n)^2] - (\mathbb{E}[X_1 + X_2 + \ldots + X_n])^2) \\
&= \frac{1}{n^2}(\mathbb{E}[X_1^2 + X_1X_2 + \ldots + X_1X_n + \ldots + X_nX_1 + X_nX_2 + \\
&\quad \ldots + X_n^2] - (\mathbb{E}[p + p + \ldots + p])^2) \\
&= \frac{1}{n^2}((p(1-p) + p^2) + p^2 + \ldots + p^2 + \ldots + p^2 + \ldots \\
&\quad +(p(1-p) + p^2) - n^2p^2) \\
&= \frac{1}{n^2}(np(1-p) + n^2p^2 - n^2p^2) \\
&= \frac{p(1-p)}{n}
\end{aligned}$$

Dicho esto, podremos probar si la varianza del estimado propuesto alcanza la cota inferior de Cramér-Rao, partamos del logaritmo de la función y sus primera y segunda derivadas:

$$\begin{aligned} lnp_{X_i}(k) &= k \cdot ln(p) + (1-k) \cdot ln(1-p) \\ \frac{\partial lnp_{X_i}(k)}{\partial p} &= \frac{k}{p} - \frac{1-k}{1-p} \\ \frac{\partial^2 lnp_{X_i}(k)}{\partial p^2} &= -\frac{k}{p^2} - \frac{1-k}{(1-p)^2} \end{aligned}$$

Tomando el valor esperado de la segunda derivada obtenemos las siguiente expresión:

$$\begin{aligned} \left[-n \cdot \mathbb{E}\left[\frac{\partial^2 lnp_{X_i}(k)}{\partial p^2}\right]\right]^{-1} &= \left[-n \cdot \mathbb{E}\left[-\frac{X_i}{p^2} - \frac{1-X_i}{(1-p)^2}\right]\right]^{-1} \\ &= \left[-n \cdot \left[-\frac{p}{p^2} - \frac{1-p}{(1-p)^2}\right]\right]^{-1} \\ &= \left[-n \cdot \left[-\frac{1}{p(1-p)}\right]\right]^{-1} \\ &= \frac{p(1-p)}{n} \end{aligned}$$

Por lo tanto, el estimador $\hat{p}$ es eficiente:

$$\begin{aligned} Var(\hat{p}) &= \frac{p(1-p)}{n} \\ &= \left[-n \cdot \mathbb{E}\left[\frac{\partial^2 lnp_{X_i}(k)}{\partial p^2}\right]\right]^{-1} \end{aligned}$$

6.2.3. Consistencia

Hasta ahora hemos asumido que los datos son una muestra fija de tamaño n, salvo en caso en que hemos analizado las condiciones asintóticas. Por ejemplo, recordemos que el estimador la varianza de una normal planteado fue:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n}$$

Del cual demostramos que:

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Pero podemos ver que es asintóticamente insesgado, puesto que a medida que $n \to \infty$ es posible determinar que el estimador es insesgado.

En esta sección introduciremos la propiedad conocida como consistencia. A diferencia del insesgamiento asintótico. La propiedad de consistencia refiere a la forma de $\hat{\theta}_n$ y como ésta cambia como función de n.

Las propiedades anteriores y la siguiente analizamos casos en los que la muestra está dada, por lo cual revisamos que sucede a los estimadores cuando el tamaño de la muestra se incrementa, es decir cuando $n \to \infty$. De esta forma construimos el concepto de cercanía en la siguiente definición:

Definición 6.14 ***Estimador consistente***. *El estimador o la estadística $\hat{\theta}_n$ $(= h(X_1, X_2, \ldots, X_n))$ es un estimador consistente del parámetro θ de una pdf dada, si y sólo si, para cada $\varepsilon > 0$, entonces:*

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$$

Dicho lo anterior, notemos que consistencia es una propiedad asintótica, es decir, una propiedad límite de un estimador. Así, decimos que cuando n es suficientemente grande, podemos decir con certeza que el error de estimación es menor a una constante positiva. Veamos un ejemplo.

Ejemplo. Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una pdf uniforme dada por:

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{para } 0 \leq x \leq \theta \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Sea $\hat{\theta}_n = X_{max}$, entonces ¿es X_{max} un estimador insesgado para θ? y ¿es consistente?

Para responder a las preguntas encontremos la función de densidad de X_{max}, su valor esperado y verificamos la consistencia.

$$F(x) = \int_D \frac{1}{\theta} dt = \int_0^x \frac{1}{\theta} dt = \left.\frac{t}{\theta}\right|_0^x = \frac{x}{\theta}$$

Por lo tanto, la función de distribución de probabilidad acumulada conjunta de para la muestra de tamaño n, será:

$$F_{max}(x) = \begin{cases} \frac{x^n}{\theta^n} & \text{para } 0 \leq x \leq \theta \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Así, la función de densidad de probabilidad acumulada será:

$$\frac{dF_{max}(x)}{dx} = \frac{nx^{n-1}}{\theta^n} = f_{max}(x)$$

$$f_{max}(x) = \begin{cases} \frac{nx^{n-1}}{\theta^n} & \text{para } 0 \leq x \leq \theta \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos el valor esperado de $\hat{\theta}_n = X_{max}$ por la expresión:

$$\begin{aligned} \mathbb{E}[\hat{\theta}_n] &= \mathbb{E}[X_{max}] \\ &= \int_0^\theta x \cdot f(x) dx \\ &= \int_0^\theta x \frac{nx^{n-1}}{\theta^n} dx \\ &= \int_0^\theta n \frac{x^n}{\theta^n} dx \\ &= \frac{n}{n+1} \frac{x^{n+1}}{\theta^n} \Big|_0^\theta \\ &= \frac{n}{n+1}\theta \neq \theta \end{aligned}$$

Por lo tanto, el estimador es sesgado, pero veamos el caso de consistencia, para lo cual partamos:

$$\begin{aligned} P(|\hat{\theta}_n - \theta| < \varepsilon) &= P(\theta - \varepsilon < \hat{\theta}_n < \theta + \varepsilon) \\ &= P(\theta - \varepsilon < \hat{\theta}_n < \theta) \\ &= \int_{\theta-\varepsilon}^\theta n \frac{x^{n-1}}{\theta^n} dx \\ &= \frac{x^n}{\theta^n} \Big|_{\theta-\varepsilon}^\theta \\ &= 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n \end{aligned}$$

Por lo tanto, comprobando la definición:

$$\begin{aligned} \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) &= \lim_{n\to\infty} \left(1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n\right) \\ &= 1 \end{aligned}$$

Hemos verificado que la estadística $\hat{\theta}_n = X_{max}$ es consistente. Más aún podemos establecer la probabilidad de que una v.a. se ubique al rededor de una vecindad de la media. Esto nos da pie a establecer un teorema conocido como la *Desigualdad de Chebyshev.*

Teorema 6.3 ***Desigualdad de Chebyshev****. Sea X una v.a. cualquiera con una pdf con parámetros de la media μ y la varianza σ^2. Así, para cualquier $\varepsilon > 0$:*

$$P(|X - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}$$

O equivalentemente:

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Ahora demostremos el teorema para el caso continuo y asumamos que el caso discreto es similar pero con el uso de sumas. Empecemos por retomar que: $\sigma^2 = Var(X)$ y que

$$\begin{aligned} P(|X - \mu| < \varepsilon) &= P(-\varepsilon < X - \mu < \varepsilon) \\ &= P(\mu - \varepsilon < X < \mu + \varepsilon) \end{aligned}$$

Usando los resultados anteriores:

$$\begin{aligned}
Var(X) &= \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx \\
&= \int_{-\infty}^{\mu-\varepsilon}(x-\mu)^2 f(x)dx + \int_{\mu-\varepsilon}^{\mu+\varepsilon}(x-\mu)^2 f(x)dx \\
&\quad + \int_{\mu+\varepsilon}^{\infty}(x-\mu)^2 f(x)dx \\
&\geq \int_{-\infty}^{\mu-\varepsilon}(x-\mu)^2 f(x)dx + \int_{\mu+\varepsilon}^{\infty}(x-\mu)^2 f(x)dx \\
&\geq \int_{-\infty}^{-\varepsilon}(x-\mu)^2 f(x)dx + \int_{\varepsilon}^{\infty}(x-\mu)^2 f(x)dx \\
&\geq \int_{|x-\mu|\geq\varepsilon}(x-\mu)^2 f(x)dx \\
&\geq \int_{|x-\mu|\geq\varepsilon}\varepsilon^2 f(x)dx \\
&= \varepsilon^2 \int_{|x-\mu|\geq\varepsilon} f(x)dx \\
&= \varepsilon^2 P(|X-\mu|\geq\varepsilon)
\end{aligned}$$

De esta forma:

$$\begin{aligned}
\varepsilon^2 P(|X-\mu|\geq\varepsilon) &\leq \sigma^2 \\
P(|X-\mu|\geq\varepsilon) &\leq \frac{\sigma^2}{\varepsilon^2} \\
P(|X-\mu|<\varepsilon) &\geq 1-\frac{\sigma^2}{\varepsilon^2}
\end{aligned}$$

¿Qué significa esto?, pensemos en un caso en que $\varepsilon = 2\sigma$, en cuyo caso:

$$\begin{aligned}
P(|X-\mu|<2\sigma) &\geq 1-\frac{\sigma^2}{(2\sigma)^2} \\
&= 0.75
\end{aligned}$$

Ejemplo. Supongammos $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una pdf discreta tal que $\mathbb{E}[X] = \mu$ y que $\sigma^2 = Var(X) < \infty$, sea:

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

¿Es $\hat{\mu}_n$ un estimador consistente?

Partamos de que la desigualdad de Chebyshev, pero previo a resolver el problema, veamos la siguiente propiedad de la suma de la varianza de la suma de variables aleatorias.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria, y sea $X = \sum_{i=1}^{n} X_i$, entonces:

$$\begin{aligned}
Var(X) &= Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)^2 \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\sum_{i=1}^{n} \mathbb{E}[X_i]\right)^2 \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - n^2\mu^2 \\
&= \mathbb{E}\left[(X_1 + X_2 + \ldots X_n)(X_1 + X_2 + \ldots X_n)\right] - n^2\mu^2 \\
&= \mathbb{E}[X_1^2 + X_1X_2 + \ldots + X_1X_n + \ldots + \\
&\quad X_nX_1 + X_nX_2 + \ldots + X_n^2] - n^2\mu^2 \\
&= \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \ldots \mathbb{E}[X_n^2] + \mathbb{E}[X_1X_2] + \ldots + \mathbb{E}[X_1X_n] \\
&\quad + \ldots + \mathbb{E}[X_{n-1}X_n] - n^2\mu^2 \\
&= \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \ldots \mathbb{E}[X_n^2] + \mu^2 + \ldots + \mu^2 + \ldots + \mu^2 - n^2\mu^2 \\
&= \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \ldots \mathbb{E}[X_n^2] + n(n-1)\mu^2 - n^2\mu^2 \\
&= \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \ldots \mathbb{E}[X_n^2] - n\mu^2 \\
&= \mathbb{E}[X_1^2] - \mu^2 + \mathbb{E}[X_2^2] - \mu^2 + \ldots \mathbb{E}[X_n^2] - \mu^2 \\
&= n\sigma^2
\end{aligned}$$

Para la demostración anterior hemos utilizado que:

$$Cov(X_i, X_j) = \mathbb{E}[X_iX_j] - \mu^2 = 0$$

Por lo tanto: $\mathbb{E}[X_iX_j] = \mu^2$, para todo $i \neq j$.

Retomando nuestro problema original:

$$\begin{aligned} P(|\hat{\mu}_n - \mu| < \varepsilon) &\geq 1 - \frac{Var(\hat{\mu}_n)}{\varepsilon^2} \\ &= 1 - \frac{Var(\frac{1}{n}\sum_{i=1}^{n} X_i)}{\varepsilon^2} \\ &= 1 - \frac{\frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)}{\varepsilon^2} \\ &= 1 - \frac{\sigma^2}{n\varepsilon^2} \end{aligned}$$

Por lo que:

$$\begin{aligned} P(|\hat{\mu}_n - \mu| < \varepsilon) &\geq 1 - \frac{\sigma^2}{n\varepsilon^2} \\ \lim_{n\to\infty} P(|\hat{\mu}_n - \mu| < \varepsilon) &= 1 \end{aligned}$$

De forma relacionada podemos eununciar el siguiente teorema:

Teorema 6.4 *Si $\hat{\theta}_n$ es un estimador insesgado de θ y $Var(\hat{\theta}_n) \to 0$, cuando $n \to \infty$, entonces, $\hat{\theta}_n$ es un estimador consistente de θ*

Ejemplo. En otros casos hemos plateado como estimador de la media de una función de densidad de probabilidad de una normal a:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Como ya mostramos en otros ejemplos, sabemos que este estimador es insesgado de la media de una normal. Entonces, ¿es $\hat{\mu}$ un estimados consistente de μ?

Hemos mostrado que:

$$Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

Por lo tanto

$$Var(\hat{\mu}) \to 0, \text{ cuando } n \to \infty$$

6.2.4. Suficiencia

Un estimador $\hat{\theta}$ se dice que es suficiente si esté utiliza toda la información relevante en una muestra para estimar a θ.

Podemos describir esta propiedad de un estimador como la probabilidad condicional de obtener la muestra aleatoria de tamaño n dado que tenemos al estimador $\hat{\theta}$ es:

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\hat{\theta}) &= \frac{f(x_1, x_2, \ldots, x_n, \hat{\theta})}{g(\hat{\theta})} \\ &= b(x_1, x_2, \ldots, x_n) \end{aligned}$$

Definición 6.15 ***Estimador Suficiente****. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una pdf, la estadística o estimador $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ es un estimador suficiente del parámetro θ de una función de densidad de probabilidad dada si y sólo si la distribución de probabilidad conjunta condicional de una muestra aleatoria dado $\hat{\theta} = \theta_e$ es independiente de θ.*

Ejemplo. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una función de densidad de probabilidad Poisson:

$$p(X_i = k_i) = \begin{cases} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} & \text{para } k_i = 0, 1, \ldots \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Sea el siguiente un estimador de λ:

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

Entonces, ¿es $\hat{\lambda}$ un estimador suficiente de λ?

Partamos de:

$$\begin{aligned}
p(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) &= \Pi_{i=1}^{n} e^{-\lambda} \frac{\lambda^{k_i}}{k_i!} \\
&= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} k_i}}{\Pi_{i=1}^{n} k_i!} \\
&= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} k_i}}{\Pi_{i=1}^{n} k_i!} \cdot \frac{n^{\sum_{i=1}^{n} k_i} (\sum_{i=1}^{n} k_i)!}{n^{\sum_{i=1}^{n} k_i} (\sum_{i=1}^{n} k_i)!} \\
&= e^{-n\lambda} \frac{(n\lambda)^{\sum_{i=1}^{n} k_i}}{(\sum_{i=1}^{n} k_i)!} \cdot \frac{(\sum_{i=1}^{n} k_i)!}{\Pi_{i=1}^{n} k_i! \cdot n^{\sum_{i=1}^{n} k_i}} \\
&= e^{-\lambda} \frac{(n\lambda)^{\hat{\lambda}}}{\hat{\lambda}!} \cdot \frac{(\sum_{i=1}^{n} k_i)!}{\Pi_{i=1}^{n} k_i! \cdot n^{\sum_{i=1}^{n} k_i}} \\
&= g(\hat{\lambda}) \cdot b(k_1, k_2, \ldots, k_n)
\end{aligned}$$

Ejemplo. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una pdf Bernoulli, entonces suponga el estimador:

$$\hat{\theta} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

¿Es ese estimador suficiente para estimar a θ?
Recordemos que una v.a. X_i tiene distribución Bernoulli si:

$$p(X = k_i) = \begin{cases} \theta^{k_i}(1-\theta)^{1-k_i} & \text{para } k_i = 0, 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Entonces la función de densidad conjunta es

$$\begin{aligned}
p(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) &= \Pi_{i=1}^{n} \theta^{k_i}(1-\theta)^{1-k_i} \\
&= \theta^{\sum_{i=1}^{n} k_i}(1-\theta)^{n-\sum_{i=1}^{n} k_i} \\
&= \theta^{k}(1-\theta)^{n-k} \\
&= \theta^{n\hat{\theta}}(1-\theta)^{n-n\hat{\theta}}
\end{aligned}$$

Donde asummimos que, como demostraremos abajo, $X = X_1 + X_2 + \ldots + X_n$ se distribulle como una Binomial:

$$p(X = k) = \binom{n}{k} \theta^{k}(1-\theta)^{n-k}$$

Donde $k = 0, 1, \ldots, n$.
La v.a. X es equivalente a resolver:

$$p(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) = \theta^{\sum_{i=1}^{n} k_i}(1-\theta)^{n-\sum_{i=1}^{n} k_i}$$

De esta forma, hemos construido una función de $p(X = k)$, para $0 \leq k \leq n$. Lo cual implica que calcular el número de veces y formas en que podemos obtener k, por lo tanto:

$$p(X = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Así:

$$g(\hat{\theta}) = \binom{n}{n\hat{\theta}}\theta^{n\hat{\theta}}(1-\theta)^{n-n\hat{\theta}}$$

Sustituyendo los resultados:

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\hat{\theta}) &= p(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n|\hat{\theta}) \\ &= \frac{\theta^{n\hat{\theta}}(1-\theta)^{n-n\hat{\theta}}}{\binom{n}{n\hat{\theta}}\theta^{n\hat{\theta}}(1-\theta)^{n-n\hat{\theta}}} \\ &= \frac{1}{\binom{n}{n\hat{\theta}}} \\ &= \frac{1}{\binom{n}{k}} \end{aligned}$$

Por lo tanto, es estimador $\hat{\theta}$ es suficiente.

Teorema 6.5 *La Estadística $\hat{\theta}$ es un estimador suficiente del parámetro θ si y sólo si la función de probabilidad conjunta o densisdad conjunta de la muestra aleatoria dada puede ser factorizada como:*

$$f(x_1, x_2, \ldots, x_n) = g(\hat{\theta}) \cdot b(x_1, x_2, \ldots, x_n)$$

Donde $b(x_1, x_2, \ldots, x_n)$ no depende de θ.

Ejemplo. Sea $\overline{X}$ un estimador de la media μ de una pdf norma con varianza σ^2. ¿Es el estimador suficiente? Partamos de una pdf para cualquier x_i:

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

Por lo tanto, la función de densidad de probabilidad conjunta será:

$$f(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

Desarrollemos:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}((x_i-\overline{X})-(\mu-\overline{X}))^2 \\ &= \sum_{i=1}^{n}(x_i-\overline{X})^2+\sum_{i=1}^{n}(\overline{X}-\mu)^2\end{aligned}$$

Utilizando el resultado anterior, podemos construir

$$f(x_1, x_2, \ldots, x_n) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2} \cdot \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-1} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\overline{X}}{\sigma}\right)^2}$$

6.3. Métodos de estimación: Método de momentos y Método de máxima verosimilitud

Existen diferentes estimadores de un mismo parámetro poblacional. Por ello existe la necesidad de un método general o métodos generales que nos permitan construir o plantear estimadores.

En esta sección vamos a analizar los dos métodos más populares para estimación de parámetros desconocidos: el método de momentos y el método de máxima verosimilitud. Existen otros métodos de estimación como el mínima distancia, o el de mínimos cuadrados en el caso de análisis de regresión, etc.

En los métodos empleados para determinar estimadores asumiremos que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una población con función de densidad de probabilidad (pdf, por sus siglas en inglés) dada por $f(\cdot;\theta)$, donde la la función de densidad de probabilidad es conocida y el elemento desconocido es el parámetro θ.

También asumiremos que $\theta \in \mathbb{R}$ o un vector de reales $\boldsymbol{\theta}$, donde $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ y $\boldsymbol{\theta} \in \mathbb{R}^n$. Denotemos a Θ como el espacio de parámetros, que es el conjunto de los posibles valores que puede tomar θ.

6.3.1. Método de Momentos

El método de momentos es el método más sencillo de resolver cuando enfrentamos el problema de platear múltiples estimadores de los parámetros. Este método genera estimadores que no son óptimos pero son facílmente computables.

El método de momentos consiste en igualar los primeros momentos de una población a los correspondientes momentos muestrales. Así podríamos obtener un sistema de tantas ecuaciones como parámetos a estimar tengamos. Tomemos como definición de momentos muestrales a:

Definición 6.16 ***Momentos Muestrales****. El k-ésimo momento muestral de un conjunto de observaciones* $\{x_1, x_2, \ldots, x_n\}$ *es la media de su k-ésimas potencias:*

$$m'_k = \frac{\sum_{i=1}^{n} x_i^k}{n}$$

Para $k = 1, 2, \ldots, K$*.*

Tomemos como definición de momentos poblacionales a:

Definición 6.17 ***Momentos Poblacionales****. El k-ésimo momento poblacional de una v.a. con pdf dada:*

$$\mathbb{E}[X^k] = \int_{-\infty}^{\infty} x^k \cdot f(x; \theta_1, \theta_2, \ldots, \theta_K) dx \text{ para } k = 1, 2, \ldots, K$$

Donde cada $\mathbb{E}[X^k]$ *será una función de* $\{\theta_1, \theta_2, \ldots, \theta_K\}$*.*

En el caso de una pdf discreta, los momentos poblacionales resultan de la suma.

El método de momentos consiste en el siguiente algoritmo. Supongamos que X es una variable aleatoria de una pdf dada con parámetros desconocidos $\{\theta_1, \theta_2, \ldots, \theta_K\} = \boldsymbol{\theta}$. Así, si la población tiene K parámetros, el método de momentos consiste en resolver para cada una de las incógnitas $\{\theta_1, \theta_2, \ldots, \theta_K\}$ el sistema de ecuaciones:

$$\begin{aligned} m'_1 &= \mathbb{E}[X^1] = g_1(\theta_1, \theta_2, \ldots, \theta_K) \\ m'_2 &= \mathbb{E}[X^2] = g_2(\theta_1, \theta_2, \ldots, \theta_K) \\ &\vdots \\ m'_K &= \mathbb{E}[X^K] = g_K(\theta_1, \theta_2, \ldots, \theta_K) \end{aligned}$$

Ejemplo. Sea $\{x_1, x_2, \ldots, x_n\}$ una muestra aleatoria de una población con pdf dada por:

$$f(x) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determine una expresión para un estimador de θ mediante el método de momentos.

Para este caso partamos de:

$$\mathbb{E}[X] = \int_0^1 x \cdot f(x)dx = \int_0^1 x \cdot \theta x^{\theta-1} dx = \left.\frac{\theta x^{\theta+1}}{\theta+1}\right|_0^1 = \frac{\theta}{\theta+1}$$

De esta forma construimos un sistema de una sola ecuación:

$$\frac{\sum_{i=1}^n x_i}{n} = \frac{\theta}{\theta+1}$$

La solución entonces estará dada por el estimador por el método de momentos:

$$\hat{\theta} = \frac{\frac{\sum_{i=1}^n x_i}{n}}{1 - \frac{\sum_{i=1}^n x_i}{n}}$$

Ejemplo. Sea $\{x_1, x_2, \ldots, x_n\}$ una muestra aleatoria de una población con pdf del tipo uniforme dada por:

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Use el método de momentos para obtener los parámetros a y b.

Para ello primero debemos determinar los dos primeros momentos poblacionales para una X cualquiera:

$$\begin{aligned} \mathbb{E}[X] &= \int_a^b x \cdot f(x)dx = \int_a^b x \cdot \frac{1}{b-a} dx = \left.\frac{x^2}{2(b-a)}\right|_a^b \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X^2] &= \int_a^b x^2 \cdot f(x)dx = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \left.\frac{x^3}{3(b-a)}\right|_a^b \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{(a^2+ab+b^2)(b-a)}{3(b-a)} = \frac{a^2+ab+b^2}{3} \end{aligned}$$

Entonces igualando los momentos poblacionales a los momentos muestrales obtebemos el siguiente sistema de ecuaciones:

$$\begin{aligned}\frac{\sum_{i=1}^{n} x_i}{n} &= \frac{b+a}{2} \\ \frac{\sum_{i=1}^{n} x_i^2}{n} &= \frac{a^2+ab+b^2}{3}\end{aligned}$$

Resolviendo el sistema de ecuaciones obtenemos el siguiente procedimiento:

$$b = 2 \cdot \frac{\sum_{i=1}^{n} x_i}{n} - a$$

Sustituyendo la expresión anterior en la segunda ecuación obtenemos:

$$\begin{aligned}\frac{\sum_{i=1}^{n} x_i^2}{n} &= \frac{a^2 + a(2\frac{\sum_{i=1}^{n} x_i}{n} - a) + (2\frac{\sum_{i=1}^{n} x_i}{n} - a)^2}{3} \\ 3\frac{\sum_{i=1}^{n} x_i^2}{n} &= a^2 + a2\frac{\sum_{i=1}^{n} x_i}{n} - a^2 + 4\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 - a4\frac{\sum_{i=1}^{n} x_i}{n} + a^2\end{aligned}$$

De donde podemos obtener la ecuación de segundo grado:

$$a^2 - a2\frac{\sum_{i=1}^{n} x_i}{n} + 4\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 - 3\frac{\sum_{i=1}^{n} x_i^2}{n} = 0$$

Utilizando una solución de las ecuaciones de segundo grado:

$$\begin{aligned}a &= \frac{2\frac{\sum_{i=1}^{n} x_i}{n} \pm \sqrt{4\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 - 4\left(4\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 - 3\frac{\sum_{i=1}^{n} x_i^2}{n}\right)}}{2} \\ &= \frac{\sum_{i=1}^{n} x_i}{n} \pm \sqrt{\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 - \left(4\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 - 3\frac{\sum_{i=1}^{n} x_i^2}{n}\right)} \\ &= \frac{\sum_{i=1}^{n} x_i}{n} \pm \sqrt{3\left(\left(\frac{\sum_{i=1}^{n} x_i^2}{n}\right) - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right)} \\ &= \frac{\sum_{i=1}^{n} x_i}{n} \pm \sqrt{3} \cdot \hat{\sigma}\end{aligned}$$

Sustituyendo en la ecuación previa tenemos las soluciones siguientes para los estimadores de a y b:

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^{n} x_i}{n} \pm \sqrt{3} \cdot \hat{\sigma} \\ \hat{b} &= \frac{\sum_{i=1}^{n} x_i}{n} \mp \sqrt{3} \cdot \hat{\sigma} \end{aligned}$$

6.3.2. Método de Máxima Verosimilitud

El método de estimación más común es el de máxima verosimilitud. La característica cental del método de máxima verosimilitud es que buscamos máximizar la probabilidad conjunta de la muestra aleatoria dada $\{x_1, x_2, \ldots, x_n\}$, en cuyo caso es equivalente a las siguientes expresiones en el caso continuo y discreto, respectivamente:

$$\begin{aligned} f(x_1, x_2, \ldots, x_n; \boldsymbol{\theta}) &= \Pi_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \\ p(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) &= \Pi_{i=1}^{n} p(X_i = k_i) \end{aligned}$$

Dicho lo anterior, podemos establecer la siguiente definición:

Definición 6.18 *Sea $\{x_1, x_2, \ldots, x_n\}$ una muestra aleatoria de una pdf dada y sea $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ un conjunto de parámetros poblaciones desconocidos. La función de verosimilitud para una muestra aleatoria estará dada para el caso continuo por la expresión:*

$$L(\boldsymbol{\theta}) = \Pi_{i=1}^{n} f(x_i; \theta)$$

y para el caso discreto por la exprresión:

$$L(\boldsymbol{\theta}) = \Pi_{i=1}^{n} p(X_i = k_i)$$

De esta forma el problema consiste en resolver:

$$\max_{\hat{\boldsymbol{\theta}}} L(\boldsymbol{\theta}) = \max_{\hat{\boldsymbol{\theta}}} ln(L(\boldsymbol{\theta}))$$

La forma de solucionar el problema es determinar las condiciones de primer orden y de segundo orden. Lo cual implica el siguiente vector gradiente:

$$\frac{\partial ln(L(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial ln(L(\boldsymbol{\theta}))}{\partial \theta_1} \\ \frac{\partial ln(L(\boldsymbol{\theta}))}{\partial \theta_2} \\ \vdots \\ \frac{\partial ln(L(\boldsymbol{\theta}))}{\partial \theta_K} \end{bmatrix}$$

De esta forma, al igualar el el vector a 0 podemos encontrar los estimadores de $\boldsymbol{\theta}$ tales que:

$$\begin{bmatrix} \frac{\partial ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_1} \\ \frac{\partial ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_2} \\ \vdots \\ \frac{\partial ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_K} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Una vez determinados los estimadores requerimos verificar si la matriz de segundas derivadas evaludas en $\hat{\boldsymbol{\theta}}$ es definida negativa:

$$\frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_1^2} & \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_1 \partial \theta_K} \\ \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_2^2} & \cdots & \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_2 \partial \theta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_K \partial \theta_1} & \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_K \partial \theta_2} & \cdots & \frac{\partial^2 ln(L(\hat{\boldsymbol{\theta}}))}{\partial \theta_K^2} \end{bmatrix}$$

Ejemplo. Sea $X \sim Bin(\theta)$. Determinemos el estimador de máxima verosimilitud de θ. Partamos de que la pdf es:

$$p(X = k) = \begin{cases} \binom{n}{k} \theta^k (1-\theta)^{n-k} & \text{para } k = 0, 1, \ldots, n \\ 0 & \text{en cualquier otro caso} \end{cases}$$

En este caso, la función de verosimilitud será:

$$p(X = k) = L(\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Tomando el logaritmo de la función de verosimilitud:

$$ln(L(\theta)) = ln\left(\binom{n}{k}\right) + kln(\theta) + (n-k)ln(1-\theta)$$

Tomando la primera derivada e igualando a cero obtenemos:

$$\frac{dln(L(\theta))}{d\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta}$$

$$\begin{aligned} \frac{k}{\hat{\theta}} - \frac{n-k}{1-\hat{\theta}} &= 0 \\ \frac{k}{\hat{\theta}} &= \frac{n-k}{1-\hat{\theta}} \\ \hat{\theta} &= \frac{k}{n} \end{aligned}$$

Las segundas derivadas serán:

$$\frac{d^2 ln(L(\theta))}{d\theta^2} = -\frac{k}{\theta^2} - \frac{n-k}{(1-\theta)^2}$$

$$\begin{aligned}
-\frac{k}{\hat{\theta}^2} - \frac{n-k}{(1-\hat{\theta})^2} &= -\frac{k}{\left(\frac{k}{n}\right)^2} - \frac{n-k}{\left(1-\frac{k}{n}\right)^2} \\
-\frac{k}{\left(\frac{k}{n}\right)^2} - \frac{n-k}{\left(1-\frac{k}{n}\right)^2} &< 0 \\
-\frac{k}{\left(\frac{k}{n}\right)^2} &< \frac{n-k}{\left(1-\frac{k}{n}\right)^2} \\
-k\left(1-\frac{k}{n}\right)^2 &< (n-k)\left(\frac{k}{n}\right)^2 \\
-k + 2\frac{k^2}{n} - k\left(\frac{k}{n}\right)^2 &< n\left(\frac{k}{n}\right)^2 - k\left(\frac{k}{n}\right)^2 \\
-nk + 2k^2 &< k^2 \\
k(k-n) &< 0
\end{aligned}$$

Ejemplo. Sea $\{x_1, x_2, \ldots, x_n\}$ una muestra aleatoria de una pdf dada por:

$$f(x_i) = \begin{cases} \frac{1}{\theta} e^{-\frac{1}{\theta}x_i} & \text{para } x > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos el estimador de θ obtenido por máxima verosimilitud. Para lo cual determinamos que la función de verosimilitud es:

$$\begin{aligned}
L(\theta) &= f(x_1, x_2, \ldots, x_n; \theta) = \Pi_{i=1}^n f(x_i; \theta) = \Pi_{i=1}^n \frac{1}{\theta} e^{-\frac{1}{\theta}x_i} \\
&= \frac{1}{\theta} e^{-\frac{1}{\theta}x_1} \cdot \frac{1}{\theta} e^{-\frac{1}{\theta}x_2} \cdots \frac{1}{\theta} e^{-\frac{1}{\theta}x_n} = \left(\frac{1}{\theta}\right)^n e^{-\frac{1}{\theta}\sum_{i=1}^n x_i}
\end{aligned}$$

Tomando el logaritmo de $L(\theta)$:

$$ln(L(\theta)) = -nln(\theta) - \frac{1}{\theta}\sum_{i=1}^n x_i$$

Tomando las condiciones de primer orden:

$$\begin{aligned}
\frac{dln(L(\theta))}{d\theta} &= -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i \\
-\frac{n}{\hat{\theta}} + \frac{1}{\hat{\theta}^2}\sum_{i=1}^{n} x_i &= 0 \\
\hat{\theta} &= \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}$$

Las condiciones de segundo orden estarán dadas por:

$$\begin{aligned}
\frac{d^2ln(L(\theta))}{d\theta^2} &= \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i \\
\frac{n}{\hat{\theta}^2} - \frac{2}{\hat{\theta}^3}\sum_{i=1}^{n} x_i &< 0 \\
\frac{n}{\hat{\theta}^2} &< \frac{2}{\hat{\theta}^3}\sum_{i=1}^{n} x_i \\
\hat{\theta} &< 2\frac{\sum_{i=1}^{n} x_i}{n} \\
\frac{\sum_{i=1}^{n} x_i}{n} &< 2\frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}$$

Ejemplo. Sea una muestra aleatoria dada por $\{x_1, x_2, \ldots, x_n\}$ de una pdf dada por:

$$f(x_i) = \begin{cases} e^{-(x_i-\theta)} & \text{para } x_i \geq \theta; \theta > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos el estimador de θ mediante el procedimiento de máxima verosimilitud.

En este caso la función de verosimilitud estará dada por:

$$\begin{aligned}
L(\theta) &= \Pi_{i=1}^{n} e^{-(x_i-\theta)} = e^{-\sum_{i=1}^{n}(x_i-\theta)} = e^{-\sum_{i=1}^{n}(x_i)+n\theta} \\
ln(L(\theta)) &= -\sum_{i=1}^{n}(x_i) + n\theta
\end{aligned}$$

Dado lo anterior, tendríamos que la condición de primer orden sería:

$$\frac{dln(L(\theta))}{d\theta} = n \neq 0 \tag{6.8}$$

Por lo tanto, este tipo de funciones no tiene un estimador de máxima verosimilitud, ¿por qué? Consideramos que $\hat{\theta} = X_{min}$ es un buen candidato de estimador.

Ejemplo. Sea una muestra aleatoria dada por $\{x_1, x_2, \ldots, x_n\}$ de una pdf dada por:

$$f(x_i) = \begin{cases} \frac{1}{b-a} & \text{para } a < x_i < b \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Dado lo anterior, establezca los estimadores de máxima verosimilitd para a y b.

Para ello, tenemos que la función de verosimilitud y su logaritmo será:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \Pi_{i=1}^{n} \frac{1}{b-a} = \left(\frac{1}{b-a}\right)^n \\ ln(L(\boldsymbol{\theta})) &= n \cdot ln\left(\frac{1}{b-a}\right) = -nln(b-a) \end{aligned}$$

Las condiciones de primer orden serán:

$$\frac{\partial ln(L(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{n}{b-a} \\ -\frac{n}{b-a} \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Donde $\boldsymbol{\theta} = \begin{pmatrix} a \\ b \end{pmatrix}$

Entonces, proponemos que dichos estimadores sean:

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} X_{min} \\ X_{max} \end{bmatrix}$$

7

Estimación por Intervalo

7.1. Introducción

Anteriormente, consideramos el proceso de estimación puntual de θ. Ahora consideremos el proceso de estimación por intervalo. Ahora plantearemos un indicador de la precisión de los estimadores. Con la estimación puntual no nos es posible determinar un intervalo alrededor del cual se ubica el verdadero valor del parámetro. Así, tenemos que tomar en cuenta las variaciones del estimador.

En esta sección estableceremos el concepto de intervalo de confianza para determinar los valores que tienen una alta probabilidad de contener al parámetro desconocido en algún punto interior. Una forma de ilustrarlo es la siguiente:

Intervalo de Confianza	=	Estimador Puntual	+/-	Margen de Error

Ilustrado el punto, enunciemos algunas definiciones.

Definición 7.1 ***Intervalo de confianza (1)***. *Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con densidad $f(\cdot, \theta)$. Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estadísticas que satisfacen $\hat{\theta}_1 < \hat{\theta}_2$, para lo cual:*

$$\mathbb{P}[\hat{\theta}_1 < \theta < \hat{\theta}_2] = 1 - \alpha$$

donde $1 - \alpha$ no depende de θ, y $1 - \alpha$ es el nivel de confianza.

Definición 7.2 ***Intervalo de confianza (2)***. *Un intervalo de confianza para un parámetro θ es un intervalo $C_n = (a, b)$ en el cual ambos extremos $a = a(X_1, X_2, \ldots, X_n)$ y $b = b(X_1, X_2, \ldots, X_n)$ son funciones de los datos o muestra aleatoria observada, tales que:*

$$\mathbb{P}(\theta \in C_n) \geq 1 - \alpha, \, \forall \theta \in \Theta$$

Donde $1 - \alpha$ es el nivel de confianza.

Definición 7.3 ***Intervalo de Confianza (3)***. *Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos posibles valores para un estimador de θ tales que:*

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha$$

Para una probabilidad especifica $1 - \alpha$ diremos que $\hat{\theta}_1 < \theta < \hat{\theta}_2$ es un intervalo de confianza al nivel de significancia de α. Diremos que $\hat{\theta}_1$ y $\hat{\theta}_2$ son los límites del intervalo de confianza.

Antes de continuar, retomemos uno de los teoremas fundamentales de la Estadística:

Teorema 7.1 ***Teorema del Límite Central***. *Sea una muestra aleatoria $\{W_1, W_2, \ldots, W_n\}$. Supongamos que la pdf asociada a cada W_i tiene una media μ y una varianza σ^2. Para cualesquiera valores a y b:*

$$\lim_{n\to\infty} P\left(a \leq \frac{W_1 + W_2 + \ldots + W_n - n\mu}{\sqrt{n}\sigma} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz$$

Del teorema anterior podemos mostrar los siguientes resultados. Partamos de la primer parte de la ecuación del teorema:

$$\begin{aligned} P\left(a \leq \frac{W_1 + \ldots + W_n - n\mu}{\sqrt{n}\sigma} \leq b\right) &= P\left(a \leq \frac{\frac{W_1 + \ldots + W_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq b\right) \\ &= P\left(a \leq \frac{\overline{W} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq b\right) \end{aligned}$$

Por lo tanto, de acuerdo con el Teorema del Límite Central, tomanndo el límite podemos afirmar que para cualquier muestra aleatoria de una pdf, cuando n tiende a infinito, lo siguiente será cierto:

$$\lim_{n\to\infty} P\left(a \leq \frac{\overline{W} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz$$

Ahora, iniciemos la discusión con un ejemplo o caso ilustrativo. Supongamos que $\{x_1, x_2, \ldots, x_n\}$ es una muestra aleatoria de tamaño n de una pdf dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2 \cdot \left(\frac{x-\mu}{\sigma}\right)^2}, \text{ para } -\infty < x < \infty$$

Supongamos por un momento que conocemos el verdadero valor de σ^2, ¿cuáles serían los posibles valores para μ dado un. nivel de confianza? Para responder, partamos de retomar que un estimador para μ podría estar dado por:

$$\hat{\mu} = \frac{\sum_{i=1}^{n}}{n} = \overline{X}$$

En este punto es necesario recorrdar que si tenemos una muestra aleatoria $\{x_1, x_2, \ldots, x_n\}$ de una distribución normal con media y variaza dada por: μ y σ^2, entonces $\overline{X}$ será una normal con media y varianza: μ y $\frac{\sigma^2}{n}$, siempre y cuando asumamos (cosa que no hemos demostrado) que la suma de normales es una normal.

Ahora, ¿qué distribución tendrá una variable aleatoria dada por?:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

En este punto solo nos falta determinar el valor de su media y varianza.

Para responder determinemos los momentos, empezando por la media:

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right] \\
&= \frac{\sqrt{n}}{\sigma}\mathbb{E}[\overline{X} - \mu] \\
&= \frac{\sqrt{n}}{\sigma}\left(\mathbb{E}[\overline{X}] - \mu\right) \\
&= 0
\end{aligned}$$

Por su parte, la varianza será:

$$
\begin{aligned}
Var(Z) &= Var\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) \\
&= \frac{1}{\frac{\sigma^2}{n}} Var(\overline{X}-\mu) \\
&= \frac{1}{\frac{\sigma^2}{n}} \left(\mathbb{E}[(\overline{X}-\mu)^2] - (\mathbb{E}[\overline{X}-\mu])^2\right) \\
&= \frac{1}{\frac{\sigma^2}{n}} \left(\mathbb{E}[\overline{X}^2 - 2\mu\overline{X} + \mu^2]\right) \\
&= \frac{1}{\frac{\sigma^2}{n}} \left(\frac{1}{n^2}\mathbb{E}[(X_1+X_2+\ldots+X_n)(X_1+X_2+\ldots+X_n)] - \mu^2\right) \\
&= \frac{1}{\frac{\sigma^2}{n}} \left(\frac{1}{n^2}\mathbb{E}[X_1^2 + X_1X_2 + \ldots + X_1X_n + \ldots + X_n^2] - \mu^2\right) \\
&= \frac{1}{\frac{\sigma^2}{n}} \left(\frac{1}{n^2}(\sigma^2+\mu^2+\mu^2+\ldots+\mu^2+\mu^2+\ldots+\sigma^2+\mu^2) - \mu^2\right) \\
&= \frac{1}{\frac{\sigma^2}{n}} \left(\frac{1}{n^2}(n\sigma^2 + n^2\mu^2) - \mu^2\right) \\
&= \frac{1}{\frac{\sigma^2}{n}} \left(\frac{\sigma^2}{n}\right) \\
&= 1
\end{aligned}
$$

Por lo tanto, si $\overline{X} \sim N(\mu, \sigma^2/n)$, entonces la transformación $Z \sim N(0,1)$. Finalmente, para casos en lo que no se trate de datos de una normal, podemos inferir del Teorema del Límite Central que esta forma Z tendrá una distribución normal siempre que n sea muy grande.

En este caso, considerando que existen tablas de valores de una $N(0,1)$, de las cuales podemos identificar que el valos a patir del cual se acumula el 2.5 % de la probabilidad en cada una de la colas. Dicho valor, como discutiremos más adelante en la sección de prueba de hipótesis es 1.96.

Por lo tanto loa siguiente es cierto:

$$\begin{aligned} P(-1.96 \leq Z \leq 1,96) &= 0.95 \\ &= P\left(-1.96 \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1,96\right) \\ &= P\left(-\overline{X} - \frac{\sigma}{\sqrt{n}} \cdot 1.96 \leq -\mu \leq -\overline{X} + \frac{\sigma}{\sqrt{n}} \cdot 1,96\right) \\ &= P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} \cdot 1.96 \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} \cdot 1,96\right) \end{aligned}$$

Donde $1-\alpha = 0.95$ o $\alpha = 0.05$. Dicho de otro modo, el intervalo de confianza es al 95 %.

7.2. Intervalos de confianza para medias

Ahora entremos al concepto y procedimiento de estimación de medias y diferencia de medias. Para ello enunciaremos un teorema que, espero para este punto, es claro que hemos demostrado en la introducción de está sección.

Teorema 7.2 *Sea $\overline{X}$ la media de una muestra aleatoria de tamaño n de una pdf normal con varianza conocida, entonces si $\overline{X}$ es un estimador μ, la probabilidad de $1-\alpha$ será descrita por:*

$$P\left(|\overline{X} - \mu| \leq Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Donde $Z_{\alpha/2}$ es el valor de tablas a partir del cual se acumula el $\alpha/2$ de probabilidad de la curva de una normal con media 0 y varianza 1. Veamos ahora un ejemplo.

Ejemplo. Sea una muestra aleatoria de $n = 20$ de una normal con varianza conocida dada por $\sigma^2 = 225$. Supongamos que estimamos la media muestral y observamos que $\overline{X} = 64.3$. Así, un intervalo de confianza al 95 % del parámetro de la media μ estará dado por:

$$64.3 - 1.96 \cdot \frac{\sqrt{225}}{\sqrt{20}} < \mu < 64.3 + 1.96 \cdot \frac{\sqrt{225}}{\sqrt{20}}$$

Donde $Z_{0.025} = 1.96$. Por lo tanto, el intervalo al 95 % de confianza para la media es: $57.7 < \mu < 70.9$.

Hasta ahora hemos asumido que conocemos la varianza de la pdf, pero en general eso no es cierto. Por esta razón ahora propondremos un procedimiento para los casos en lo que no conocemos la varianza de la función. Para ello partamos de los siguientes teoremas.

Teorema 7.3 ***Distribución t-Student***. *Sean* $Z \sim N(0,1)$ *y* $Y \sim \chi^2_{[n]}$ *dos v.a's independientes, entonces:*

$$\frac{X}{\sqrt{\frac{Y}{n}}} \sim t_{[n]}$$

Teorema 7.4 ***Distribución t-Student para una muestra aleatoria***. *Sean* $\{x_1, x_2, \ldots, x_n\}$ *una muestra aleatoria de una pdf normal, de forma que cada* x_i *se distribuye como una* $N(\mu, \sigma^2)$*, entonces:*

$$T = \frac{\overline{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{[n-1]}$$

Donde:

$$\begin{aligned} \overline{X} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1} \end{aligned}$$

¿Cómo podemos llegar a este resultado? Hagamos algunas precisiones del teorema anterior, partamos de que:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \tag{7.1}$$

Consideremos que dada la ecuación (7.1), entonces para un muestreo de tamaño $n = 1$ ¿cómo se distribuye una transformación dada por?:

$$Z = \frac{x_i - \overline{X}}{\sigma}$$

Verifiquemos los momentos:

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}\left[\frac{x_i - \overline{X}}{\sigma}\right] \\ &= \frac{1}{\sigma}\mathbb{E}[x_i - \overline{X}] \\ &= 0 \end{aligned}$$

$$
\begin{aligned}
Var(Z) &= Var\left(\frac{x_i - \overline{X}}{\sigma}\right) \\
&= \frac{1}{\sigma^2} Var\left(x_i - \overline{X}\right) \\
&= \frac{1}{\sigma^2}\left(\mathbb{E}[(x_i - \overline{X})^2] - (\mathbb{E}[x_i - \overline{X}])^2\right) \\
&= \frac{1}{\sigma^2}\left(\mathbb{E}[x_i^2 - 2x_i\overline{X} + \overline{X}^2]\right) \\
&= \frac{1}{\sigma^2}\left((\sigma^2 + \mu^2) - \frac{2}{n}(\mu^2 + \mu^2 + \ldots + \sigma^2 + \mu^2 + \ldots + \mu^2)\right. \\
&\quad \left. + \frac{1}{n^2}(\sigma^2 + \mu^2 + \mu^2 + \ldots + \mu^2 + \ldots + \sigma^2 + \mu^2)\right) \\
&= \frac{1}{\sigma^2}\left((\sigma^2 + \mu^2) - \frac{2}{n}(\sigma^2 + n\mu^2) + \frac{1}{n^2}(n\sigma^2 + n^2\mu^2)\right) \\
&= \frac{1}{\sigma^2}\left(\sigma^2 - \frac{\sigma^2}{n}\right) \\
&= \frac{n-1}{n}
\end{aligned}
$$

Por lo tanto, tenemos que hacer un ajuste a la tranformación, propongamos a:

$$Z = \sqrt{\frac{n}{n-1}}\left(\frac{x_i - \overline{X}}{\sigma}\right) \sim N(0,1)$$

De esta forma, una propiedad que no vamos a demostrar pero que utilizaremos es que el cuadrado de una v.a. con distribución $N(0,1)$ es una $\chi^2_{[1]}$, por lo que:

$$
\begin{aligned}
Y &= Z^2 \\
&= \frac{n}{n-1}\left(\frac{x_i - \overline{X}}{\sigma}\right)^2 \sim \chi^2_{[1]}
\end{aligned}
$$

Utilizando lo anterior y asumiendo que la suma de n v.a.'s chi-cuadrado

son chi-cuadrado con n grados de libertad:

$$\begin{aligned} Y &= \sum_{i=1}^{n} \frac{n}{n-1} \left(\frac{x_i - \overline{X}}{\sigma} \right)^2 \sim \chi^2_{[n]} \\ &= \sum_{i=1}^{n} \frac{n}{n-1} \left(\frac{x_i - \overline{X}}{\sigma} \right)^2 \\ &= \sum_{i=1}^{n} \frac{n}{\sigma^2} \left(\frac{(x_i - \overline{X})^2}{n-1} \right) \\ &= \frac{\sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n-1}}{\frac{\sigma^2}{n}} \sim \chi^2_{[n]} \end{aligned} \tag{7.2}$$

Por lo tanto, el teorema anterior afirma que una distribución t-Student resultaría de la división de la ecuación (7.1) y la raíz cuadrada de la ecuación (7.2) dividida por sus grados de libertad:

$$\begin{aligned} T &= \frac{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n-1}}{\frac{\sigma^2}{n}}}} \\ &= \frac{\overline{X} - \mu}{\sqrt{\frac{\sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n-1}}{n}}} \\ &= \frac{\overline{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{[n-1]} \end{aligned} \tag{7.3}$$

En adelante utilizaremos la ecuación (7.3) para trabajar casos en los que no conocemos la varianza y además tenemos muestras pequeñas de una nor-

mal. En este sentido vamos a proponer un intervalo de confianza para T:

$$\begin{aligned} P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) &= 1-\alpha \\ &= P\left(-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}} < t_{\alpha/2,n-1}\right) \\ &= P\left(-\overline{X} - t_{\alpha/2,n-1}\cdot\frac{\hat{\sigma}}{\sqrt{n}} < -\mu \right. \\ &\quad \left. < -\overline{X} + t_{\alpha/2,n-1}\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right) \\ &= P\left(\overline{X} - t_{\alpha/2,n-1}\cdot\frac{\hat{\sigma}}{\sqrt{n}} < \mu \right. \\ &\quad \left. < \overline{X} + t_{\alpha/2,n-1}\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right) \end{aligned}$$

Por lo tanto podemos enunciar el siguiente teorema:

Teorema 7.5 *Sean $\overline{X}$ y $\hat{\sigma}^2$ valores de la media y varianza muestrales de una una muestra de tamaño n de una población normal, entonces, si $\overline{X}$ es un estimador de μ y $\hat{\sigma}^2$ es un estimador de σ^2, ña probabilidad $1-\alpha$ será descrita por:*

$$P\left(\overline{X} - t_{\alpha/2,n-1}\cdot\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right) = 1-\alpha \qquad (7.4)$$

Donde $t_{\alpha/2,n-1}$ es el valor de tablas a partir del cual se acumula el $\alpha/2$ de probabilidad de la curva de una distribución t-Student con $n-1$ grados de libertad.

Ejemplo. Sean $\overline{X} = 66.3$, $\hat{\sigma} = 8.4$ y $n = 12$ de una muestra aleatoria de una normal, utilicemos estos datos para determinar un intervalo de confianza para la media. Sustituyendo en la ecuación (7.4), tenemos:

$$66.3 - t_{\alpha/2,n-1}\cdot\frac{8.4}{\sqrt{12}} < \mu < 66.3 + t_{\alpha/2,n-1}\cdot\frac{8.4}{\sqrt{12}}$$

En este punto podemos elegir algún nivel de significancia, por ejemplo para una significancia al 95 % las tablas indican que el respectivo es: $t_{[0.025,11]} = 2.201$.

En el cuadro 7.1 resumimos cómo utilizar las herramientas aquí descritas. Como regla general utilizaremos una transformación Z siempre que sea conocida la varianza y una transformación T cuando no lo sea. No obstante, como regla de dedo se suele utilizar T para muestras pequeñas de $n < 30$ de una normal.

Por otro lado, por el teorema del límite central, podemos utilizar las transformaciones Z y T para datos que no provienen de una normal siempre que las muestras sean grandes, que por conveción diremos que es el caso cuando $n > 30$.

Cuadro 7.1: Uso de las transformaciones Z y T

Caso y valor de n	Muestras de poblaciones normales de cualquier n	Muestras de cualquier distribución para $n \geq 30$
σ^2 Conocida	Z	Z
σ^2 Desconocida	T	T

7.3. Intervalos de confianza para diferencias de medias

Para muestras aleatorias independientes de una población normal, de forma similar al caso de intervalos para medias, propongamos a Z como:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{7.5}$$

Donde $\overline{X}_1$ y $\overline{X}_2$ son las medias muestrales de poblaciones normales de tamaño n_1 y n_2, respectivamente. Asimismo, asumimos que conocemos la varianza de cada una de las distribuciones normales, las cuales son σ_1^2 y σ_2^2, respectivamente.

De forma similar a una transformación para una media, verifiquemos cuál es la media y varianza de la ecuación (7.5). Así, partamos de que esa expresión es la diferencia dada por:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\overline{X}_1 - \mu_1) - (\overline{X}_2 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

De esta expresión es sencillo verificar que su valor esperado o media es cero:

$$\mathbb{E}\left[\frac{(\overline{X}_1-\mu_1)-(\overline{X}_2-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right]=\frac{1}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\mathbb{E}[(\overline{X}_1-\mu_1)-(\overline{X}_2-\mu_2)]=0$$

Por lo que hace a la varianza tenemos que:

$$\begin{aligned}
Var\left[\frac{(\overline{X}_1-\mu_1)-(\overline{X}_2-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right] &= \frac{1}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\cdot Var[(\overline{X}_1-\mu_1)-(\overline{X}_2-\mu_2)] \\
&= \frac{1}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\Big(\mathbb{E}[((\overline{X}_1-\mu_1)-(\overline{X}_2-\mu_2))^2] \\
&\quad -(\mathbb{E}[(\overline{X}_1-\mu_1)-(\overline{X}_2-\mu_2)])^2\Big) \\
&= \frac{1}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\Big(\mathbb{E}[(\overline{X}_1-\mu_1)^2] \\
&\quad -2\mathbb{E}[(\overline{X}_1-\mu_1)(\overline{X}_2-\mu_2)] \\
&\quad +\mathbb{E}[(\overline{X}_2-\mu_2)^2]\Big) \\
&= \frac{1}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\left(\frac{\sigma_1^2}{n_1}-2\mathbb{E}[\overline{X}_1\overline{X}_2]+2\mathbb{E}[\mu_2\overline{X}_1]\right. \\
&\quad \left.+2\mathbb{E}[\mu_1\overline{X}_2]-2\mu_1\mu_2+\frac{\sigma_2^2}{n_2}\right) \\
&= \frac{1}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right) \\
&= 1
\end{aligned}$$

Por lo tanto, la transformación en la ecuación (7.5) se distribulle como normal con media cero y varianza 1:

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

De esta forma podemos establecer el siguiente intervalo de confianza para la diferencia de medias $\mu_1-\mu_2$ de la siguiente forma.

Teorema 7.6 *Sean* $\overline{X}_1$ *y* $\overline{X}_2$ *los valores de la medias muestrales de dos muestras aleatorias independientes de tamaño* n_1 *y* n_2 *de poblaciones normales con varianzas* σ_1^2 *y* σ_2^2 *conocidas, entonces:*

$$(\overline{X}_1 - \overline{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

es un intervalo de confianza al $1 - \alpha$ *de la diferencia de medias.*

Al respecto, hagamos algunas observaciones. Por vistud del Teorema del Límite Central, este intervalo de confianza puede ser usado para muestras aleatorias de poblaciones que no son normales cuando las varianzas son conocidas y sucede que $n_1 > 30$ y $n_2 > 30$. Veámos un ejemplo.

Ejemplo. Consideremos dos muestras aleatorias para las cuales observamos los resultados del Cuadro 7.2

Cuadro 7.2: Estadísticas de dos muestras aleatorias

Estadística	Muestra 1	Muestra 2
$\overline{X}$	418	402
σ	26	22
n	40	50

De esta forma podemos construir el siguiente intervalo de confianza:

$$(418 - 402) - z_{\alpha/2}\sqrt{\frac{26^2}{40} + \frac{22^2}{50}} < \mu_1 - \mu_2 < (418 - 402) + z_{\alpha/2}\sqrt{\frac{26^2}{40} + \frac{22^2}{50}}$$

$$16 - z_{\alpha/2}5.16 < \mu_1 - \mu_2 < 16 + z_{\alpha/2}5.16$$

En este punto lo único que nos hace falta es elegir el nivel de confianza que queremos determinar $(\alpha/2)$. De forma similar podemos construir un intervalo de confianza cuando las varianzas son desconocidas. En este caso plantearemos el problema a través de la transformación:

$$\begin{aligned} T &= \frac{Z}{\sqrt{\frac{Y}{n_1+n_2-2}}} \\ &= \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{[n_1+n_2-2]} \end{aligned} \tag{7.6}$$

Donde en la ecuación (7.6):

$$\hat{\sigma}_p^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

Por lo tanto, propondremos el siguiente intervalo de confianza dado por aquel en el que la probabilidad $1 - \alpha$ está dada de acuerdo con el siguiente teorema.

Teorema 7.7 *Sean* $\overline{X}_1$, $\overline{X}_2$, $\hat{\sigma}_1$, $\hat{\sigma}_2$ *valores de las medias y desviaciones estándar de muestras aleatorias independientes de tamaño* n_1 *y* n_2 *de una población normal con las mismas varianzas, entonces:*

$$\begin{aligned}&(\overline{X}_1 - \overline{X}_2) - \hat{\sigma}_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{\alpha/2, n_1+n_2-2} \\ &< \mu_1 - \mu_2 < \\ &(\overline{X}_1 - \overline{X}_2) + \hat{\sigma}_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{\alpha/2, n_1+n_2-2}\end{aligned}$$

Es un intervalo de confianza al $(1 - \alpha)$ *por ciento de la diferencia de medias de dos muestras de poblaciones. Donde:*

$$\hat{\sigma}_p = \sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}$$

Ejemplo. Supongamos que estamos analizando muestras aleatorias independientes del contenido de cierto químico en dos productos dados. Supongamos que obtenemos los resultados del Cuadro 7.3:

Cuadro 7.3: Tabla de muestras aleatorias

	Muestra 1	Muestra 2
$\overline{X}$	3.1	2.7
$\hat{\sigma}$	0.5	0.7
n	10	8

En primer lugar obtendremos:

$$\hat{\sigma}_p = \sqrt{\frac{(10 - 1)(0.5)^2 + (8 - 1)(0.7)^2}{10 + 8 - 2}} = 0.596$$

Sustituyendo en el intervalo:

$$\begin{aligned}&(3.1-2.7)-(0.596)\sqrt{\frac{1}{10}+\frac{1}{8}}\cdot t_{[\alpha/2,n_1+n_2-2]}\\&<\mu_1-\mu_2<\\&(3.1-2.7)+(0.596)\sqrt{\frac{1}{10}+\frac{1}{8}}\cdot t_{[\alpha/2,n_1+n_2-2]}\end{aligned}$$

Algunas observaciones:

- En este caso se cumple el resumen mostrado en el Cuadro 7.1.
- Si en el intervalo de confianza se encuentra el 0 podemos concluir que no hay diferencia estadíticamente significativamente.

7.4. Intervalos de confianza para proporciones y diferencia de proporciones

En muchos casos deseamos estimar intervalos de confianza de proporciones, probabilidades, procentajes o tasas. En estos casos es razonable pensar estos casos como observaciones de una v.a. binomial $X = k$ con un parámetro θ bajo una transformación dada por:

$$Z = \frac{k - n\theta}{\sqrt{n\theta(1-\theta)}}$$

Gracias al Teorema del Límite Central $Z \sim N(0,1)$ para casos en que n es grande. Bajo estas condciones podemos establecer que:

$$\begin{aligned}P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= 1-\alpha\\&= P\left(-z_{\alpha/2} < \frac{k-n\theta}{\sqrt{n\theta(1-\theta)}} < z_{\alpha/2}\right)\\&= P\left(\frac{k}{n} - \sqrt{\frac{\theta(1-\theta)}{n}}\cdot z_{\alpha/2}\right.\\&\qquad\left.< \theta < \frac{k}{n} + \sqrt{\frac{\theta(1-\theta)}{n}}\cdot z_{\alpha/2}\right)\end{aligned}$$

Por lo tanto, podemos establecer el siguiente teorema:

Teorema 7.8 *Sea $X = k$ una v.a. binomial con parámetro θ y con una muestra n grande. Asumamos un estimador de θ dado por:*

$$\hat{\theta} = \frac{X}{n} \tag{7.7}$$

Entonces,

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

Donde $z_{\alpha/2}$ es el valor de tavlas a partir de la cual se acumula el $\alpha/2$ de probabilidad de la curva de una distribución normal con media 0 y varianza 1.

Ejemplo. Supongamos una muestra aleatoria de $n = 400$ de los cuales observamos que $X = 136$ indivuduos cumplen con cierta característica. Dicho esto podemos establecer que $\hat{\theta} = 136/400 = 0.34$ y que un intervalo estará dado por:

$$0.34 - \sqrt{\frac{0.34(1-0.34)}{400}} \cdot z_{\alpha/2} < \theta < 0.34 + \sqrt{\frac{0.34(1-0.34)}{400}} \cdot z_{\alpha/2}$$

Ejemplo. Supongamos que deseamos estimar un intervalo de confianza para analizar la opinión pública sobre algún tema en particular. Concretamente, asumamos que tenemos información sobre la opinión de las personas sobre la despenalización del consumo de diversas drogas. Supongamos que hemos realizado una encuesta en la población de México con representatividad nacional, en la cual se ha preguntado:

En general, ¿piensas que es apropiado que el Estado genere y establezca las condiciones para que las personas adultas (mayores de 18 años) puedan acceder o comprar libremente mariguana y algunos de sus productos derivados?

Supongamos que la encuesta más reciente indica que de 1,200 adultos sleccionados aleatoriamente, 396 se pronunciaron a favor de la despenalización y otros 804 manifestaron estar en contra. Digamos que buscamos aprovar una Ley en este sentido, por ello resulta relevante que determinemos un intervalo de confianza para la proporción de personas que dijo que si.

Sea p la proporción de personas que dijo que si. Como este es un proceso Bernoulli, sabemos que un estimador de dicho parámetro podría ser:

$$\hat{p} = \frac{396}{1200} = 0.33$$

De esta forma, $1 - \hat{p} = 0.67$. Como hemos demostrado anteriormente, la estimación del error estándar de la proporción estará dado por:

$$se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.33 \times 0.67}{1200}} = \sqrt{0.000184} = 0.0136$$

En conclusión, el intervalo de confianza estará dado por:

$$\begin{aligned} 0.33 - 0.0136 \times 1.96 < p < 0.33 + 0.0136 \times 1.96 \\ 0.30 < p < 0.36 \end{aligned}$$

Así, podemos concluir que el porcentaje de la población que podría estar a favor de la medida oscila entre 30 % y 36 %. Por lo que con un 95 % de confianza no llega a ser ni el 50 %.

En otro caso, cuando buscamos estimar la diferencia entre 2 parámetros de una binomial θ_1 y θ_2 con base en 2 muestras aleatorias independientes de tamaños n_1 y n_2 provenientes de dos poblaciones normales. Estos casos suelen estar relacionados con, por ejemplo, proporciones de hobres y mujeres que votan a afvor de cierto candidato en alguna elección.

Para darle contexto al planteamiento, sean $X_1 = k_1$ y $X_2 = k_2$ la cantidad de éxitos en las muestras indpendientes de dos poblaciones, de esta forma podemos establecer dos estimadores:

$$\hat{\theta}_1 = \frac{k_1}{n_1} \text{ y } \hat{\theta}_2 = \frac{k_2}{n_2}$$

Digamos que queremos determinar un intervalo de confianza para la diferencia de entre ambos estimadores: $\hat{\theta}_1 - \hat{\theta}_2$, de la cual revisamos los momentos de distribución de ésta que estará dada por una media:

$$\mathbb{E}[\hat{\theta}_1 - \hat{\theta}_2] = \frac{1}{n_1} \cdot \mathbb{E}[X_1] - \frac{1}{n_2} \cdot \mathbb{E}[X_2] = \frac{1}{n_1} \cdot n_1\theta_1 - \frac{1}{n_2} \cdot n_2\theta_2 = \theta_1 - \theta_2$$

Y una varianza:

$$\begin{aligned}
Var[\hat{\theta}_1 - \hat{\theta}_2] &= \mathbb{E}[(\hat{\theta}_1 - \hat{\theta}_2)^2] - (\mathbb{E}[\hat{\theta}_1 - \hat{\theta}_2])^2 \\
&= \mathbb{E}[\hat{\theta}_1^2] - 2\mathbb{E}[\hat{\theta}_1\hat{\theta}_2] + \mathbb{E}[\hat{\theta}_2^2] - (\theta_1 - \theta_2)^2 \\
&= \mathbb{E}\left[\left(\frac{X_1}{n_1}\right)^2\right] - 2\theta_1\theta_2 + \mathbb{E}\left[\left(\frac{X_2}{n_2}\right)^2\right] - \theta_1^2 + 2\theta_1\theta_2 - \theta_2^2 \\
&= \frac{1}{n_1^2}\mathbb{E}[X_1^2] + \frac{1}{n_2^2}\mathbb{E}[X_2^2] - \theta_1^2 - \theta_2^2 \\
&= \frac{1}{n_1^2}(n_1\theta_1(1-\theta_1) + n_1^2\theta_1^2) + \frac{1}{n_1^2}(n_2\theta_2(1-\theta_2) + n_2^2\theta_2^2) \\
&\quad -\theta_1^2 - \theta_2^2 \\
&= \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}
\end{aligned}$$

Donde hemos utilizado que para cualquier v.a. X_i con distribución binomial con parámetros θ_i y n_i se cumple lo siguiente:

$$\begin{aligned}
Var[X_i] &= n_i\theta_i(1-\theta_i) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 \\
&= n_i\theta_i(1-\theta_i) = \mathbb{E}[X_i^2] - n_i^2\theta_i^2
\end{aligned}$$

Lo que implica que:

$$\mathbb{E}[X_i^2] = n_i\theta_i(1-\theta_i) + n_i^2\theta_i^2$$

De esta forma, de acuerdo con el Teorema del Límite Central (para grandes muestras) de las v.a.'s X_1 y X_2 podemos establecer la siguiente transformación:

$$Z = \frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}}}$$

La cual es una v.a. que tiene una distribución normal y para la cual podemos establecer que:

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Dicho esto, plateamos el siguiente teorema:

Teorema 7.9 *Sean $X_1 = k_1$ y $X_2 = k_2$ v.a.'s con parámetros n_1, θ_1, n_2 y θ_2, respectivamente. Supongamos que n_1 y n_2 son 'grandes' y que tenemos dos estimadores de los parámetros desconocidos:*

$$\hat{\theta_1} = \frac{k_1}{n_1} \text{ y } \hat{\theta_2} = \frac{k_2}{n_2}$$

Entonces,

$$(\hat{\theta}_1 - \hat{\theta}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$$
$$< \theta_1 - \theta_2 <$$
$$(\hat{\theta}_1 - \hat{\theta}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$$

es un intervalo de confianza al $(1-\alpha)100\,\%$ de confianza.

Ejemplo. Supongamos 2 muestras aleatorias independientes, para las cuales se observa la información del Cuadro 7.4.

Cuadro 7.4: Dos muestras aleatorias de poblaciones binomiales

Estadística	Muestra 1	Muestra 2
X_i	132	90
n_i	200	150

Construyamos un intervalo de confianza para la diferencia de medias $\theta_1 - \theta_2$, por lo que tenemos:

$$\hat{\theta_1} = \frac{132}{200} = 0.66 \text{ y } \hat{\theta_2} = \frac{90}{150} = 0.60$$

De acuerdo con las tablas estadísticas, al $99\,\%$ de confianza tenemos que $z_{\alpha/2} = z_{0.005} = 2.57$. Con esta información podemos determinar que el intervalo de confianza es:

$$(0.66 - 0.60) - 2.57 \cdot \sqrt{\frac{0.66(1-0.66)}{200} + \frac{0.60(1-0.60)}{150}}$$
$$< \theta_1 - \theta_2 <$$
$$(0.66 - 0.60) + 2.57 \cdot \sqrt{\frac{0.66(1-0.66)}{200} + \frac{0.60(1-0.60)}{150}}$$
$$-0.074 < \theta_1 - \theta_2 < 0.194$$

Notemos que el intervalo incluye al cero (0), ¿cuál es el significado de esto?

7.5. Intervalos de confianza para varianzas y razón de varianzas

Para finalizar esta sección discutiremos intervalos de confianza para las varianzas. Partamos de que:

$$\frac{\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n-1}}{\frac{\sigma^2}{n}} = n \cdot \frac{\sum_{i=1}^{n}\frac{(x_i-\overline{X})}{n-1}}{\sigma^2} = n \cdot \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{[n]}$$

Entonces, podemos afirmar que:

$$(n-1) \cdot \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{[n-1]}$$

Utilizando lo anterior, podemos establecer que:

$$P\left[\chi^2_{[1-\alpha/2,n-1]} < \frac{(n-1)\hat{\sigma}^2}{\sigma^2} < \chi^2_{[\alpha/2,n-1]}\right] = 1-\alpha$$

$$P\left[\frac{(n-1)\hat{\sigma}^2}{\chi^2_{[\alpha/2,n-1]}} < \sigma^2 < \frac{(n-1)\hat{\sigma}^2}{\chi^2_{[1-\alpha/2,n-1]}}\right] = 1-\alpha$$

Utilizando lo anterior, podemos enunciar el siguiente teorema:

Teorema 7.10 *Sea $\hat{\sigma}^2$ el valor de la varianza muestral de una muestra de tamaño n de una población normal, entonces:*

$$\frac{(n-1)\hat{\sigma}^2}{\chi^2_{[\alpha/2,n-1]}} < \sigma^2 < \frac{(n-1)\hat{\sigma}^2}{\chi^2_{[1-\alpha/2,n-1]}}$$

Donde:

$$\hat{\sigma}^2 = \sum_{i=1}^{n}\frac{(x_i-\overline{X})^2}{n-1}$$

Es un intervalo de confianza al $(1-\alpha)100\,\%$ de confianza para σ^2.

Ejemplo. Supongamos una muestra aleatoria de tamaño $n = 16$ de una población normal para la cual determinamos que $\hat{\sigma}^2 = 2.2$, establezcamos un intervalos de confianza al 99 %. Para lo cual lozalizamos en tablas los valores:

$$\chi^2_{[0.005,15]} = 32.8 \text{ y } \chi^2_{[0.995,15]} = 4.6$$

De esta forma, el intervalo de la varianza quedará delimitado como:

$$\frac{(16-1)(2.2)^2}{32.8} < \sigma^2 < \frac{(16-1)(2.2)^2}{4.6}$$
$$2.21 < \sigma^2 < 15.78$$

De forma similar al caso de diferencias de medias, en varianzas tenemos un concepto similar que es el de razones de varianzas el cual supone dos muestras aleatorias independientes de tamaños n_1 y n_2 de poblaciones normales, cuyas varianzas muestrales son: $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$. De esta forma, retomando la propiedad de que una p.d.f. dada por una F de Fisher, resulta del cociente entre dos variables aleatorias Chi cuadrado divididas por sus grados de libertad.

Así, supongamos dos Chi cuadradas con $n_1 - 1$ y $n_2 - 1$ grados de libertad:

$$\frac{\frac{\chi^2_{[n_1-1]}}{n_1-1}}{\frac{\chi^2_{[n_2-1]}}{n_2-1}} = \frac{\frac{(n_1-1)\cdot\frac{\hat{\sigma}_1^2}{\sigma_1^2}}{n_1-1}}{\frac{(n_2-1)\cdot\frac{\hat{\sigma}_2^2}{\sigma_2^2}}{n_2-1}} = \frac{\frac{\hat{\sigma}_1^2}{\sigma_1^2}}{\frac{\hat{\sigma}_2^2}{\sigma_2^2}}$$

De esta manera planteamos:

$$F = \frac{\frac{\hat{\sigma}_1^2}{\sigma_1^2}}{\frac{\hat{\sigma}_2^2}{\sigma_2^2}} \sim F_{[n_1-1,n_2-1]}$$

De esta forma podemos platear que si:

$$P\left[f_{[1-\alpha/2,n_1-1,n_2-1]} < \frac{\frac{\hat{\sigma}_1^2}{\sigma_1^2}}{\frac{\hat{\sigma}_2^2}{\sigma_2^2}} < f_{[\alpha/2,n_1-1,n_2-1]}\right] = 1-\alpha$$

y que:

$$f_{[1-\alpha/2,n_1-1,n_2-1]} = \frac{1}{f_{[\alpha/2,n_2-1,n_1-1]}}$$

De esta forma podemos plantear el siguiente teorema:

Teorema 7.11 *Sean $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$ varianzas muestrales de 2 muestras aleatorias independientes de tamaño n_1 y n_2, respectivamente, de poblaciones normales, entonces:*

$$\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}\frac{1}{f_{[\alpha/2,n_2-1,n_1-1]}} < \frac{\sigma_2^2}{\sigma_1^2} < f_{[\alpha/2,n_1-1,n_2-1]}\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}$$

Es un intervalo de confianza al $(1-\alpha)100\,\%$ de confianza para $\frac{\sigma_2^2}{\sigma_1^2}$.

Ejemplo. Supongamos dos muestras aleatorias de poblaciones normales para las cuales observamos las estadísticas del Cuadro 7.5. Determines un intervalo de confianza para el $98\,\%$.

Cuadro 7.5: Estadísticas para razón de varianzas

Estadística	Muestra 1	Muestra 2
$\hat{\sigma}_i$	0.5	0.7
n	10	8

De la búsqueda en las tablas estadísticas tenemos:

$$f_{[0.01,9,7]} = 6.72 \text{ y } f_{[0.01,7,9]} = 5.61$$

Entonces el intervalo de confianza será:

$$\frac{0.25}{0.49}\cdot\frac{1}{6.72} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{0.25}{0.49}\cdot 5.61$$

$$0.076 < \frac{\sigma_1^2}{\sigma_2^2} < 2.862$$

8

Pruebas de Hipótesis

8.1. Elementos de una prueba estadística

En esta sección discutiremos el concepto de **hipótesis prueba estadística**, la cual es una conjetura o afirmación sobre alguna(s) característica de una distribución de probabilidad de una o más variables aleatorias.

Está técnica esdística es utilizada para validar proposiciones que son enunciadas como dicotómicas y son referidas como pruebas de hipótesis. La cual se descompone en dos enunciados.

- El primero es denominado como una hipótesis nula (H_0), la cual buscamos aceptar o rechazar.
- El segundo es la hipótesis alternativa (H_a), la cual sería válida una vez que hemos rechazado a la hipótesis nula.

Formalmente, supondremos que es posible hacer una partición del espacio de paramétros Θ en dos connjuntos disjuntos Θ_0 y Θ_1, y que pretendemos probar:

$$H_0 : \theta \in \Theta_0 \text{ Vs } H_0 : \theta \notin \Theta_1 \tag{8.1}$$

Sea X una variable aleatoria y sea $\mathcal{X}$ el rango de X. Así, probaremos una hipótesis definiendo un subconjunnto apropiado de resultados $R \subset \mathcal{X}$ llamado la región de rechazo.

Si $X \in R$ rechazamos la hipótesis nula, en otro caso, nno rechazamos la hipótesis nula. Es deci:

$$\begin{aligned} X \in R &\Rightarrow \text{Rechazamos } H_0 \\ X \notin R &\Rightarrow \text{NO Rechazamos } H_0 \end{aligned}$$

Usualmente, la región de rechazo R es de la forma:

$$R = \{x : T(x) > c\} \tag{8.2}$$

Donde T es una estadística de prueba y c el valor crítico. El problema en la prueba de hipótesis es encontrar una estadística apropiada T y el valor crítico apropiado c.

Definición 8.1 ***Hipótesis y Pruebas de hipótesis.***

1. *Una hipótesis de la forma $\theta = \theta_0$ es denominada prueba simple*

2. *Una prueba de hipótesis de la forma:*

$$H_0 : \theta = \theta_0 \ Vs \ H_1 : \theta \neq \theta_0$$

   es llamada una prueba de 2 colas

3. *Una hipótesis de la forma $\theta > \theta_0$ o $\theta < \theta_0$ es denominada prueba compuesta*

4. *Una prueba de hipótesis de la forma:*

$$H_0 : \theta \leq \theta_0 \ Vs \ H_1 : \theta > \theta_0$$

   o

$$H_0 : \theta \geq \theta_0 \ Vs \ H_1 : \theta < \theta_0$$

   es llamada una prueba de 1 cola. En general para estas pruebas podemos utilizar una representación de la hipótesis nula utilizando el signo $=$ en lugar de $\leq$ o $\geq$ pero preservando que la hipótesis alternativa es como se mostró anteriormente

8.2. Potencia de una prueba

La posibilidad de establecer una conclusión incorrecta es un subproducto inevitable de las pruebas de hipótesis. Sin importar la manera de cómo establezcamos la prueba, no existe garantía de que la prueba nos lleve a la conclusión correcta.

Existen 2 tipos de errores que podemos cometer cuando realizamos una prueba de hipótesis:

1. **Error Tipo I**. Podemos rechazar H_0 cuando H_0 es cierta. La probabilidad que identifica a este error es denotada por α, y

2. **Error Tipo II**. Podemos equivocarnos en no rechazar H_0 cuando H_0 es falsa. La probabilidad que identifica a este error es denotada por β.

Otra forma de establecerlo es decir:

1. Si el verdadero valor de un parámetro θ es θ_0 y concluimos incorrectamente que $\theta = \theta_1$, estaremos en un escenario del Error Tipo I, y

2. Si el verdaro valor de θ es θ_1 y concluimos incorrectamente que $\theta = \theta_0$, estaremos en el escenario de un Error Tipo II.

Al mismo tiempo existen 2 tipos de **desiciones correctas**:

1. Podemos no rechazar H_0 cuando H_0 es cierta, y

2. Podemos rechazar H_0 cuando H_0 es falsa.

En resumen tenemos lo siguiente:

Cuadro 8.1: Tabla Tipo de Error

		Estado verdadero de la naturaleza	
		H_0 cierta	H_1 cierta
Desición	No rechazo H_0	Desición Correcta	Error Tipo II: β
	Rechazo H_0	Error Tipo I: α	Desición Correcta

Tratemos de ejemplificar la situación y supongamos que queremos probar la hipótesis nula de que la media de una población normal con varianza $\sigma^2 = 1$ es μ_0, contra la hipótesis alternativa de que la media es μ_1, donde $\mu_1 > \mu_0$,

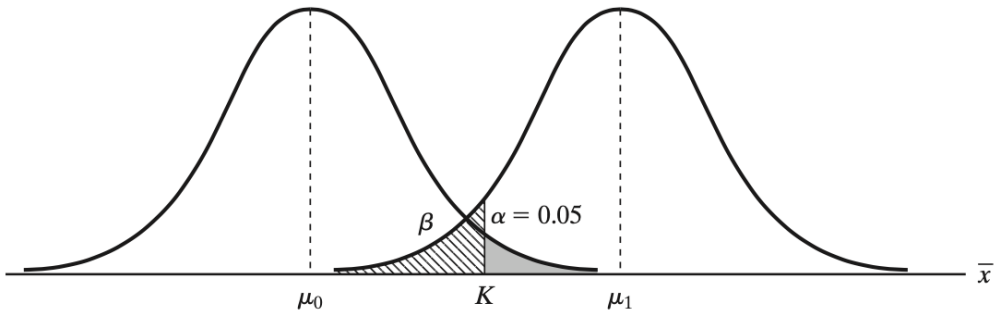

Figura 8.1: Ilutración de las probabilidades de Error Tipo I y Error Tipo II. Retomado de Miller y Miller (2014; p. 340) I. Miller y M. Miller 2014

¿cuál es el valor de la constante K tal que $\overline{X} > K$ proporciona una región crítica de tamaño $\alpha = 0.05$?

La respuesta a esta pregunta se puede responder partiendo de la transformación Z e igualandóla al valor de tablas que acula una probabilidad de $\alpha = 0.005$:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - \mu_0}{\frac{1}{\sqrt{n}}} = \sqrt{n}(\overline{X} - \mu_0) = z_{0.05}$$

A partir de que $z_{0.005} = 1.645$, podemos encontrar que:

$$\overline{X} = \mu_0 + \frac{1.645}{\sqrt{n}}$$

Así, la constante será $K = \mu_0 + \frac{1.645}{\sqrt{n}}$. En la Figura 8.1 ilustramos el valor de esta constante. Asimismo, en la Figura 8.1 se ilustra los tipos de error y las probabilidades asociadas a cada uno.

En ese mismo sentido, definieremos el concepto de **Potencia de una Prueba**, la cual será el valor daddo por $1 - \beta$. De forma similar definiremos el **Nivel de Significancia de una Prueba** al valor de α. Concepto del cual abundaremos en la sección siguiente en la discusión de la regla de desición.

8.3. Nivel de significancia de una prueba de hipótesis

La prueba de hipótesis tiene una lógica similar al caso de los intervalos de confianza para medias o para razones de varianzas, entre otros. En este senti-

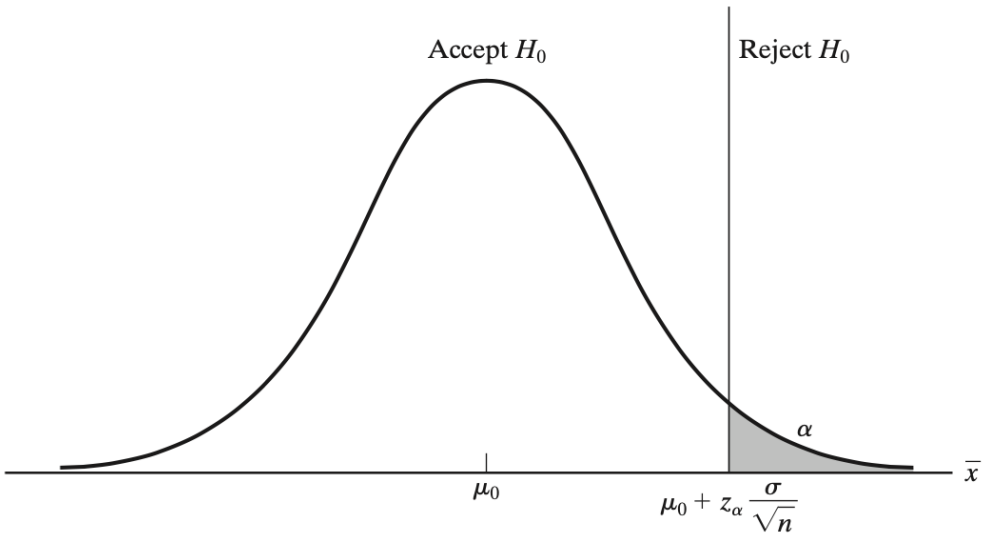

Figura 8.2: Hipótesis de una cola (derecha) y su zona de rechazo. Retomado de Miller y Miller (2014; p. 360) I. Miller y M. Miller 2014

do, la primera prueba de hipótesis que plantearemos es aquellas que asumen que conocemos la varianza de la población, σ^2. Después iremos relajando ese supuesto de forma similar al caso cuando analizamos intervalos de confianza.

Sea $\{X_1, X_2, \ldots, X_n\}$ una muestra aleatoria de tamaño n de una población con función de densidad de probabilidad normal estándar ($N(0,1)$), donde σ es conocida. Sea:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Entonces tenemos los siguientes casos como regla de rechazo:

1. Para probar $H_0 : \mu = \mu_0$ Vs $H_a : \mu > \mu_0$ al nivel de significancia α, rechazamos H_0 si $Z \geq z_{\alpha/2}$. En términos de la distribución orginal de $\overline{X}$, la zona de rechazo será como se describe en la Figura 8.2.

2. Para probar $H_0 : \mu = \mu_0$ Vs $H_a : \mu < \mu_0$ al nivel de significancia α, rechazamos H_0 si $Z \leq -z_{\alpha/2}$. En términos de la distribución orginal de $\overline{X}$, la zona de rechazo será como se describe en la Figura 8.3.

3. Para probar $H_0 : \mu = \mu_0$ Vs $H_a : \mu \neq \mu_0$ al nivel de significancia α, rechazamos H_0 si $Z \leq -z_{\alpha/2}$ o $Z \geq z_{\alpha/2}$. En términos de la distribución

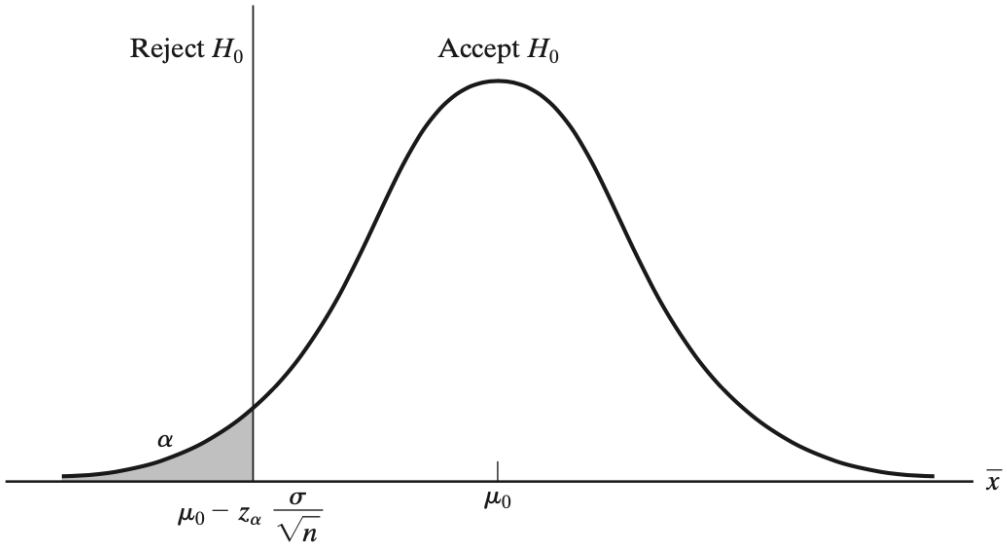

Figura 8.3: Hipótesis de una cola (izquierda) y su zona de rechazo. Retomado de Miller y Miller (2014; p. 361) I. Miller y M. Miller 2014

orginal de $\overline{X}$, las dos zonas de rechazo serán como se describe en la Figura 8.4.

En este sentido, requerimos de establecer un concepto de p-value. El p-value está asociado con una estadística de prueba y es la probabilidad de obtener un valor para la estadística extrema relativa a H_1, dado que H_0 es cierta. En otras palabras, es la probabilidad acumulada en el o los extremos de la curva antes y después de la constante K.

8.4. Pruebas para medias, varianzas y proporciones

8.4.1. Pruebas para medias

En este caso cabe aclarar que todas las pruebas estan basadas en muestras aleatorias tomadas de poblaciones con distribuciones normales o, en su caso, son muestras suficientemente grandes ($n > 30$) para utilizar la aproximación normal que nos da el uso del Teorema del Límite Central.

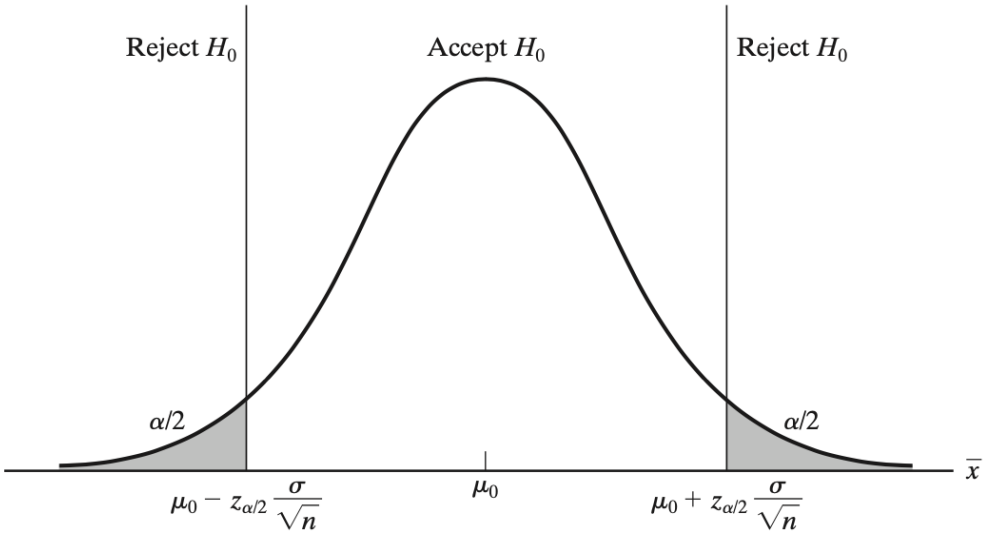

Figura 8.4: Hipótesis de dos colas y sus zonas de rechazo. Retomado de Miller y Miller (2014; p. 360) I. Miller y M. Miller 2014

Partamos suponiendo que queremos probar:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_a &: \begin{cases} \mu \neq \mu_0 \\ \text{o} \\ \mu > \mu_0 \\ \text{o} \\ \mu < \mu_0 \end{cases} \end{aligned} \tag{8.3}$$

En la base de una muestra aleatoria de tamaño n de una población normal con una varianza conocida, σ^2.

Dada la hipótesis nula proponemos las siguientes regiones críticas o de su

rechazo, respectivamente,s con las hipótesis alternativas en la ecuación (8.3):

$$\begin{cases} |Z| \geq z_{\alpha/2} \\ \text{o} \\ Z \geq z_{\alpha} \\ \text{o} \\ Z \leq -z_{\alpha} \end{cases}$$

Ejemplo. Supongamos una muestra aleatoria de las calificaciones de 25 alumnos de una materia y escuela determinadas. Supongamos que la población de donde se seleccionó la muestra tiene una distribución normal con desviación estándar 0.16. Dicha muestra reporta un promedio de 8.091, pruebe la hipótesis nula:

$$\begin{aligned} H_0 &: \mu = 8 \\ H_a &: \mu \neq 8 \end{aligned}$$

Asumamos que queremos una región critica o de rechazo del 1 %, es decir, es una prueba con una confianza del 99 %. Dicho esto, el valor critico será $z_{\alpha/2} = z_{0.005} = 2.575$. Utilizaremos una transformación: c Entonces, dado que $Z = 2.84 > 2.575$ concluimos que podemos rechazar la hipótesis nula de que la media de las calificaciones es 8 al 99 % de confianza.

Visto como un intervalo de confianza podemos establecer:

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= 1 - \alpha \\ P(-2.575 < Z < 2.575) &= 0.99 \end{aligned}$$

Donde el valor de $Z = 2.84$ es un valor fuera del intervalo o en la zona de rechazo. De esta forma estamos expuestos un Error Tipo I.

¿Qué pasa cuando no son datos de una normal? En esos casos podemos emplear una transformación Z cuando la muestra es suficientmeente grande, dado el Teorema del Límite Central.

Ejemplo. Supongamos una muestra aleatoria de tamaño 100 de una población cuya función de densidad de probabilidad cualquiera con desviación estándar $\sigma = 1.295$. Supongamos que la media muestral es 21.819, queremos probar que:

$$\begin{aligned} H_0 &: \mu = 22 \\ H_a &: \mu < 22 \end{aligned}$$

Supongamos que queremos una región crítica al 5 % o una prueba con un nivel de confianza del 95 %. De esta forma $z_\alpha = z_{0.05} = 1.645$. Por lo tanto, rechazaremos la hipótesis nula si $Z < -1.645$. Así:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{21.819 - 22}{\frac{1.295}{\sqrt{100}}} = -13.976$$

Dado que $Z = -13.976 < -1.645$, entonces podemos rechazar la hipótesis nula.

¿Qué pasa cuando desconocemos la varianza? En esos casos, al igual que en los intervalos de confianza, utilizaremos una estadística:

$$T = \frac{\overline{X} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{[n-1]}$$

En estos casos, de forma similar queremos probar las hipótesis de la ecuación (8.3), para la cual proponemos las siguientes regiones críticas o de su rechazo, respectivamente, con las hipótesis alternativas en la ecuación (8.3):

$$\begin{cases} |T| \geq t_{[\alpha/2,n-1]} \\ \text{o} \\ T \geq t_{[\alpha,n-1]} \\ \text{o} \\ T \leq -t_{[\alpha,n-1]} \end{cases}$$

Ejemplo. Sea una muestra aleatoria de una población normal de tañamo $n = 5$, de donde: $\overline{X} = 183.1$, $\hat{\sigma} = 8.2$ y $\alpha = 0.05$, y probemos la hipótesis:

$$\begin{aligned} H_0 &: \mu = 185 \\ H_a &: \mu < 185 \end{aligned}$$

De acuerdo con las tablas estadísticas tenemos: $t_{[\alpha,n-1]} = t_{[0.05;4]} = 2.132$, por lo tanto rechazamos la hipótesis nula si $T < -2.132$. Así:

$$T = \frac{\overline{X} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} = T = \frac{183.1 - 185}{\frac{8.2}{\sqrt{5}}} = 0.49$$

Por lo tanto, no podemos rechazar la hipótesis nula.

8.4.2. Pruebas para diferencias de medias

Supongamos 2 muestras aleatorias de tamaños n_1 y n_2 tomadas de 2 poblaciones normales con medias μ_1 y μ_2, respectivamente. Asumamos en un principio que conocemos las varianzas σ_1^2 y σ_2^2. De esta forma podemos platear la siguiente estructura de hipótesis nula e hipótesis alternativa:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = \delta \\ H_a &: \begin{cases} \mu_1 - \mu_2 \neq \delta \\ \text{o} \\ \mu_1 - \mu_2 > \delta \\ \text{o} \\ \mu_1 - \mu_2 < \delta \end{cases} \end{aligned} \tag{8.4}$$

De forma similar al caso de hipótesis para medias podemos establecer la siguiente transformación para evaluar la estrtuctura de hipótesis planteada en la ecuación (8.4):

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{8.5}$$

Dada la hipótesis nula en la ecuación (8.4) proponemos las siguientes regiones críticas o de su rechazo, respectivamente,s con las hipótesis alternativas en la ecuación (8.5):

$$\begin{cases} |Z| \geq z_{\alpha/2} \\ \text{o} \\ Z \geq z_{\alpha} \\ \text{o} \\ Z \leq -z_{\alpha} \end{cases}$$

Ejemplo. Supongamos dos muestras aleatorias tomadas de poblaciones normales, para las cuales observamos los datos del cuadro 8.2. Asumiendo un nivel de confianza del 95 %, determine si la diferencia de medias es de 0.20.

De esta forma, la hipótesis del problema se puede platear como:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0.20 \\ H_a &: \mu_1 - \mu_2 \neq 0.20 \end{aligned}$$

Cuadro 8.2: Estadísticas de dos muestras aleatorias

Estadística	Muestra 1	Muestra 2
$\overline{X}$	2.61	2.38
n	50	40
σ	0.12	0.14

Con los datos anteriores tenemos que:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{2.61 - 2.38 - 0.20}{\sqrt{\frac{(0.12)^2}{50} + \frac{(0.14)^2}{40}}} = 1.08$$

De acuerdo con las tablas de la normal, $z_{\alpha/2} = z_{0.05/2} = 1.96$, dado que $Z = 1.08 < z_{0.05/2}$ y $Z = 1.08 > -z_{0.05/2}$, concluimos que no podemos rechazar H_0, por lo que tenemos evidencia estadística que señala que la diferencia entre las medias poblacionales es de 0.20.

¿Cómo camabia las hipótesis cuando desconocemos las variazas de las poblaciones de donde provienen las muestras? En estos casos utilizaremos una estructura de hipótesis similar a la ecuación (8.4) y una trasnformación dada por:

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{[n_1+n_2-2]} \tag{8.6}$$

Donde:

$$\hat{\sigma}_p^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

$$\hat{\sigma}_i^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n - 1}$$

De forrma similar establecemos las reglas de rechazo para la ecuación (8.6) como:

$$\begin{cases} |T| \geq t_{[\alpha/2, n_1+n_2-2]} \\ \text{o} \\ T \geq t_{[\alpha, n_1+n_2-2]} \\ \text{o} \\ T \leq -t_{[\alpha, n_1+n_2-2]} \end{cases}$$

Ejemplo. Supongamos dos muestras aleatorias de dos poblaciones normales, de las cuales no conocemos la varianza. Supongamos que obtenemos los resultados del Cuadro 8.3. Supongamos que queremos determinar si con una confianza del 95 % las media de la población 1 es mayor que la media de la población 2.

Cuadro 8.3: Estadísticas de dos muestras aleatorias

Estadística	Muestra 1	Muestra 2
$\overline{X}$	546	492
n	4	4
$\hat{\sigma}$	31	26

De esta forma, la hipótesis del problema se puede platear como:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_a &: \mu_1 - \mu_2 > 0 \end{aligned}$$

Con los datos anteriores tenemos que:

$$\hat{\sigma}_p = \sqrt{\frac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}} = \sqrt{\frac{3(31)^2+3(26)^2}{4+4-2}} = 28.609$$

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \delta}{\hat{\sigma}_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{546-492-0}{28.609\sqrt{\frac{1}{4}+\frac{1}{4}}} = 2.67$$

De acuerdo con las tablas de la t-Student, $t_{\alpha/2,n_1+n_2-2} = t_{0.025;6} = 1.943$, dado que $T = 2.67 > t_{0.025;6} = 1.943$, concluimos que podemos rechazar H_0, por lo que tenemos evidencia estadística que señala que la media de la población 1 es mayor que la media de la población 2.

8.4.3. Pruebas para varianzas

Este tipo de pruebas tiene el objetivo de medir o comparar la variabilidad de los datos de una muestra aleatoria de tamaño n de una población con funciòn de densidad de probabilidad normal.

También se emplean para el análisis de comperación de varianzas, con el objeto de identificar que cumplen con el supuesto de misma varianza en el caso de diferencias de medias.

Empecemos por platear la estructura de las hipótesis de la siguiente forma:

$$\begin{aligned} H_0 &: \sigma^2 = \sigma_0^2 \\ H_a &: \begin{cases} \sigma^2 \neq \sigma_0^2 \\ \text{o} \\ \sigma^2 > \sigma_0^2 \\ \text{o} \\ \sigma^2 < \sigma_0^2 \end{cases} \end{aligned} \tag{8.7}$$

Para los cual establecemos la siguiente transformación para evaluar la ecuación (8.7):

$$\chi^2 = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{[n-1]} \tag{8.8}$$

Donde

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{n-1}$$

Dada la hipótesis nula en la ecuación (8.7) proponemos las siguientes regiones críticas o de su rechazo, respectivamente,s con las hipótesis alternativas en la ecuación (8.8):

$$\begin{cases} \chi^2 \geq \chi^2_{[\alpha/2,n-1]} \text{ o } \chi^2 \leq \chi^2_{[1-\alpha/2,n-1]} \\ \text{o} \\ \chi^2 \geq \chi^2_{[\alpha,n-1]} \\ \text{o} \\ \chi^2 \leq \chi^2_{[1-\alpha,n-1]} \end{cases}$$

Ejemplo. Supongamos uan muestra aleatoria de una población con distribución normal de tamaño $n = 18$ y una varianza muestral de $\hat{\sigma} = 0.68$. Realicemos una prueba de hipótesis al 95 % de confianza para la afimarción de que la varianza muestal es más grande que la varianza poblacional $\sigma^2 = 0.36$.

De esta forma podemos establece la siguiente estructura de la hipótesis:

$$\begin{aligned} H_0 &: \sigma^2 = 0.36 \\ H_a &: \sigma^2 > 0.36 \end{aligned}$$

Para lo cual formulamos la siguiete expresión:

$$\chi^2 = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2} = \frac{(18-1)0.68}{0.36} = 32.11$$

De acuerdo con las tabla de Chi cuadrado, encontramos que $\chi^2_{[0.05;17]} = 27.587$.

Por lo tanto, rechazamos la hipótesis nula ya que $\chi^{=}32.11 > \chi^2_{[0.05;17]} = 27.587$. De esta forma, concluimos la varianza de la población de donde proviene la muestra es más grande que 0.36.

¿Qué pasa ciando queremos comparar 2 varianzas de 2 poblaciones distintas? Supongamos $\hat{\sigma}_1^2$ y $\hat{\sigma}_2^2$ varianzas muestrales de poblaciones con distribuciones normales. Asumamos que las muestras son de tamaño n_1 y n_2.

De esta forma podemos establecer la siguiente estructura de hipótesis:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_a &: \begin{cases} \sigma_1^2 \neq \sigma_2^2 \\ \text{o} \\ \sigma_1^2 > \sigma_2^2 \\ \text{o} \\ \sigma_1^2 < \sigma_2^2 \end{cases} \end{aligned} \tag{8.9}$$

Para los cual establecemos la siguiente transformación para evaluar la ecuación (8.9):

$$F = \frac{\frac{\hat{\sigma}_1^2}{\sigma_1^2}}{\frac{\hat{\sigma}_2^2}{\sigma_2^2}} = \frac{\hat{\sigma}_1^2\sigma_2^2}{\hat{\sigma}_2^2\sigma_1^2} \sim f_{[n_1-1,n_2-1]} \tag{8.10}$$

Dada la hipótesis nula en la ecuación (8.9) proponemos las siguientes regiones críticas o de su rechazo, respectivamente,s con las hipótesis alternativas en la ecuación (8.10):

$$\begin{cases} F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \geq f_{[\alpha/2,n_1-1,n_2-1]} \text{ si } \hat{\sigma}_1^2 > \hat{\sigma}_2^2 \text{ o } F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} \geq f_{[\alpha/2,n_2-1,n_1-1]} \text{ si } \hat{\sigma}_2^2 > \hat{\sigma}_1^2 \\ \text{o} \\ F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \geq f_{[\alpha,n_1-1,n_2-1]} \text{ si } \hat{\sigma}_1^2 > \hat{\sigma}_2^2 \\ \text{o} \\ F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} \geq f_{[\alpha,n_2-1,n_1-1]} \text{ si } \hat{\sigma}_2^2 > \hat{\sigma}_1^2 \end{cases}$$

Ejemplo. Supongamos que tenemos 2 muestras aleatorias de 2 poblaciones con distribuciones normales, para las cuales observamos los datos del Cuadro 8.4.

Cuadro 8.4: Table para razón de varianzas

Estadística	Muestra 1	Muestra 2
$\hat{\sigma}_i^2$	19.2	3.5
n	13	16

Utilice una prueba al 98 % de confianza para determinar si las siguientes hipótesis son ciertas:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_a &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

Construimos una transformación:

$$F = \frac{\hat{\sigma}_2^2 \sigma_1^2}{\hat{\sigma}_1^2 \sigma_2^2} \sim f_{[n_1-1, n_2-1]} = \frac{19.2}{3.5} = 5.49 \tag{8.11}$$

De acuerdo con la tabla de F de Fisher, tenemos que $f_{[0.001:12,15]} = 3.67$. Así, $F = 3.49 > f_{[0.001:12,15]} = 3.67$, por lo que podemos rechazar la hipótesis nula de que las varianzas sean iguales.

8.4.4. Pruebas para proporciones

Finalmente, platearemos una prueba para el caso de proporciones, para el cual partimos de observaciones de variables aleatorias binomiales y suponemos la siguiente estructura de hipótesis.

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_a &: \begin{cases} \theta \neq \theta_0 \\ \text{o} \\ \theta > \theta_0 \\ \text{o} \\ \theta < \theta_0 \end{cases} \end{aligned} \tag{8.12}$$

Para lo cual utilizaremos la transformación:

$$Z = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}$$

De esta forma podemos establecer las siguientes reglas de rechazo y zonas criticas:

$$\begin{cases} |Z| \geq z_{\alpha/2} \\ \text{o} \\ Z \geq z_{\alpha} \\ \text{o} \\ Z \leq -z_{\alpha} \end{cases}$$

Ejemplo. Supongamos que sabemos que menos del 20 % de una población tiene un padecimiento o enfermedad. Suponga que tomamos una muestra aleatoria de 200 individuos de dicha población y observamos que 22 tienen el padecimiento. Al respecto se pregunta si el procentaje de personas con el padecimiento es menor al 22 %. Así formulamos las siguientes hipótesis:

$$\begin{aligned} H_0 &: \theta = 0.2 \\ H_a &: \theta < 0.2 \end{aligned}$$

Así:

$$Z = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = \frac{22 - 200(0.20)}{\sqrt{200(0.20)(1-(0.20))}} = -3.18$$

De acuerdo con las tablas de la normal, $z_\alpha = 2.33$, de esta forma, $Z = -3.18 < z_\alpha = -2.33$ por lo que podemos decir que el procentaje ha disminuido.

8.5. Pruebas de razón de verosimilitud

Este tipo de pruebas son más potentes para hipótesis simples y compuestas cuando se conoce la distribución de las observaciones. Sea $L(\theta)$ una función de verosimilitud de θ parámetros. Así, definamos a λ como la razón de verosimilitud dada por:

$$\lambda = \frac{L(\theta_0)}{L(\theta_u)} \tag{8.13}$$

Donde $L(\theta_0)$ es la función de verosimilitud del modelo rentrigido evaluada en el estimador generado con dicha función, y $L(\theta_u)$ es la función de verosimilitud para el modelo no restringido evaluada en el estimados generado con dicha función.

Asumiremos una estructura de hipótesis dada por:

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

Definición 8.2 ***Prueba de razón de verosimilitud****. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de $f(\cdot)_0$ y $f(\cdot)_1$. Una prueba del tipo:*

$$\begin{aligned} H_0 &: x_i \sim f(\cdot)_0 \\ H_1 &: x_i \sim f(\cdot)_1 \end{aligned}$$

Está definida para seer una razón de verosimilitud simple en la que la regla de rechazo y aceptación es: rechazamos H_0 si $\lambda < k$; aceptamos H_0 si $\lambda > k$ y ees no concluyente si $\lambda = k$.

Donde k es una constante no negativa y:

$$\lambda = \lambda(x_1, x_2, \ldots, x_n) = \frac{\Pi_{i=1}^n f(x_i)_0}{\Pi_{i=1}^n f(x_i)_1} = \frac{L_0(X_1, X_2, \ldots, X_n)}{L_1(X_1, X_2, \ldots, X_n)} = \frac{L_0}{L_1}$$

Como se trata de una razón de verosimilitud que compara un escenario no restringido con uno restringido, claramente $0 \leq \lambda \leq 1$, por lo que la hipótesis nual será cierta si λ es cercana a cero.

En condiciones de regularidad de la función $f(y;\theta)$ y para n grande utilizaremos la siguiente estadística:

$$-2ln\lambda \sim \chi^2_Q$$

Donde Q es el número de restricciones a probar. Dado un valor de región crítica α, diremos que las regiones de rechazo serán:

$$\chi^2 \geq \chi^2_{[\alpha, n-1]}$$

Ejemplo. Supongamos que una empresa desea probar si el número de clientes por hora que visitan dos de sus tiendas es estadísticamente el mismo. Para ello supongamos que extraemos dos muestras de tamaño 100 para ambas

tiendas de forma que obtenemos que la media para la primera muestra es $\overline{X} = 20$ y $\overline{Y} = 22$.

Supongamos que la distribución del número de cleintes que visitan las tiendas es Poisson. Asuma un error tipo I, $\alpha = 0.01$. Imaguinemos que queremos probar la hipótesis de que:

$$H_0 \quad : \quad \theta_1 = \theta_2 \tag{8.14}$$
$$H_1 \quad : \quad \theta_1 \neq \theta_2 \tag{8.15}$$

Recordemos que una función Poisson es de la siguiente forma:

$$f(x) = \frac{e^{-\theta}\theta^x}{x!}$$

Así, la funciónn de verosimilitud para dos muestras aleatorias independientes estará dada por:

$$\begin{aligned} L(\theta_1, \theta_2) &= \Pi_{i=1}^n f(x_i) \cdot \Pi_{i=1}^n f(y_i) \\ &= \frac{e^{-\theta}\theta^{x_i}}{x_i!} \cdot \frac{e^{-\theta}\theta^{y_i}}{y_i!} \\ &= \frac{1}{x_1!, x_2!, \ldots, x_n! \cdot y_1!, y_2!, \ldots, y_n!} \cdot e^{(-n\theta_1)+(-n\theta_2)} \theta_1^{\sum_{i=1}^n x_i} \theta_2^{\sum_{i=1}^n y_i} \end{aligned}$$

Bajo la hipótesis nula $\theta_1 = \theta_2 = \theta$, la función de verosimilitud sería:

$$L(\theta) = \frac{1}{x_1!, x_2!, \ldots, x_n! \cdot y_1!, y_2!, \ldots, y_n!} \cdot e^{-n2\theta} \theta^{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$$

Recordemos que sabemos como es el estimador cuando no se mantiene la hipótesis nula tendríamos que los estimadores de máxima verosimilitud serían:

$$\begin{aligned} \hat{\theta}_1 &= \overline{X} \\ \hat{\theta}_2 &= \overline{Y} \end{aligned}$$

Sólo nos faltaría determinar el estimador de θ cuando se cumple la hipótesis nula:

$$lnL(\theta) = ln\left(\frac{1}{k}\right) - 2n\theta + \left(\sum_{i=1}^n x_i + \sum_{i=1}^n y_i\right) ln(\theta)$$

Donde:

$$k = x_1!, x_2!, \ldots, x_n! \cdot y_1!, y_2!, \ldots, y_n!$$

Así, las condicionnes de primer orden serían:

$$\frac{\partial}{\partial\theta} lnL(\theta) = \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}{\theta} - 2n$$

De esta forma la solución estará dada por:

$$\hat{\theta} = \frac{1}{2}\frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}{n} = \hat{\theta}_0$$

De esta forma podemos establecer la estadística de razón de verosimilitud como:

$$\begin{aligned}
\lambda &= \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_u)} \\
&= \frac{\frac{1}{x_1!,x_2!,\ldots,x_n!\cdot y_1!,y_2!,\ldots,y_n!} \cdot e^{-n2\hat{\theta}}\hat{\theta}^{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}}{\frac{1}{x_1!,x_2!,\ldots,x_n!\cdot y_1!,y_2!,\ldots,y_n!} \cdot e^{(-n\hat{\theta}_1)+(-n\hat{\theta}_2)}\hat{\theta}_1^{\sum_{i=1}^{n} x_i}\hat{\theta}_2^{\sum_{i=1}^{n} y_i}} \\
&= \frac{e^{-n2\hat{\theta}}\hat{\theta}^{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}}{e^{(-n\hat{\theta}_1)+(-n\hat{\theta}_2)}\hat{\theta}_1^{\sum_{i=1}^{n} x_i}\hat{\theta}_2^{\sum_{i=1}^{n} y_i}} \\
&= \frac{e^{-n2\hat{\theta}}\hat{\theta}^{n\overline{X}+n\overline{Y}}}{e^{(-n\hat{\theta}_1)+(-n\hat{\theta}_2)}\hat{\theta}_1^{n\overline{X}}\hat{\theta}_2^{n\overline{Y}}} \\
&= \frac{21^{100\times 20+100\times 22}}{20^{100\times 20}22^{100\times 22}}
\end{aligned}$$

Al finalizar,

$$-2ln(\lambda) = 9.53$$

De acuerdo con las tablas:

$$\chi^2_{0.01} = 6.635$$

Por lo tanto, rechazamos la hipótesis nula.

8.6. Bondad de ajuste

9

Análisis de varianza y diseño de experimentos

9.1. Análisis de varianza: Introducción y motivación

El análisis de varianza está basado en un prueba estadística del tipo F-Fisher. Para ilustrarla supongamos un experimento en el que deseamos comparar el efecto promedio de un factor en el caso de k grupos ($k \geq 2$). En este caso podemos establecer un procedimiento que inicia con la prueba de hipótesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \tag{9.1}$$

Para probar si la hipótesis en la ecuación (9.1) es válidad podemos utilizar una prueba basada en una razón de verosimilitud:

$$\lambda = \frac{L(\omega_e)}{L(\omega_e)}$$

No obstante, en esta sección incorporaremos un análisis basado en otra estadística que construímos a continuación. Sea y_{ij} una observación independiente de una variable aleatoria para la cual observamos una muestra con media μ_j para el grupo $j = 1, 2, \cdots, K$ y una varianza dada por σ^2 para todos los grupos j. Esto es:

$$f_{Y_{ij}}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu_j}{\sigma}\right)^2} \tag{9.2}$$

Donde $-\infty < y < \infty$.

Supongamos n observaciones en la muestra, donde:

$$\mu = \frac{1}{n}\sum_{j=1}^{K} n_j \mu_j$$

Si H_0 es cierta entonces $\mu = \mu_j$, $\forall j$. Para establecer una estadística apropiada, iniciemos estimando cada uno de los μ_j para cada una de las muestras aleatorias de una distribución normal indexadas de la forma: $y_{1j}, y_{2j}, \cdots, y_{n_j j}$.

Tomemos como estimador de μ_j a $\overline{Y}_j$ definido de la forma habitual. Asimismo, propongamos como un estimador de μ a $\overline{\mathbf{Y}}$:

$$\overline{\mathbf{Y}} = \frac{1}{n}\sum_{j=1}^{K} n_j \overline{Y}_j$$

Dado lo anterior, podemos construir una estadística que denominaremos la suma de cuadrados del tratamiento ($SSTR$, por sus siglas en inglés), la cual es la suma de los cuadrados de las desviaciones de las medias muestrales respecto del promedio de medias muestrales:

$$SSTR = \sum_{j=1}^{K}\sum_{i=1}^{n_j}(\overline{Y}_j - \overline{\mathbf{Y}})^2 \tag{9.3}$$

Podemos desarollar la ecuación (9.3) de forma que tendríamos:

$$\begin{aligned} SSTR &= \sum_{j=1}^{K}\sum_{i=1}^{n_j}(\overline{Y}_j - \overline{\mathbf{Y}})^2 \\ &= \sum_{j=1}^{K} n_j(\overline{Y}_j - \overline{\mathbf{Y}})^2 \end{aligned}$$

A la expresión de la ecuación (9.3) también se le conoce como la variavilidad total de las medias muestrales respecto de la media total. A esta también

se le puede descomponer en una expansión al rededor de $\overline{Y}_j$, $\overline{\mathbf{Y}}$ y μ:

$$\begin{aligned}
SSTR &= \sum_{j=1}^{K} n_j(\overline{Y}_j - \overline{\mathbf{Y}})^2 \\
&= \sum_{j=1}^{K} n_j((\overline{Y}_j - \mu) - (\overline{\mathbf{Y}} - \mu))^2 \\
&= \sum_{j=1}^{K} n_j((\overline{Y}_j - \mu)^2 + (\overline{\mathbf{Y}} - \mu)^2 - 2(\overline{Y}_j - \mu)(\overline{\mathbf{Y}} - \mu)) \\
&= \sum_{j=1}^{K} n_j(\overline{Y}_j - \mu)^2 + (\overline{\mathbf{Y}} - \mu)^2 \sum_{j=1}^{K} n_j - 2(\overline{\mathbf{Y}} - \mu) \sum_{j=1}^{K} n_j(\overline{Y}_j - \mu) \\
&= \sum_{j=1}^{K} n_j(\overline{Y}_j - \mu)^2 + n(\overline{\mathbf{Y}} - \mu)^2 - 2(\overline{\mathbf{Y}} - \mu) \sum_{j=1}^{K} (n_j\overline{Y}_j - n_j\mu) \\
&= \sum_{j=1}^{K} n_j(\overline{Y}_j - \mu)^2 + n(\overline{\mathbf{Y}} - \mu)^2 - 2(\overline{\mathbf{Y}} - \mu) \left(\sum_{i=1}^{n} y_i - n\mu \right) \\
&= \sum_{j=1}^{K} n_j(\overline{Y}_j - \mu)^2 + n(\overline{\mathbf{Y}} - \mu)^2 - 2(\overline{\mathbf{Y}} - \mu)n(\overline{\mathbf{Y}} - \mu) \\
&= \sum_{j=1}^{K} n_j(\overline{Y}_j - \mu)^2 - n(\overline{\mathbf{Y}} - \mu)^2
\end{aligned}$$

Una vez determinado una expresión alterna para la ecuación (9.3) podemos verificar su valor esperado:

$$\begin{aligned}
\mathbb{E}[SSTR] &= \mathbb{E}\left[\sum_{j=1}^{K} n_j(\overline{Y}_j - \mu)^2 - n(\overline{\mathbf{Y}} - \mu)^2 \right] \\
&= \sum_{j=1}^{K} n_j \mathbb{E}[(\overline{Y}_j - \mu)^2] - n\mathbb{E}[(\overline{\mathbf{Y}} - \mu)^2]
\end{aligned}$$

Para continuar requerimos recordar que:

$$\mathbb{E}[\overline{\mathbf{Y}}] = \mu$$

Así:

$$Var(\overline{\mathbf{Y}}) = \mathbb{E}[(\overline{\mathbf{Y}} - \mu)^2] = \frac{\sigma^2}{n}$$

Por otro lado, analicemos la siguiente expresión:

$$Var(\overline{Y}_j - \mu) = \mathbf{E}[(\overline{Y}_j - \mu)^2] - (\mathbf{E}[\overline{Y}_j - \mu])^2$$

Por lo tanto:

$$\begin{aligned} \mathbf{E}[(\overline{Y}_j - \mu)^2] &= Var(\overline{Y}_j - \mu) + (\mathbf{E}[\overline{Y}_j - \mu])^2 \\ &= \frac{\sigma^2}{n_j} + (\mathbf{E}[\overline{Y}_j - \mu])^2 \end{aligned}$$

Dados los resultados anteriores podemos determinar que:

$$\begin{aligned} \mathbb{E}[SSTR] &= \sum_{j=1}^{K} n_j \mathbb{E}[(\overline{Y}_j - \mu)^2] - n\mathbb{E}[(\overline{\mathbf{Y}} - \mu)^2] \\ &= \sum_{j=1}^{K} n_j (Var(\overline{Y}_j - \mu) + (\mathbf{E}[\overline{Y}_j - \mu])^2) - nVar(\overline{\mathbf{Y}}) \\ &= \sum_{j=1}^{K} n_j \frac{\sigma^2}{n_j} + \sum_{j=1}^{K} n_j (\mu_j - \mu)^2 - n\frac{\sigma^2}{n} \\ &= K\sigma^2 + \sum_{j=1}^{K} n_j (\mu_j - \mu)^2 - \sigma^2 \\ &= (K-1)\sigma^2 + \sum_{j=1}^{K} n_j (\mu_j - \mu)^2 \end{aligned}$$

9.2. Planteamiento de la prueba de hipótesis

Prueba de $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ cuando σ^2 es conocida
La hipótesis plantea que todas las medias son iguales. De ser cierta:

$$\mathbb{E}[SSTR] = (K-1)\sigma^2$$

Por el contrario, en caso de que la hipótesis no sea cierta ocurriría que:

$$\mathbb{E}[SSTR] > (K-1)\sigma^2$$

De esta forma, podemos ver que cuando la hipótesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ es ccierta, entonces:

$$\frac{SSTR}{\sigma^2} = \sum_{j=1}^{K} n_j \frac{(\overline{Y}_j - \mu)^2}{\sigma^2} - n\frac{(\overline{\mathbf{Y}} - \mu)^2}{\sigma^2} \sim \chi^2_{[K-1]}$$

Entoncces, suponiendo un nivel de significancia α y que σ^2 es conocida, podemos rechazar la hipótesis nula siempre que:

$$\frac{SSTR}{\sigma^2} \geq \chi^2_{[1-\alpha,K-1]}$$

Pruebas de hipótesis cuando σ^2 es desconocida

En este caso cada una de las K muestras pueden proveernos un estimados independiente e insesgado para σ^2. Así, podemos retomar que un estimador de la varianza muestral estará dado por:

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \overline{Y}_j)^2$$

Ahora, multiplicando cada una de las $\hat{\sigma}_j^2$ por $n_j - 1$ y sumando las K varianzas muestrales podemos encontrar una estadística que denominaremos como la suma de los errores al cuadrado (SSE, por sus siglas en inglés):

$$SSE = \sum_{j=1}^{K} (n_j - 1)\hat{\sigma}_j^2 = \sum_{j=1}^{K} \sum_{i=1}^{n_j} (y_{ij} - \overline{Y}_j)^2 \tag{9.4}$$

Notemos que sin importar si H_0 es cierta, una expresión como la siguiente tiene características como:

$$\frac{SSE}{\sigma^2} = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \frac{(y_{ij} - \overline{Y}_j)^2}{\sigma^2} \sim \chi^2_{[n-K]} \tag{9.5}$$

Definamos, para complementar el análisis, a la estadística de la variabilidad total o summa de cuadrados total ($STOT$, por sus siglas en inglés) como:

$$SSTOT = SSTR + SSE$$

Es decir,

$$SSTOT = \sum_{j=1}^{K}\sum_{i=1}^{n_j}(y_{ij} - \overline{\mathbf{Y}})^2 = \sum_{j=1}^{K}\sum_{i=1}^{n_j}((\overline{Y}_j - \overline{\mathbf{Y}}) + (y_{ij} - \overline{Y}_j))^2 \qquad (9.6)$$

Podemos demostrar que:

$$\begin{aligned} SSTOT &= \sum_{j=1}^{K}\sum_{i=1}^{n_j}(y_{ij} - \overline{\mathbf{Y}})^2 \\ &= \sum_{j=1}^{K}\sum_{i=1}^{n_j}(\overline{Y}_j - \overline{\mathbf{Y}})^2 + \sum_{j=1}^{K}\sum_{i=1}^{n_j}(y_{ij} - \overline{Y}_j)^2 \end{aligned}$$

Dicho lo anterior podemos establecer que para probar la hipótesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ podemos utilizar una estadística F-Fisher:

$$F = \frac{SSTR/(K-1)}{SSE/(n-K)} \sim F_{[K-1,n-K]} \qquad (9.7)$$

Por lo tanto, rechzaremos la hipótesis nula siempre que:

$$\frac{SSTR/(K-1)}{SSE/(n-K)} \geq F_{[1-\alpha,K-1,n-K]}$$

9.3. Tablas ANOVA

Los cálculos del análisis de varianza suelen presentarse como tablas. Altamente estructuradas, estas tablas son especialmente útiles en la implementación de pruebas estadísticas como la hipótesis: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$. Así, la tabla sepuede construir como se muestra en el Cuadro 9.1.

Cuadro 9.1: Tabla ANOVA

Fuente	df	SS	MS	F	Prob.
Tratamiento	$K-1$	$SSTR$	$MSTR = \frac{SSTR}{K-1}$	$\frac{MSTR}{MSE}$	$P(F_{[K-1,n-K]} \geq F)$
Error	$n-K$	SSE	$MSE = \frac{SSE}{n-K}$		
Total	$n-1$	$SSTOT$			

Veámos algunos ejemplos.

Ejemplo. Suponga que investiga una posible relación entre el tabaquismo y la frecuencia cardíaca. Suponga que tiene información de 4 grupos de individuos que van desde los no fumadores hasta los fumadores empedernidos. Cada grupo fue representado por seis sujetos. A lo largo de la fila inferior del Cuadro 9.2 se muestran las frecuencias cardíacas promedio calculadas para los sujetos de cada grupo tres minutos después de haber realizado una sesión de ejercicio físico sostenido.

Cuadro 9.2: Tabla Fumadores

Estadística	No fumadores	Fumador leve	Fumador moderado	Fumador empedernido
	69	55	66	91
	52	60	81	72
	71	78	70	81
	58	58	77	67
	59	62	57	95
	65	66	79	84
$\sum_{i=1}^{n_j} y_{ij}$	374	379	430	490
$\overline{Y}_j$	62.3	63.2	71.7	81.7

Asumiendo un nivel de significancia de $\alpha = 5\,\%$ determine si es posible aceptar la hipótesis de que no existe diferrencia estadística entre las medias de los grupos de individuos, i.e.,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Para rechazar esta hipótesis debemos verificar si:

$$F = \frac{SSTR/(K-1)}{SSE/(n-K)} \geq F_{[1-\alpha,K-1,n-K]}$$

Calculando los respectivos indicadores:

$$\overline{\mathbf{Y}} = \frac{1}{n}\sum_{j=1}^{K}\sum_{i=1}^{n_j} y_{ij} = \frac{1}{24}\sum_{j=1}^{4}\sum_{i=1}^{6} y_{ij} = \frac{374+379+430+490}{24} = 69.7$$

$$\begin{aligned} SSTR &= \sum_{j=1}^{K} n_j(\overline{Y}_j - \overline{\mathbf{Y}})^2 = \sum_{j=1}^{4} 6(\overline{Y}_j - \overline{\mathbf{Y}})^2 \\ &= 6(62.3-69.7) + \cdots + 6(81.7-69.7) = 1464.125 \end{aligned}$$

$$
\begin{aligned}
SSE &= \sum_{j=1}^{K}\sum_{i=1}^{n_j}(y_{ij}-\overline{Y}_j)^2 = \sum_{j=1}^{4}\sum_{i=1}^{6}(y_{ij}-\overline{Y}_j)^2 \\
&= [(69-62.3)^2+\cdots+(6562.3)^2]+\cdots \\
&\quad +[(91-81.7)^2+\cdots+(84-81.7)^2] = 1594.833
\end{aligned}
$$

De esta forma, la estadística F estará dada por:

$$
\begin{aligned}
F &= \frac{SSTR/(K-1)}{SSE/(n-K)} = \frac{1464.125/(4-1)}{1594.833/(24-4)} = 6.12 \\
F_{[1-\alpha,K-1,n-K]} &= F_{[0.95,4-1,24-4]} = 3.10
\end{aligned}
$$

Dado lo anterior, podemos no podemos aceptar la hipótesis nula. Estos datos apoyan la afirmación de que fumar influye en la frecuencia cardíaca de una persona.

Desde la perspectiva de una tabla ANOVA tendríamos:

Cuadro 9.3: Tabla ANOVA

Fuente	df	SS	MS	F	Prob.
Tratamiento	$4-1$	1464.125	488.04	6.12	0.004
Error	$20-4$	1594.833	79.74		
Total	$24-1$	3058.958			

9.4. Diseño del Análisis de Varianza para bloques

En cualquier experimento, reducir la magnitud del error del experimento es un objetivo deseable. Una σ^2 más pequeña suele ser una oportunidad de rechazar una hipótesis nula falsa. Existen 2 formas de reducir el error. Por un lado existe un mécanismo menos formal, el cual consiste en refinar el experimento minimizando error asociado a cada individuo. Por otro lado, existe un mécanismo más formal el cual permite redicir el error realizando experimentos aleatorios por bloques. La Figurra 9.1 muestra ejemplos de cómo realizar los experimentos por bloques.

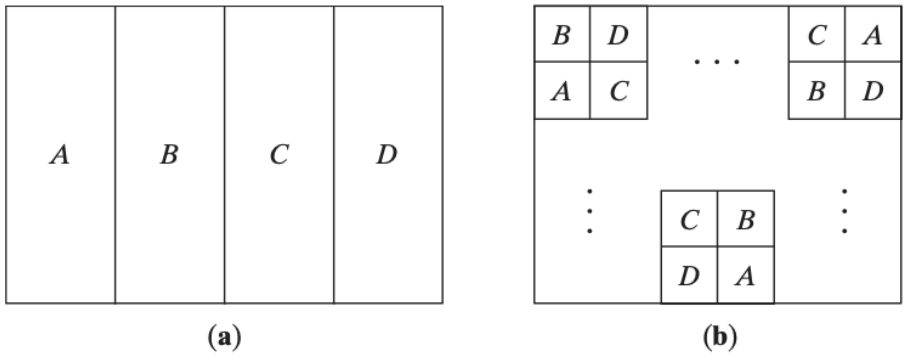

Figura 9.1: Dos diseños experimentales diferentes

9.5. La prueba F para diseño de bloques aleatorizados

Superficialmente, la estructura de bloques aleatorizados tiene una forma similar a K tratamientos en una muestra. Por otro lado, desde un punto de vista estadístico los bloques aleatorizados son fundamentalmente bloques de K muestras de datos.

En otras secciones hemos supuesto que las K muestras eran independientes. En este caso, las observaciones dentro de una regla de dedo (que comprende a un bloque) son independientes, a partir de que cada uno refleja alguna de las condiciones inherentes al bloque.

Por otro lado, desde el punto de vista estadístico, los bloques aleatorizados de datos son fundamentalmente diferentes de una muestra de K muestras. En ese caso, las K muestras son independientes. Aquí las observaciones dentro de una muestra dada (que comprende a un bloque) son independientes. Esta condición hace que el análisis de varianza se tenga que tratar de forma diferente.

En este caso, nuestro objetivo es probar que:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$

La misma que hemos planteado previamente. No obstante, el modelo matemático planteado es diferente, en el cual asummimos un término adicional.

Así, incorporamos al análisis el efecto del bloque i-ésimo β_i como:

$$Y_{ij} = \mu_j + \beta_i + \varepsilon_{ij} \tag{9.8}$$

Donde ε_{ij} se distrribuye de forma normal con media cero (0) y varianza σ^2 para $i = 1, 2, \ldots, b$ y $j = 1, 2, \ldots, K$. Es decir, mantenemos el supuesto del caso del análisis ANOVA de que las observaciones tienen una distribución normal.

De forma similar al caso de ANOVA, denotaremos a la media sobre todos los tratamientos o bloques como:

$$\mu = \frac{1}{K} \sum_{j=1}^{K} \mu_j \tag{9.9}$$

En este caso podemos utilizar un método similar al que seguimos en el caso de ANOVA. Para determinar los distintos tipos de sumas de los cuadrados de los errores o desviaciones. No obstante, necesitamos recalcularlos ya que el error en un componente de bloques aleatorizados las medidas reflejan tanto el efecto del bloque como el del error aleatorio.

Para abordar los dos requerimos estimar el conjunto de efectos por bloques: $\beta_1, \beta_2, \cdots, \beta_b$.

Sea $\overline{Y}_i = \frac{1}{K} \sum_{j=1}^{K} Y_{ij}$ la media de K observaciones en el i-ésimo bloque. Supongamos que los datos no contienen un error aleatorio, esto es, $\varepsilon_{ij} = 0, \forall i, j$. Entonces diremos:

$$\begin{aligned} \overline{Y}_i &= \frac{1}{K} \sum_{j=1}^{K} (\mu_j + \beta_i) \\ &= \frac{1}{K} \sum_{j=1}^{K} \mu_j + \frac{1}{K} \sum_{j=1}^{K} \beta_i \\ &= \mu + \beta_i \end{aligned}$$

Si $\overline{Y}$ es sustituido por el paramétro del cual es su estimador, entonces podemos expresar el efecto de bloque como:

$$\beta_i = \overline{Y}_i - \overline{Y}$$

Sabiendo lo anterior, podemos alterar la forma en que hemos calculado SSE para incorporar o separar el efecto del bloque:

$$\begin{aligned}
\sum_{i=1}^{b}\sum_{j=1}^{K}(Y_{ij}-\overline{Y}_{j})^2 &= \sum_{i=1}^{b}\sum_{j=1}^{K}[(Y_{ij}-\overline{Y}_{j})+(\overline{Y}_{i}-\overline{Y})-(\overline{Y}_{i}-\overline{Y})]^2 \\
&= \sum_{i=1}^{b}\sum_{j=1}^{K}[(\overline{Y}_{i}-\overline{Y})+(Y_{ij}-\overline{Y}_{j}-\overline{Y}_{i}+\overline{Y})]^2 \\
&= \sum_{i=1}^{b}\sum_{j=1}^{K}(\overline{Y}_{i}-\overline{Y})^2 \\
&\quad +\sum_{i=1}^{b}\sum_{j=1}^{K}(Y_{ij}-\overline{Y}_{j}-\overline{Y}_{i}+\overline{Y})^2 \\
&\quad +2\sum_{i=1}^{b}\sum_{j=1}^{K}(\overline{Y}_{i}-\overline{Y})(Y_{ij}-\overline{Y}_{j}-\overline{Y}_{i}+\overline{Y}) \\
&= \sum_{i=1}^{b}\sum_{j=1}^{K}(\overline{Y}_{i}-\overline{Y})^2 \\
&\quad +\sum_{i=1}^{b}\sum_{j=1}^{K}(Y_{ij}-\overline{Y}_{j}-\overline{Y}_{i}+\overline{Y})^2
\end{aligned}$$

En este caso hemos empleado un resultado para el térmmino:

$$\begin{aligned}
&\sum_{i=1}^{b}\sum_{j=1}^{K}(\overline{Y}_{i}-\overline{Y})(Y_{ij}-\overline{Y}_{j}-\overline{Y}_{i}+\overline{Y}) \\
= &\sum_{i=1}^{b}(\overline{Y}_{i}-\overline{Y})\sum_{j=1}^{K}(Y_{ij}-\overline{Y}_{j}-\overline{Y}_{i}+\overline{Y}) \\
= &\sum_{i=1}^{b}(\overline{Y}_{i}-\overline{Y})\left(\sum_{j=1}^{K}Y_{ij}-\sum_{j=1}^{K}\overline{Y}_{j}-\sum_{j=1}^{K}\overline{Y}_{i}+\sum_{j=1}^{K}\overline{Y}\right) \\
= &\sum_{i=1}^{b}(\overline{Y}_{i}-\overline{Y})\left(K\overline{Y}_{i}-K\overline{Y}-K\overline{Y}_{i}+K\overline{Y}\right) \\
= &\ 0
\end{aligned}$$

En conclusión, la SSTOT se puede descomponer en la suma de dos cuadrados: i) la suma de cuadrados de bloques (SSB) y ii) la suma de errores cuadrados (SSE) que estará dado por:

$$\begin{aligned} SSB &= \sum_{i=1}^{b}\sum_{j=1}^{K}(\overline{Y}_i - \overline{Y})^2 \\ SSE &= \sum_{i=1}^{b}\sum_{j=1}^{K}(Y_{ij} - \overline{Y}_j - \overline{Y}_i + \overline{Y})^2 \end{aligned}$$

Así, estableceremos que:

$$\begin{aligned} SSTOT &= \sum_{i=1}^{b}\sum_{j=1}^{K}(Y_{ij} - \overline{Y})^2 \\ SSTR &= \sum_{i=1}^{b}\sum_{j=1}^{K}(\overline{Y}_j - \overline{Y})^2 \end{aligned}$$

Una propiedad similar al caso de ANOVA es que:

$$SSTOT = SSTR + SSB + SSE$$

En el Cuadro 9.4 reportamos una tabla que ilustra las distintas relaciones que existen en la medias por bloque y tratamiento.

Cuadro 9.4: Tabla de bloques y diseño experimental

	Nivel de tratammiento				Total de	Media de	Efecto verdadero
	1	2	...	K	bloques	bloques	del bloque
Bloques 1	Y_{11}	Y_{12}	$\cdots$	Y_{1K}	T_1	$\overline{Y}_1$	β_1
Bloques 2	Y_{21}	Y_{22}	$\cdots$	Y_{2K}	T_2	$\overline{Y}_2$	β_2
Bloques $\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Bloques b	Y_{b1}	Y_{b2}	$\cdots$	Y_{bK}	T_b	$\overline{Y}_b$	β_b
Total de muestra	T_1	T_2	$\cdots$	T_K	T		
Media de muestra	$\overline{Y}_1$	$\overline{Y}_2$	$\cdots$	$\overline{Y}_K$		$\overline{Y}$	
Verdadera media	μ_1	μ_2	$\cdots$	μ_K			

Teorema 9.1 *Supongamos K niveles de tratamiento, con $\mu_1, \mu_2, \cdots, \mu_K$ que son medias de un conjunto de b bloques, donde los efectos del bloque son $\beta_1, \beta_2, \cdots, \beta_b$. Entonces:*

- *Cuando $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ es cierta, entonces:*

$$\frac{SSTR}{\sigma^2} \sim \chi^2_{[K-1]}$$

- *Cuando $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b$ es cierta, entonces:*

$$\frac{SSB}{\sigma^2} \sim \chi^2_{[b-1]}$$

- *Si μ_j's y/o β_i's son iguales, entonces:*

$$\frac{SSE}{\sigma^2} \sim \chi^2_{[(b-1)(K-1)]}$$

Teorema 9.2 *Supongamos que K tratamientos con medias $\mu_1, \mu_2, \cdots, \mu_K$ son medias de los bloques, entonces:*

- *Si $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ es cierta:*

$$F = \frac{SSTR/(K-1)}{SSE/(b-1)(K-1)} \sim F_{[K-1,(b-1)(K-1)]}$$

- *Al nivel de significancia α, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ será rechazada si:*

$$F \geq F_{[1-\alpha,k-1,(b-1)(K-1)]}$$

Teorema 9.3 *Supongamos que k tratamientos con medias sobre un conjunto de b bloques, dado el efecto del bloque son $\beta_1, \beta_2, \cdots, \beta_b$ entonces:*

- *Si $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b$ es cierta:*

$$F = \frac{SSB/(b-1)}{SSE/(b-1)(K-1)} \sim F_{[b-1,(b-1)(K-1)]}$$

- *Al nivel de significancia α, $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b$ será rechazada si:*

$$F \geq F_{[1-\alpha,b-1,(b-1)(K-1)]}$$

Cuadro 9.5: Tabla ANOVA para Bloques

Fuente	df	SS	MS	F	Prob.
Tratamiento	$K-1$	$SSTR$	$\frac{SSTR}{K-1}$	$\frac{SSTR/(K-1)}{SSE/(b-1)(K-1)}$	$P(F_{[K-1,(b-1)(K-1)]} \geq$
Bloques	$b-1$	SSB	$\frac{SSB}{b-1}$	$\frac{SSB/(b-1)}{SSE/(b-1)(K-1)}$	$P(F_{[b-1,(b-1)(K-1)]} \geq$
Error	$(b-1)(K-1)$	SSE	$\frac{SSE}{(b-1)(K-1)}$		
Total	$n-1$	$SSTOT$			

Los cálculos del análisis de varianza suelen presentarse como tablas. Altamente estructuradas, estas tablas son especialmente útiles en la implementación de pruebas estadísticas como las hipótesis: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ y $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b$. Así, la tabla sepuede construir como se muestra en el Cuadro 9.5.

Antes de ver un ejemplo, plantearemos que podemos utilizar el siguiente conjunto de fórmulas:

$$\begin{aligned} SSTR &= \sum_{j=1}^{K} \frac{T_j^2}{b} - \frac{T^2}{bK} \\ SSB &= \sum_{i=1}^{b} \frac{T_i^2}{K} - \frac{T^2}{bK} \\ SSTOT &= \sum_{i=1}^{b}\sum_{j=1}^{K} Y_{ij}^2 - \frac{T^2}{bK} \\ SSE &= SSTOT - SSTR - SSB \end{aligned}$$

Ejemplo. Supongamos que tenemos una muestra de la concentración cierto compuesto químico en la fermentación del vino. Asumamos que consideramos cuatro vinos diferentes y que medimos la concentración con dos procedimientos diferentes: una técnica de espectroscopía infrarroja (EI) y otra basada en un proceso de aireación-oxidación (AO). El Cuadro 9.6 muestra los resultados.

Al respecto pruebe si las medias de con concentración derivadas de los dos procedimientos (μ_{EI} y μ_{AO}, respectivamente) son iguales:

$$\begin{aligned} H_0 &: \mu_{EI} = \mu_{AO} \\ H_1 &: \mu_{EI} \neq \mu_{AO} \end{aligned}$$

Cuadro 9.6: Medidas de concentración (mg / ml)

	EI	AO
Vino blanco 1	112.9	115.1
Vino blanco 2	123.1	125.6
Vino tinto 1	135.2	132.4
Vino tinto 2	140.2	143.7

Para reponder consideremos un nivel de significancia del 95 %.
El Cuadro 9.7 muestra los totales por bloque de tratamiento.

Cuadro 9.7: Tratamientos

	EI	AO	Totales (T_i)
Vino blanco 1	112.9	115.1	228.0
Vino blanco 2	123.1	125.6	248.7
Vino tinto 1	135.2	132.4	267.6
Vino tinto 2	140.2	143.7	283.9
Totales (T_j)	511.4	516.8	1028.2

Usando la información anterior podemos terminar que las diferentes sumas de cuadrados serán:

$$\begin{aligned} SSTR &= \sum_{j=1}^{K} \frac{T_j^2}{b} - \frac{T^2}{bK} \\ &= \frac{511.4^2}{4} + \frac{516.8^2}{4} - \frac{1028.2^2}{8} \\ &= 3.645 \end{aligned}$$

$$\begin{aligned} SSB &= \sum_{i=1}^{b} \frac{T_i^2}{K} - \frac{T^2}{bK} \\ &= \frac{228.0^2}{2} + \frac{248.7^2}{2} + \frac{267.67^2}{2} + \frac{283.9^2}{2} - \frac{1028.2^2}{8} \\ &= 872.925 \end{aligned}$$

$$
\begin{aligned}
SSTOT &= \sum_{i=1}^{b}\sum_{j=1}^{K} Y_{ij}^2 - \frac{1028.2^2}{8} \\
&= (112.9)^2 + (123.1)^2 + \cdots + (143.7)^2 - \frac{1028.2^2}{8} \\
&= 888.515
\end{aligned}
$$

$$
\begin{aligned}
SSE &= SSTOT - SSTR - SSB \\
&= 888.515 - 3.645 - 872.925 \\
&= 11.945
\end{aligned}
$$

El Cuadro 9.8 muestra los resultados en forma de tabla ANOVA.

Cuadro 9.8: Tabla ANOVA para Bloques

Fuente	df	SS	MS	F
Tratamiento	1	3.645	$\frac{3.645}{1} = 3.645$	$\frac{3.645}{3.982} = 0.92$
Bloques	3	872.925	$\frac{872.925}{3} = 290.975$	$\frac{290.975}{3.982} = 73.07$
Error	3	11.945	$\frac{11.945}{3} = 3.982$	
Total	7	888.515		

Los valores críticos para las respectivas hipóttesis nulas son: i) Igualdad de medias: $F_{[1,3]} = 10.13 > F = 0.92$ y ii) Igualdad de efecto de bloque $F_{[3,3]} = 9.28 < F = 73.07$

9.6. Análisis de datos categóricos

9.6.1. Introducción y motivación

Existen múltiples experimentos en los que las mediciones analizadas son cualitativas o categóricas, más que cuantitativas. En este caso hablamos de experimentos multinomiales que son:

1. Experimentos que consisten en ensayos idénticos.

2. El resultadode cada ensayo cae en exactamente K categorías.

3. La probabilidad de que el resultado de un solo ensayo caiga en una categoría i es: p_i; $i = 1, 2, \ldots, K$. Donde: $p_1 + p_2 + \cdots + p_K = 1$.

4. Los ensayos son idependientes.

5. Estamos interesados en $n_1, n_2, \ldots, n_K$, donde n_i, $i = 1, 2, \ldots, K$, es igual al número de ensayos para los cuales el resultado cae en la categoría i y $n = n_1 + n_2 + \ldots + n_K$.

9.6.2. Prueba de Ji Cuadrado

En un modelo multinomial cada una de las n_i tiene distribución binomial con parametros n y p_i. Entonces:

$$\mathbb{E}[n_i] = np_i; \text{ parar } i = 1, 2, \ldots, K$$

En nuestro caso tenemos probabilidades por grupo $p_1, p_2, \ldots, p_K$ y podemos calcular el valor esperado para cada grupo o categoría. Entonces nuestra hipótesis nula es que los datos siguen una distribución con probabilidades por categoría que hemos planteado.

Si nuestra hipótesis es verdadera, cada n_i no debe desviarse de su valor esperado np_i, $i = 1, 2, \ldots, K$. Entonces podemos calcular K desviaciones:

$$n_i - \mathbb{E}[n_i] = n_i - np_i; \text{ para } i = 1, 2, \ldots, K$$

En 1900, PEarson propuso la siguiente estadística de prueba:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{K} \frac{(n_i - \mathbb{E}[n_i])^2}{\mathbb{E}[n_i]} \\ &= \sum_{i=1}^{K} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{[K-1]} \end{aligned}$$

Ejemplo. Supongamos que durante 50 semanas ($n = 50$) recolectamos el número de accidentes Y en un crucero de la CDMX, con los resultados que se muestran en el Cuadro 9.9. Pruebe la hipótesis de que la variable aleatoria Y tiene una distribución de Poisson, suponiendo que las observaciones son independientes entre cada semana observada y que $\lambda = 0.48$. Use un $\alpha = 0.05$.

Cuadro 9.9: Número de accidentes por semana

$Y = y$	Fequency
0	32
1	12
2	6
3 o más	0

Para dar solución a este problema partamos la función de distribución Poisson:

$$p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}; \text{ para } y = 0, 1, 2, \ldots$$

Las probabilidades para las 4 categorías estarán dadas por:

$$\begin{aligned} p_1 &= P(Y = 0) = e^{-\lambda} = e^{-0.48} = 0.619 \\ p_2 &= P(Y = 1) = \lambda e^{-\lambda} = 0.48 e^{-0.48} = 0.297 \\ p_3 &= P(Y = 2) = \frac{\lambda^2 e^{-\lambda}}{2!} = \frac{0.48^2 e^{-0.48}}{2!} = 0.071 \\ p_4 &= P(Y \geq 2) = 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2 e^{-\lambda}}{2!} \\ &= 1 - 0.619 - 0.297 - 0.071 = 0.013 \end{aligned}$$

Dicho esto podemos encontar cada uno de los varloes esperados $\mathbb{E}[n_i] = np_i$. Los calculos se muestran en el Cuadro 9.10.

Cuadro 9.10: Número de accidentes por semana y calculos relacionados

$Y = y$	Fequency (n_i)	p_i	$\mathbb{E}[n_i] = np_i$	$(n_i - \mathbb{E}[n_i])^2/\mathbb{E}[n_i]$
0	32	0.619	30.95	0.036
1	12	0.297	14.85	0.547
2	6	0.071	3.55	1.691
3 o más	0	0.013	0.65	0.650
Sumas	50	1	50	2.923

Ahora, en este caso tenemos que $\chi^2_{[3]} = 7.81 > \chi^2 = 2.923$, por lo que no podemos rechazar la hipótesis nula de que los datos sigan una distribución Poisson.

9.7. Consideraciones para el diseño de experimentos

9.7.1. Elementos que se relaciona con el tamaño de la muestra

Si deseamos comparar dos poblaciones con base en un total de n observaciones, ¿cuántas observaciones deberíamos tomar de cada población?

Generalmente, dentro de experimentos o diseño de experimentos un tema recurrente es el del método de muestreo para reducir la variación del experimento y, por lo tanto, para adquirir una cantidad mínima de información al mínimo costo posible; considerando información suficiente para el análisis (por ejemplo, análisis de información de muestras pareadas).

9.7.2. Diseño de experimentos para aumentar la precisión

Para ilustrar este punto utilizaremos un ejemplo. ¿Cuántas observaciones se requieren seleccionar de las poblaciones 1 y 2, es decir, n_1 y n_2 ($n_1 + n_2 = n$) para maximizar la información de los datos necesarios para analizar la diferencia de medias $\mu_1 - \mu_2$?

Para responder retomemos el concepto de error estándar que utilizamos en intervalos de confianza y en pruebas de hipótesis cuando conociamos la varianzas σ_1^2 y σ_2^2:

$$\sigma_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Así, cuanto menor sea $\sigma_{\overline{Y}_1 - \overline{Y}_2}$, menor será el correspondiente error del cálculo y mayor la cantidad de información de la muestra.

9.7.3. El experimento de observaciones pareadas

En muchos experimentos las muestran son pareadas más que independientes. En esta situaciones ocurren comúnmente cuando el análisis se basa en el seguimiento de una misma observación o unidad de muestreo en diferentes momentos. Por ejemplo, imaginemos que buscamos obtener el peso

de una misma persona antes y después de que participe en un programa de reducción de peso.

Otro ejemplo puede ser un experimento médico en el que pareamos individuos que son del mismo género, pesos y edades similares. Así, un inndividuo de cada par es seleccionado al azar para que reciba uno de dos medicamentos competidores para el tratamiento de algún padecimiento, mientras que el otro inndividuo recibe el otro medicamento.

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\overline{\mathbf{Y}})^2 &= \mathbf{Y}'\mathbf{M}^0\mathbf{Y} \\
\sum_{i=1}^{n}(y_i-\overline{\mathbf{Y}}) &= \sum_{i=1}^{n}(y_i^2-2y_i\overline{\mathbf{Y}}+\overline{\mathbf{Y}}^2) \\
&= \sum_{i=1}^{n}y_i^2-2\overline{\mathbf{Y}}\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}\overline{\mathbf{Y}}^2 \\
&= \sum_{i=1}^{n}y_i^2-2\frac{\sum_{i=1}^{n}y_i}{n}\sum_{i=1}^{n}y_i+n\overline{\mathbf{Y}}^2 \\
&= \sum_{i=1}^{n}y_i^2-\frac{2}{n}\left(\sum_{i=1}^{n}y_i\right)^2+n\left(\frac{\sum_{i=1}^{n}y_i}{n}\right)^2 \\
&= \sum_{i=1}^{n}y_i^2-\frac{2}{n}\left(\sum_{i=1}^{n}y_i\right)^2+\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2 \\
&= \sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{e}'\mathbf{e} &= (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y}-2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
&= 18310.63-2\begin{bmatrix}559.60 & 7375.44 & 324972.70\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix} \\
&\quad +\begin{bmatrix}\beta_1 & \beta_2 & \beta_3\end{bmatrix}\begin{bmatrix}25.00 & 219.00 & 9932.00\\ 219.00 & 3055.00 & 129099.00\\ 9932.00 & 129099.00 & 6402888.00\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
Cov(x,y) &= \mathbb{E}[(x-\mathbb{E}[x])(y-\mathbb{E}[y])] \\
Cov(p_i,u_i) &= \mathbb{E}[(p_i-\mathbb{E}[p_i])(u_i-\mathbb{E}[u_i])] \\
&= \mathbb{E}\left[\left(\frac{\beta_0-\alpha_0}{\alpha_1-\beta_1}+\frac{v_i-u_i}{\alpha_1-\beta_1}-\frac{\beta_0-\alpha_0}{\alpha_1-\beta_1}\right)(u_i)\right] \\
&= \mathbb{E}\left[\left(\frac{v_i-u_i}{\alpha_1-\beta_1}\right)(u_i)\right] \\
&= \ldots
\end{aligned}
$$

10

Regresión Lineal

10.1. Introducción y motivación

En esta sección analizaremos el concepto de regresión lineal. El cual está basado en la observación cotidiana de cómo es la relación entre dos o más variables, una dependiente, y, y otras independientes (o dadas), x, dado que hemos obtenido una muestra aleatoria. Por ejemplo, ¿qué determina el cáncer pulmonar en las personas? Algunas respuestas son: fumar, dieta, contaminación, condiciones genéticas, etc.

Otros ejemplos son, si la latitud explica algunas condiciones climáticas. Las figuras 10.1, 10.2, 10.3, 10.4 muestran la relación geometrica entre la latitud y diversas variables del clima para una muestra de 500 ciudades en la fecha que se señala.

10.2. El concepto de regresión entre dos variables

Definición 10.1 *Sea $f(Y|x)$ la función de densidad de probabilidad de una variable aleatoria Y dado un valor de $X = x$, y sea $\mathbb{E}[Y|x]$ el valor esperado asociado con la función $f(Y|x)$. La función:*

$$\mu_{Y|x} = \mathbb{E}[Y|x]$$

se denomina como la curva de regresión de Y en $X = x$. De forma similar podemos definir la curva de regresión de X en $Y = y$.

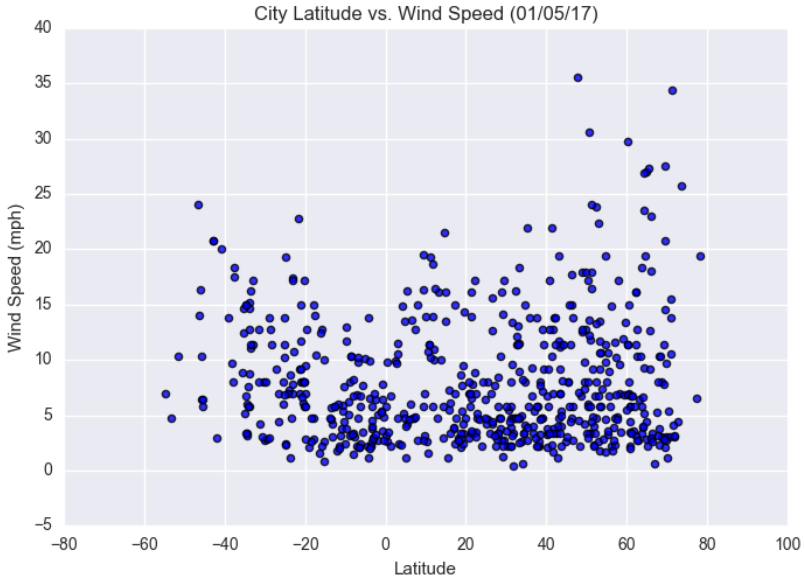

Figura 10.1: Relación entre Latitud y Velocidad del Viento

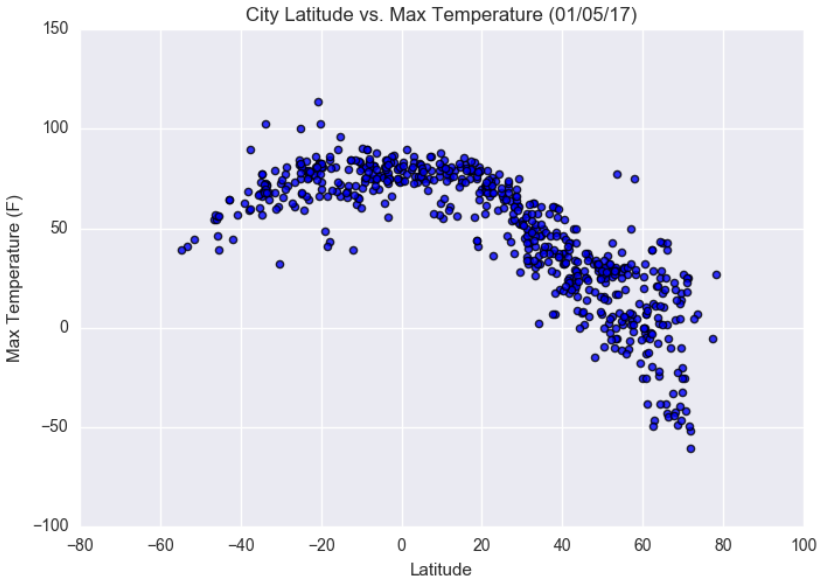

Figura 10.2: Relación entre Latitud y Temperatura

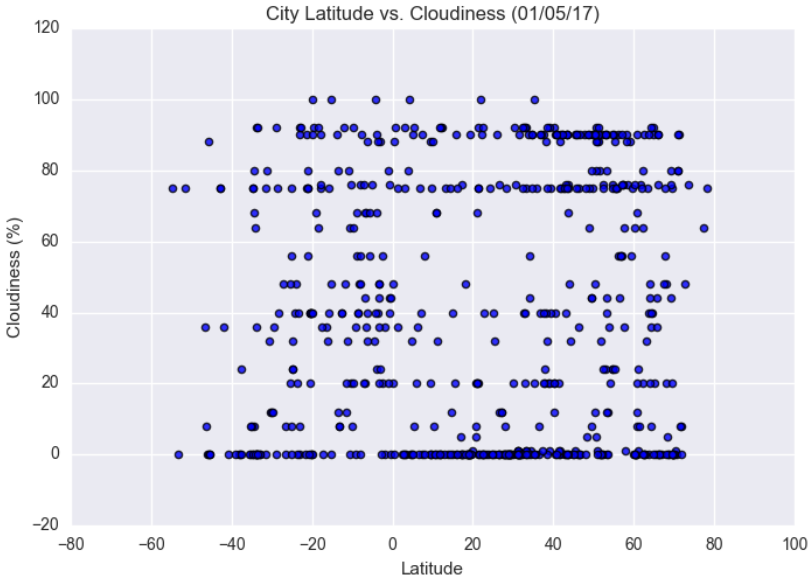

Figura 10.3: Relación entre Latitud y Nubocidad

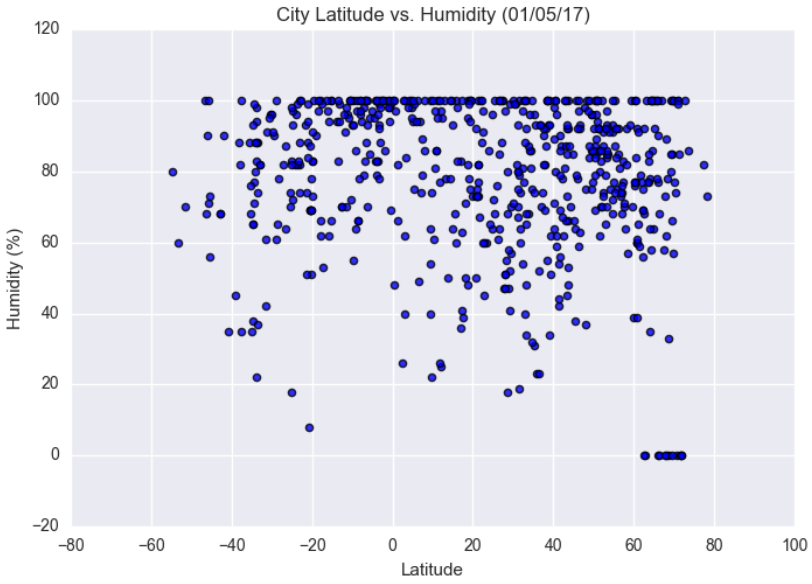

Figura 10.4: Relación entre Latitud y Humedad

Ejemplo. Supongamos una función dada por:

$$f(Y|x) = \begin{cases} \frac{x+y}{x+\frac{1}{2}} & \text{para } 0 \leq y \leq 1, 0 \leq x \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos la función de regresión de Y en $X = x$. Para ello primero determinemos si $f(Y|x)$ es una función de densidad de probabilidad condicional:

$$\begin{aligned} \int_D f(Y|x)dy &= \int_0^1 \frac{x+y}{x+\frac{1}{2}}dy = \int_0^1 \frac{x}{x+\frac{1}{2}}dy + \int_0^1 \frac{y}{x+\frac{1}{2}}dy \\ &= \frac{xy}{x+\frac{1}{2}}\bigg|_0^1 + \frac{y^2/2}{x+\frac{1}{2}}\bigg|_0^1 = \frac{x}{x+\frac{1}{2}} + \frac{\frac{1}{2}}{x+\frac{1}{2}} \\ &= \frac{x+\frac{1}{2}}{x+\frac{1}{2}} = 1 \end{aligned}$$

Por lo tanto, observamos que $f(Y|x)$ es una función de densidad de probabilidad. Ahora determinemos $\mu_{Y|x}$:

$$\begin{aligned} \mu_{Y|x} &= \mathbb{E}[Y|x] = \int_0^1 y\frac{x+y}{x+\frac{1}{2}}dy = \int_0^1 \frac{xy+y^2}{x+\frac{1}{2}}dy = \frac{\frac{xy^2}{2}+\frac{y^3}{3}}{x+\frac{1}{2}}\bigg|_0^1 \\ &= \frac{\frac{x}{2}+\frac{1}{3}}{x+\frac{1}{2}} = \frac{\frac{3x}{2}+1}{3x+\frac{3}{2}} = \frac{3x+2}{6x+3} \end{aligned}$$

Podemos platear una definición más amplia:

Definición 10.2 *Sea $f(x,y)$ una función de densidad conjunta de las variables aleatorias X y Y, la función de regresión consiste en determinar la densidad condicional de Y dado $X = x$, la cual resulta de evaluar la integral:*

$$\mu_{Y|x} = \mathbb{E}[Y|x] = \int_D y \cdot f(Y|x)dy$$

El resultado de la integral es denominado como ecuación de regresión de Y en X. De forma similar, la regresión de X en Y estará dada por:

$$\mu_{X|y} = \mathbb{E}[X|y] = \int_D x \cdot f(X|y)dx$$

Donde las funciones de densidad condicional resultan de:

$$\begin{aligned} f(Y|x) &= \frac{f(x,y)}{g(x)} \\ f(X|y) &= \frac{f(x,y)}{g(y)} \end{aligned}$$

Y las funciones marginales son resultado de:

$$\begin{aligned} g(x) &= \int_{D_y} f(x,y)dy \\ g(y) &= \int_{D_x} f(x,y)dx \end{aligned}$$

Ejemplo. Dadas dos variables aleatorias X y Y con función de densidad conjuntandada por:

$$f(x,y) = \begin{cases} x \cdot e^{-x(1+y)} & \text{para } x > 0, y > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos la ecuación de regresión de Y en X. Para ello requerimos de:

$$\begin{aligned} g(x) &= \int_{D_y} f(x,y)dy = \int_{D_y} x \cdot e^{-x(1+y)}dy = -e^{-x(1+y)}\Big|_0^\infty \\ &= \begin{cases} e^{-x} & \text{para } x > 0 \\ 0 & \text{en cualquier otro caso} \end{cases} \end{aligned}$$

De esta forma podemos determinar la función de densidad condicional como:

$$f(Y|x) = \frac{f(x,y)}{g(x)} = \frac{x \cdot e^{-x(1+y)}}{e^{-x}} = \begin{cases} x \cdot e^{-xy} & \text{para } y > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Podemos observar que es una función de densidad ya que:

$$\int_{D_y} f(Y|x)dy = \int_0^\infty x \cdot e^{-xy}dy = e^{-xy}\Big|_0^\infty = 1$$

Finalmente, la ecuación de regresión de Y en X estará dada por:

$$\begin{aligned}\mu_{Y|x} &= \mathbb{E}[Y|x] = \int_{D_y} y \cdot f(Y|x)dy = \int_0^\infty y \cdot x \cdot e^{-xy}dy \\ &= -y \cdot e^{-xy}\Big|_0^\infty + \int_0^\infty e^{-xy}dy = -y \cdot e^{-xy}\Big|_0^\infty - \frac{1}{x}e^{-xy}\Big|_0^\infty \\ &= \begin{cases} \frac{1}{x} & \text{para } x > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}\end{aligned}$$

Ejemplo. Supongamos una función de densidad conjunta de X_1, X_2, X_3 variables aleatorias dada por:

$$f(x_1, x_2, x_3) = \begin{cases} (x_1 + x_2)e^{-x_3} & \text{para } 1 > x_1 > 0, 1 > x_2 > 0, x_3 > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos la ecuación de regresión de X_2 en X_1 y X_3. Para ello requerimos de la función de densidad marginal dada por:

$$\begin{aligned}g(x_1, x_3) &= \int_{D_{x_2}} f(x_1, x_2, x_3)dx_2 = \int_0^1 (x_1 + x_2)e^{-x_3}dx_2 \\ &= \left(x_1x_2 + \frac{x_2^2}{2}\right)e^{-x_3}\Big|_0^1 \\ &= \begin{cases} \left(x_1 + \frac{1}{2}\right)e^{-x_3} & \text{para } 1 > x_1 > 0, x_3 > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}\end{aligned}$$

Determinemos entonces la función de densidad condicional, la cual estará dada por:

$$\begin{aligned}f(X_2|x_1, X_3) &= \frac{f(x_1, x_2, x_3)}{g(x_1, x_3)} = \frac{(x_1 + x_2)e^{-x_3}}{\left(x_1 + \frac{1}{2}\right)e^{-x_3}} \\ &= \begin{cases} \frac{x_1+x_2}{x_1+\frac{1}{2}} & \text{para } 1 > x_1 > 0, 1 > x_2 > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}\end{aligned}$$

Entoncees, la ecuación de regresión estará dada por:

$$\begin{aligned}\mu_{X_2|x_1,x_3} &= \int_{D_{x_2}} x_2 \cdot f(X_2|x_1, X_3)dx_2 = \int_0^1 x_2\frac{x_1 + x_2}{x_1 + \frac{1}{2}}dx_2 = \frac{\frac{x_1x_2^2}{2} + \frac{x_2^3}{3}}{x_1 + \frac{1}{2}}\Big|_0^1 \\ &= \frac{\frac{x_1}{2} + \frac{1}{3}}{x_1 + \frac{1}{2}} = \begin{cases} \frac{x_1+\frac{2}{3}}{2x_1+1} & \text{para } 1 > x_1 > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}\end{aligned}$$

10.3. Modelo de regresión lineal

10.3.1. Supuestos del modelo de regresión lineal

Un caso especial es el de la regresión lineal. Lo que hasta ahora hemos visto es el caso general en el que $\mu_{Y|x} = \mathbb{E}[Y|x]$, el cual modificaremos bajo los siguientes supuestos – **Supuestos del Modelo Lineal**:

1. **Normalidad**. $f(Y|x)$ es una función de densidad de probabilidad normal $\forall x$

2. **Homocedásticidad**. La desviación estánndar, σ, asociada con $f(Y|x)$ es la misma $\forall x$

3. **Linealidad**. Las medias de todas las distribuciones condicionales son una combinación lineal dada por:

   $$\mu_{Y|x} = \mathbb{E}[Y|x] = \beta_0 + \beta_1 x$$

   Donde β_0 y β_1 son parámetros por determinar, como veremos más adelante

4. Todas las distribuciones condicionales representan variables aleatorias independientes como se ilustra en la Figura 10.5.

Ahora enunciaremos un teorema relvante para el caso del modelo lineal.

Teorema 10.1 *Sean Y y X dos variables aleatorias que cumplen con los supuestos del modelo lineal. Si la regresión de Y en X es linneal, entonces:*

$$\mu_{Y|x} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

Si la regresión de X en Y es lineal, entonces:

$$\mu_{X|y} = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y)$$

Para demostrar el teorema anterior partamos de recordar que la media condicional estará dada por:

$$\mu_{Y|x} = \int_{D_y} y \cdot f(Y|x)dy = \beta_0 + \beta_1 x \tag{10.1}$$

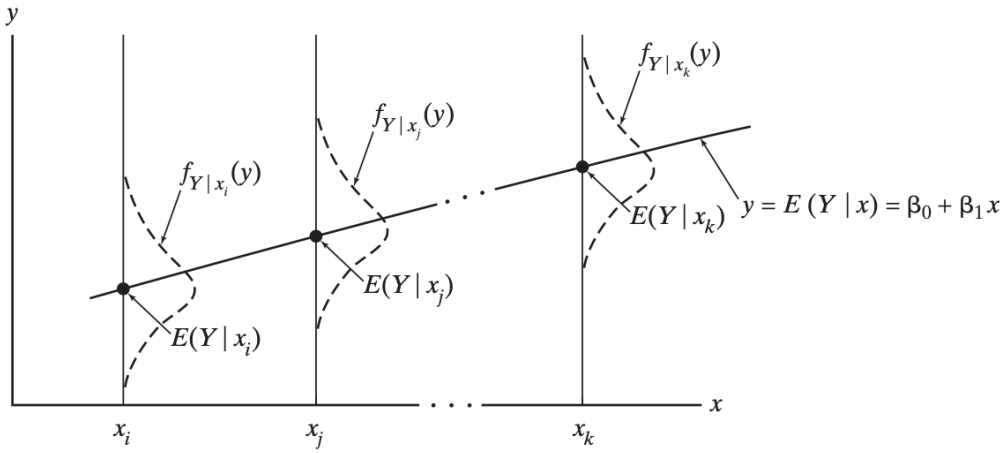

Figura 10.5: Supuestos del modelo lineal (Retomado de Larsen (2012, p 545)Larsen y Marx 2012), donde se muestra una notación con las siguientes equivalencias $f_{Y|x}(y) = f(Y|x)$ y $y = \mu_{Y|x} = \mathbb{E}[Y|x]$

Multiplicando cada uno de los lados de la última parte de la ecuación (10.1) por $g(x)$, la función de densidad de probabilidad marginal, e integrando ambos lados tenemos lo siguiente:

$$\begin{aligned}
\int_{D_x}\int_{D_y} y \cdot f(Y|x)g(x)dydx &= \beta_0 \int_{D_x} g(x)dx + \beta_1 \int_{D_x} x \cdot g(x)dx \\
\iff \int_{D_x}\int_{D_y} y \cdot f(x,y)dydx &= \beta_0 + \beta_1\mu_X \\
\iff \int_{D_y}\int_{D_x} y \cdot f(x,y)dxdy &= \beta_0 + \beta_1\mu_X \\
\iff \int_{D_y} y \cdot g(y)dy &= \beta_0 + \beta_1\mu_X \\
\iff \mu_Y &= \beta_0 + \beta_1\mu_X \qquad (10.2)
\end{aligned}$$

Retomando la ecuación (10.1) y multiplicando por un factor dado por

$x \cdot g(x)$ e integrando, tenemos:

$$\begin{aligned}
\int_{D_x}\int_{D_y} y \cdot f(Y|x) \cdot x \cdot g(x) dy dx &= \beta_0 \int_{D_x} x \cdot g(x) dx + \beta_1 \int_{D_x} x^2 \cdot g(x) dx \\
\iff \int_{D_x}\int_{D_y} x \cdot y \cdot f(x,y) dy dx &= \beta_0 \mu_X + \beta_1 \mathbb{E}[X^2] \\
\iff \mathbb{E}[X \cdot Y] &= \beta_0 \mu_X + \beta_1 \mathbb{E}[X^2] \qquad (10.3)
\end{aligned}$$

Aquí recordamos un par de resultados que conocemos:

$$Cov(X,Y) = \mathbb{E}[X \cdot Y] - \mathbb{E}[X]\mathbb{E}[Y] = \sigma_{XY} = \mathbb{E}[X \cdot Y] - \mu_X \mu_Y$$

$$\mathbb{E}[X \cdot Y] = \sigma_{XY} + \mu_X \mu_Y$$

$$Var(X) = \mathbb{E}[X^2] - \mu_X^2 = \sigma_X^2$$

$$\mathbb{E}[X^2] = \sigma_X^2 + \mu_X^2$$

Retomando las ecuaciones (10.2) y (10.3) y los resultados mostrados tenemos que resolver el siguiente sistema de ecuaciones:

$$\begin{aligned}
\mu_Y &= \beta_0 + \beta_1 \mu_X \\
\sigma_{XY} + \mu_X \mu_Y &= \beta_0 \mu_X + \beta_1 (\sigma_X^2 + \mu_X^2)
\end{aligned}$$

De la primera despejamos:

$$\beta_0 = \mu_Y - \beta_1 \mu_X$$

Sustituimos en la segunda para encontrar:

$$\begin{aligned}
\sigma_{XY} + \mu_X \mu_Y &= (\mu_Y - \beta_1 \mu_X)\mu_X + \beta_1(\sigma_X^2 + \mu_X^2) \\
&= \mu_X \mu_Y - \beta_1 \mu_X^2 + \beta_1 \sigma_X^2 + \beta_1 \mu_X^2 \\
\sigma_{XY} &= \beta_1 \sigma_X^2
\end{aligned}$$

De donde:

$$\beta_1 = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_X \sigma_Y}{\sigma_X \sigma_Y}\frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}\frac{\sigma_X \sigma_Y}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X}$$

Y:

$$\beta_0 = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$$

Sustituyendo en la ecuación de regresión:

$$\begin{aligned}\mu_{Y|x} &= \beta_0 + \beta_1 x \\ &= \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X + \rho\frac{\sigma_Y}{\sigma_X}x \\ &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)\end{aligned}$$

Con lo cual hemos demotrado la primera parte del teorema. La segunda parte se demustra de forma simlar.

10.3.2. Estimación de los parámetros del modelo lineal bivariado

Antes de plantear el proceso de estimación respondamos ¿cuáles son los parámetros $\boldsymbol{\theta'} = (\theta_1, \theta_2, \ldots, \theta_K)'$ del modelo lineal dado por $\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$? Así, $\boldsymbol{\theta'} = (\beta_0, \beta_1, \sigma^2)'$. Dado que asumimos que Y tienen una distribución de probabilidad, en este caso normal, podemos determinar los parámetros por medio del uso de un procedimiento de máxima verosimilitud.

Sean $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ un conjunto de datos que satisfacen los supuestos del modelo lineal, entonces podemos establecer la siguiente transformación Z y la función de verosimilitud:

$$Z_i = \frac{y_i - \mathbb{E}[Y|x_i]}{\sigma}$$

$$\begin{aligned}L &= \Pi_{i=1}^n f(Y_i|x_i) = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}Z_i^2} = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \mathbb{E}[Y|x_i]}{\sigma}\right)^2} \\ &= \Pi_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x}{\sigma}\right)^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} \qquad (10.4)\end{aligned}$$

Tomando el logaritmo de la ecuación (10.4) tenemos:

$$
\begin{aligned}
ln(L) &= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2 \quad (10.5)\\
-2ln(L) &= n \cdot ln(2\pi) + n \cdot ln(\sigma^2) + \sum_{i=1}^{n}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2 \quad (10.6)
\end{aligned}
$$

Ambas ecuaciones (10.5) y (10.6) son equivalentes. La ecuación (10.6) es conocida por ser la expresión de la forma log-normal o una expresión de Chi cuadrado. De la ecuación (10.5) obtendremos el siguiente vector gradiente:

$$
\frac{\partial ln(L)}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial ln(L)}{\partial \beta_0} \\ \frac{\partial ln(L)}{\partial \beta_1} \\ \frac{\partial ln(L)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i \\ -\frac{n}{2\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{bmatrix}
$$

Así, podemos obtener las siguientes expresiones para los parámetros:

$$
\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) &= 0\\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\beta_0 - \sum_{i=1}^{n}\beta_1 x_i &= 0\\
\sum_{i=1}^{n} y_i - n\beta_0 - \sum_{i=1}^{n}\beta_1 x_i &= 0\\
\beta_0 &= \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} \quad (10.7)
\end{aligned}
$$

Para la segunda ecuación del gradiente y utilizando 10.7 tenemos:

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i &= 0\\
\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n}\beta_0 x_i - \sum_{i=1}^{n}\beta_1 x_i^2 &= 0\\
\sum_{i=1}^{n} y_i x_i - \beta_0\sum_{i=1}^{n} x_i - \beta_1\sum_{i=1}^{n} x_i^2 &= 0\\
\sum_{i=1}^{n} y_i x_i - \left(\frac{\sum_{i=1}^{n} y_i - \beta_1\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i - \beta_1\sum_{i=1}^{n} x_i^2 &= 0\\
n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i\sum_{i=1}^{n} x_i - \beta_1 n\sum_{i=1}^{n} x_i^2 &= 0\\
-\sum_{i=1}^{n} y_i\sum_{i=1}^{n} x_i + n\sum_{i=1}^{n} y_i x_i + \beta_1\left(\left(\sum_{i=1}^{n} x_i\right)^2 - n\sum_{i=1}^{n} x_i^2\right) &= 0 \quad (10.8)
\end{aligned}$$

Utilizando las ecuaciones (10.7) y (10.8) llegamos a las siguientes soluciones para los estimadores de los parámetros:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (10.9)$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} \quad (10.10)$$

Finalmente:

$$-\frac{n}{2\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 = 0$$

De donde obtenemos que:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{n} \quad (10.11)$$

Estos estimadores de las ecuaciones (10.10), (10.9) y (10.11) requieren de verificar sus propiedades de sesgo, consistencia y eficiencia. No obstante, lo dejaremos para más adelante en el caso de regresión múltiple.

10.4. Estimación de parámetros por mínimos cuadrados

Hasta el momento hemos discutido el problema de regresión, lineal o no lineal, partiendo de que conocemos la función de densidad de probabilidad de las variables aleatorias involucradas. En la realidad no siempre es factible conocerlas, y por lo tanto tampoco es factible conocer la función de densidad conjunta de dichas variables.

Pero nuestro problema sigue siendo el mismo, queremos determinar los parámetros β_0 y β_1 que pueden ayudarnos a ajustar una curva de regresión. En esta sección discutiremos el método de mínimso cuadrados ordinarios, el cual fue implementado por Adrian Legendre.

Una motivación del método es geométrica. La cual considera la relación entre dos o más variables aleatorias, digamos que observamos puntos en el plano cartesiano: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ y supongamos una constante m que determina el grado del polinomio:

$$\hat{y}_i = p(x_i) = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_i^j$$

Para cualquier grado del polinomio podemos definir a los residuales para cualquier punto $i = 1, 2, \ldots, n$ en el plano como:

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \sum_{j=1}^{m} \hat{\beta}_j x_i^j \end{aligned}$$

Sin pérdida de generalidad, puesto que el siguiente plateamiento es válido para cualquier grado de polinomio, diremos que:

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

En la Figura 10.6 ilustramos el concepto de error.

El método de mínimos cuadrados ordinarios consiste en resolver un problema de minimización de la una función que depende de los residuales que hemos definido. El la función introducimos los residuales al cuadrado con el objeto de aproximar el problema como uno de minimización de distancias

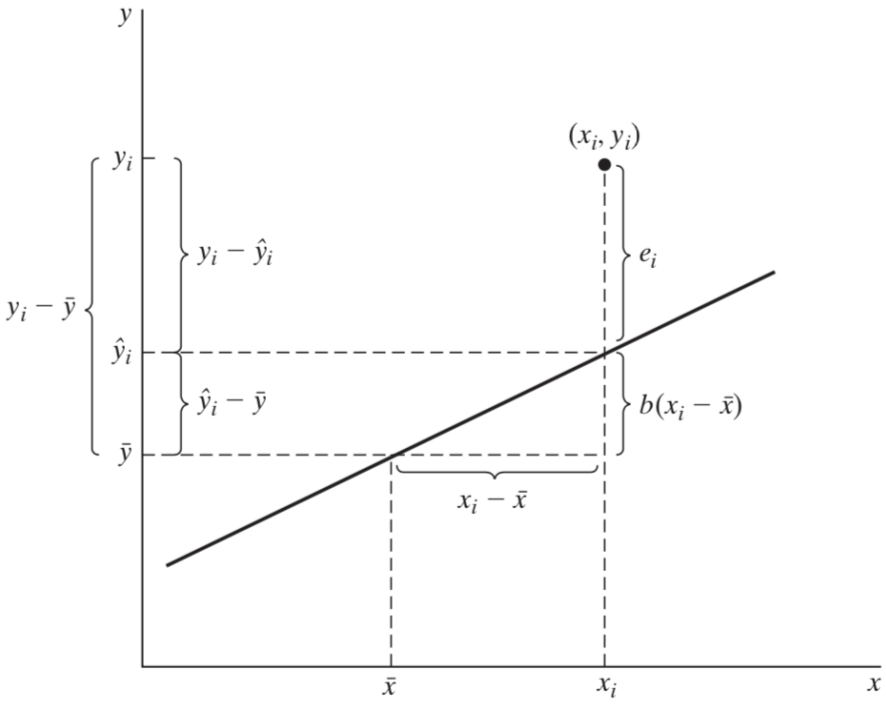

Figura 10.6: Ilustración del error de estimación, retomado de Greene (2012, pp. 40) Greene 2012

entre la coordenada y_i y la coordenada x_i de forma conjunta para todos los elementos en la muestra, $i = 1, 2, \ldots, n$.

Así, partimos del siguiente planteamiento en el cual desconocemos los parámetros β_0 y β_1:

$$\begin{aligned} S(\beta_0, \beta_1) &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

Para mimimizar la función anterior requerimos de las condciones de primer orden dadas como:

$$\begin{aligned} \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \end{aligned}$$

A estas expresiones se les suele conocer como ecuaciones normales. Si igualamos a cero y reolvemos para encontrar los valores que minimizan las distancias tendremos que:

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

De donde podemos encontrar que:

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

Continuando:

$$\begin{aligned}\hat{\beta}_0 &= \frac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n} \\ &= \frac{\sum_{i=1}^n y_i}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^n x_i}{n} \\ &= \overline{Y} - \hat{\beta}_1 \overline{X}\end{aligned}$$

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

Por lo tanto, las soluciones serían:

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

Estas soluciones son las mismas que obtuvimos por el procedimiento de Máxima Verosimilitud.

10.5. Modelos no lineales

¿Qué pasa cuando la relación observada entre las variables no es lineal? Casi siempre pasa eso. A continuación discutiremos tres casos de ecuaciones que nos ayudaran a describir una relación no lineal mediante una transformación que hace la relaciones lineales. Cada uno responde a condiciones o características que la experiencia del investigador tiene que solventar.

Cada una de las ecuaciones se distingue por la tasa de cambio de la variable y en relación a los cambios de la variable x. Partamos de que la forma más simple de una relación lineal está dada por la ecuación (10.12). Note que hemos eliminado los subíndices para indicar que la ecuacion se

cumple para cada uno de los elementos en la muestra, $i = 1, 2, \ldots, n$. Sin pérdida de generalidad digamos que en lugar de β_0 y β_1 empleamos a y b como constantes, y asumimos $\mu_{Y|x} = y$, como una forma de ilustrar cada uno de los fenómenos que discutimos a continuación.

$$y = a + bx \tag{10.12}$$

Esta ecuación (10.12) tiene la característica de que tiene una tasa de variación constante para y en relación a x, de esta forma:

$$\frac{dy}{dx} = b \Longleftrightarrow dy = bdx \Longleftrightarrow \int dy = \int bdx \Longleftrightarrow y = a + bx$$

Los siguientes modelos no lineales tienen una motivación similar, dependen de la forma en que se asuma la variación de y en función de la variación de x.

10.5.1. Regresión exponencial

Supongamos que y depende de x y que los cambios en y darivados de x son proporcionales a y, es decir:

$$\frac{dy}{dx} = by$$

Donde b es una constante. De esta forma podemos encontar que:

$$\frac{dy}{dx} = by \Longleftrightarrow \frac{dy}{y} = bdx \Longleftrightarrow \int \frac{dy}{y} = \int bdx \Longleftrightarrow ln(y) = a + bx$$

De esta forma, tenemos que:

$$y = e^{bx}e^{a} \Longleftrightarrow y = ke^{bx}$$

Por lo tanto, cuando estimamos una relación: $ln(y) = a + bx$ estamos asumiendo que existe la sigueinte relación: $\frac{dy}{dx} = by$. A este tipo de ecuaciones se les conoce como log-lineales. En la Figura 10.7 se ilustra la relación log-lineal - exponencial.

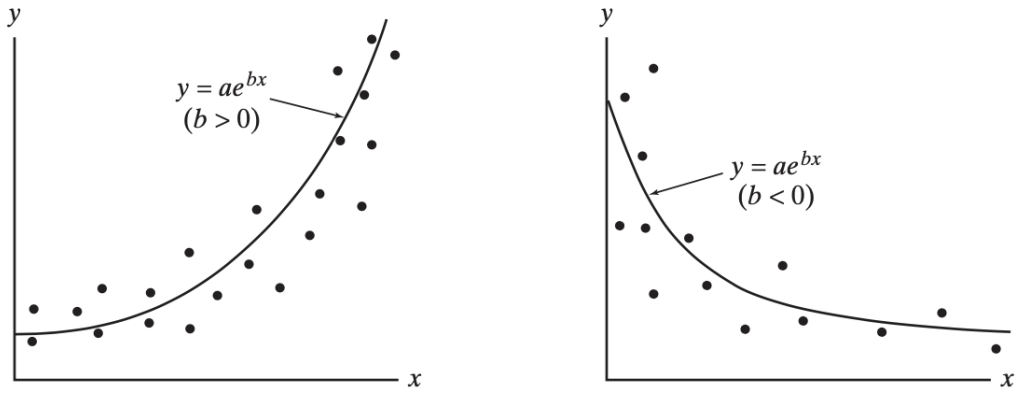

Figura 10.7: Funciones exponenciales, Retomado de Larsen y Marx (2012, p. 532) Larsen y Marx 2012

10.5.2. Regresión logarítmica

En este caso suponemos que los cambios en y por causa de cambios en x son proporcionales a la razón que guardan y y x, es decir:

$$\frac{dy}{dx} = b\frac{y}{x}$$

Por lo que podemos establecer:

$$\frac{dy}{dx} = b\frac{y}{x} \iff \frac{dy}{y} = b\frac{dx}{x} \iff \int \frac{dy}{y} = \int b\frac{dx}{x} \iff ln(y) = a + ln(x)$$

Así, podemos establecer que la forma funcional de la relación de y y x es como sigue y como se muestra en la Figura 10.8.

$$y = kx^b$$

10.5.3. Funciones logísticas

Este tipo de ecuaciones tienen la característica de que permiten modelar crecimientos poblacionales, aceptación de políticas públicas, aceptación de tecnologías, epidemias, etc. En este caso suponemos que enfrentamos que los cambios en y dados los cambios en x son proporcionales a y y a las distancia

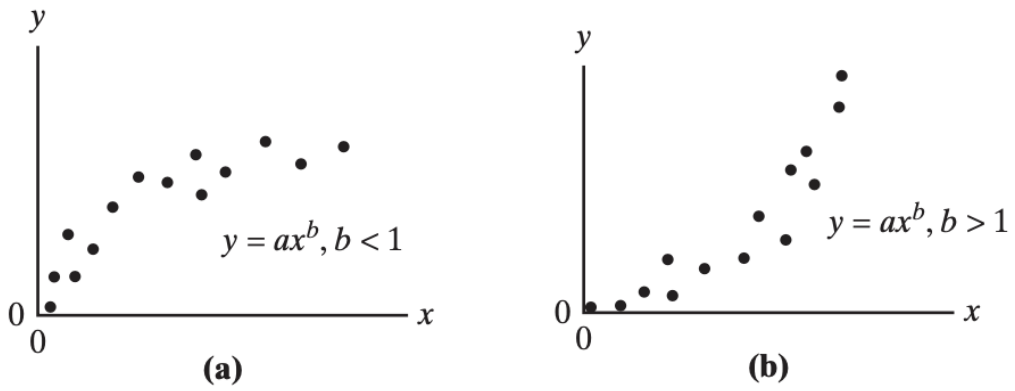

Figura 10.8: Funciones logarítmicas, Retomado de Larsen y Marx (2012, p. 536) Larsen y Marx 2012

que y tiene respecto de un factor de saturación L (límite superior o punto de saturación poblacional, cobertura universal, infecciones en el total de la población, etc.).

En este caso suponemos que la variación estará dada por:

$$\frac{dy}{dx} = ky(L - y)$$

Donde k y L son constantes. Una vez solucionada la ecuación diferencial anterior, encontraremos que, si $L = 1$, entonces l a solución será como sigue, la Figura 10.9 ilustra ecuaciones como la siguiente:

$$y = \frac{1}{1 + e^{a+bx}} \tag{10.13}$$

Ahora mostremos como llegar a la solución de la ecuación logística:

$$\frac{dy}{dx} = ky(L - y) \Longleftrightarrow \frac{dy}{y(L - y)} = kdx \Longleftrightarrow \left(\frac{\frac{1}{L}}{y} + \frac{\frac{1}{L}}{(L - y)}\right) dy = kdx$$

$$\Longleftrightarrow \frac{1}{L}\left(\frac{dy}{y} + \frac{dy}{(L - y)}\right) = kdx \Longleftrightarrow \int \frac{1}{L}\left(\frac{dy}{y} + \frac{dy}{(L - y)}\right) = \int kdx$$

$$\Longleftrightarrow \left(\int \frac{dy}{y} + \int \frac{dy}{(L - y)}\right) = \int Lkdx \Longleftrightarrow (ln(y) - ln(L - y)) = Lkx + C$$

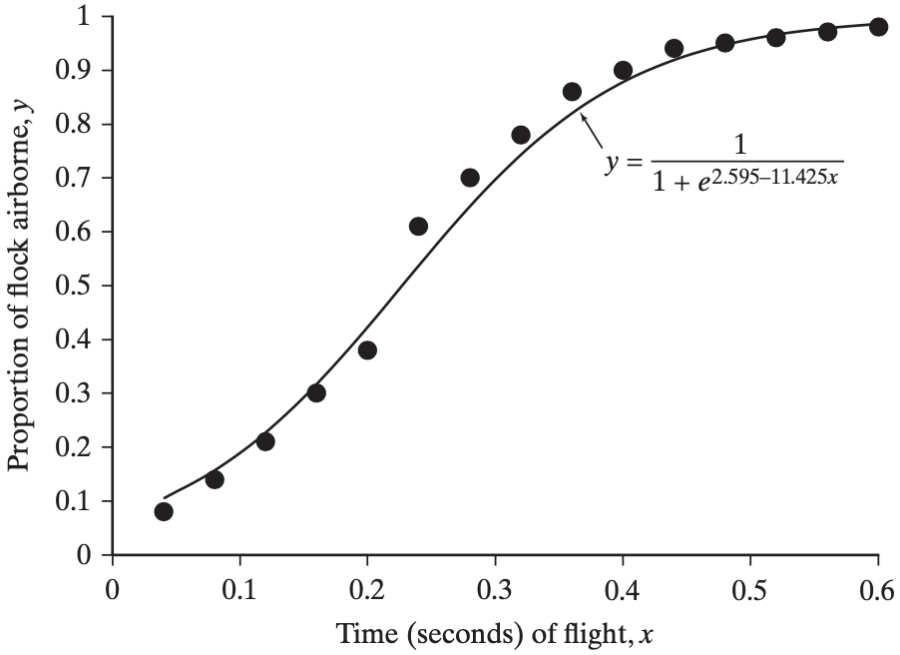

Figura 10.9: Funciones logísticas, Retomado de Larsen y Marx (2012, p. 540) Larsen y Marx 2012

$$\iff (ln(L-y) - ln(y)) = -LC - Lkx \iff ln\left(\frac{L-y}{y}\right) = -LC - Lkx$$

$$\iff \left(\frac{L-y}{y}\right) = e^{-LC-Lkx} \iff y = \frac{L}{1+e^{-LC-Lkx}}$$

Por lo tanto, la solución general de la ecucación será:

$$y = \frac{L}{1+e^{a+bx}}$$

Donde asumimos que la forma lineal de la ecuación logística es:

$$ln\left(\frac{L-y}{y}\right) = a + bx$$

10.6. Modelo de regresión lineal múltiple

En esta sección ampliaremos la discusión de regresión bivariada para introducir más varaibles de forma que ahora supondremos una media condicional como:

$$\mu_{Y|x_1,x_2,\ldots,x_K} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K = \hat{y}$$

En este caso, como se notará, tenemos $K+1$ parámetros a estimar mediante el método de mínimos cuadrados ordinarios. Por ello, nuestro conjunto de datos será descrito por:

$$\{(x_{i1}, x_{i2}, \ldots, x_{iK}; y_i) : i = 1, 2, \ldots, n\}$$

En este caso podemos ilustrar el término de error con la Figura 10.10. Donde asumiremos que se cumplen los supuestos del modelo lineal, pero para múltiples variables independientes. De esta forma podemos establecer la función:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

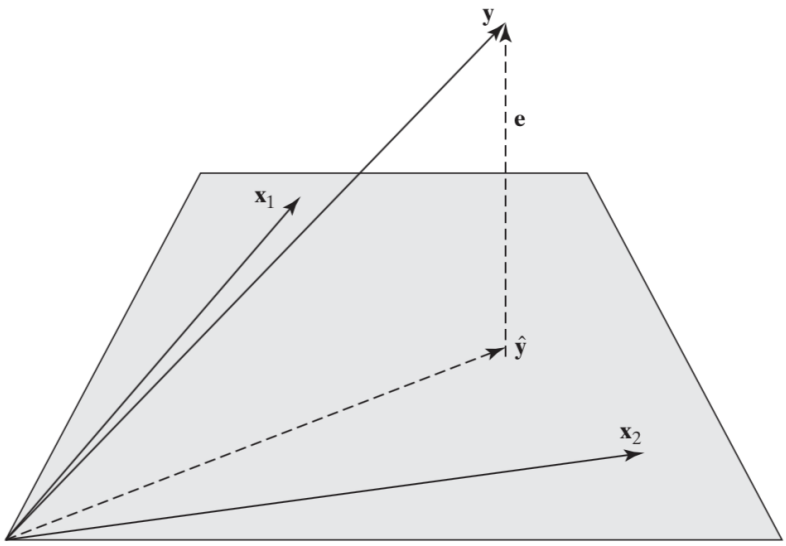

Figura 10.10: Ilustración de la poroyección de y en x_1 y x_2 y el significado del error e, retomado de Greene (2012, pp. 20) Greene 2012

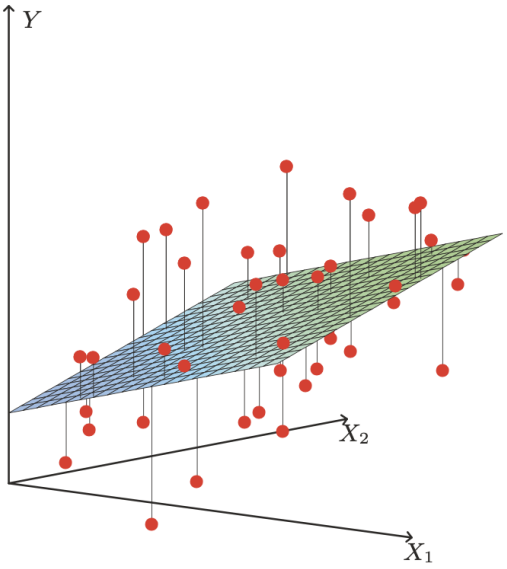

Figura 10.11: Ilustración del hiperplano generado dee regresar X_1 y X_2 en Y, retomado de Hastie, Tibshirani y Friedman (2009, pp. 45) Hastie, Tibshirani y Friedman 2017

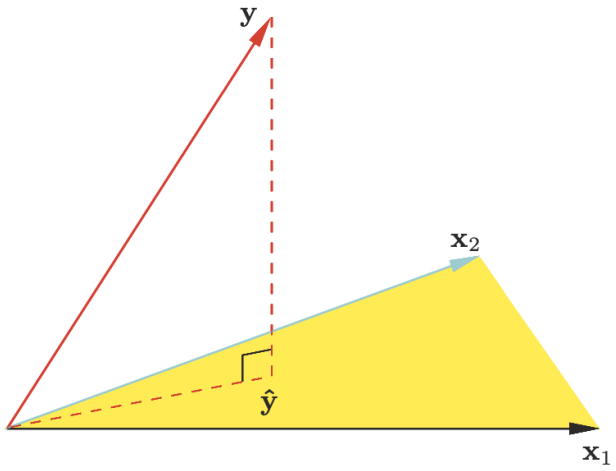

Figura 10.12: Ilustración de la poroyección de y en x_1 y x_2 y el significado del error e, retomado de Hastie, Tibshirani y Friedman (2009, pp. 46) Hastie, Tibshirani y Friedman 2017

Donde:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

Entonces nuestro problema consiste en:

$$\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_K x_{iK})^2$$

Las $K+1$ condiciones de primer orden estarán dadas por:

$$\begin{aligned}
\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_K x_{iK}) \\
\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_1} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_K x_{iK})x_{i1} \\
&\vdots \\
\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_K} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_K x_{iK})x_{iK}
\end{aligned}$$

De donde tenemos $K+1$ ecuaciones normales. Resolviendo el siguiente sistema podríamos obtener los estimadores de los parámetros:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_K x_{iK}) &= 0 \\
\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_K x_{iK})x_{i1} &= 0 \\
&\vdots \\
\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_K x_{iK})x_{iK} &= 0
\end{aligned}$$

¿Cómo resolvemos este sistema? Utilicemos una alternativa. Para ello introduciremos notación matricial. Sean:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1K} \\ 1 & x_{21} & x_{22} & \ldots & x_{2K} \\ 1 & x_{31} & x_{32} & \ldots & x_{3K} \\ \vdots & \vdots & \vdots & \ddots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nK} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

Bajo esta notación podemos ver que:

$$\mu_{\mathbf{Y}|\mathbf{x}} = \mathbb{E}[\mathbf{Y}|\mathbf{x}] = \mathbf{X}\boldsymbol{\beta}$$

Definamos el vector de errores de la estimación como:

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \\ \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} &= \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & \dots & x_{2K} \\ 1 & x_{31} & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \end{aligned}$$

De esta forma, definiremos la función:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} e_i^2 = \begin{bmatrix} e_1 & e_2 & e_3 & \dots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{e}'\mathbf{e} \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Así, nuestro problema de optimización consiste en:

$$\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

En los siguientes pasos utilzaré una propiedad de la derivación de matrices. Sean A, B y C matrices matrices con dimensiones que permiten los

productos: $A \cdot B$ y $B' \cdot C \cdot B$, y además supongamos que la matriz C es simétrica, $C = C'$. Entonces lo siguiente ees cierto:

$$\begin{aligned} \frac{\partial(A \cdot B)}{\partial B} &= A' \\ \frac{\partial(B' \cdot C \cdot B)}{\partial B} &= 2 \cdot C' \cdot B = 2 \cdot C \cdot B \end{aligned}$$

Dicho lo anterior, tenemos que el gradiente será:

$$\begin{aligned} \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) \\ &= -2\mathbf{X}'\mathbf{Y} + 2(\mathbf{X}'\mathbf{X})'\boldsymbol{\beta} \\ &= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Donde hemos utilizado que la matriz $\mathbf{X}'\mathbf{X}$ es simétrica (más adelante discutimos esta condición). Finalmente, las condiciones de primer orden serán:

$$-\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

Si la inversa de la matriz $\mathbf{X}'\mathbf{X}$, entonces la solución de mínimos cuadrados será:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{10.14}$$

Para verificar si es un mínimo:

$$\frac{\partial}{\partial \boldsymbol{\beta}}(-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = 2\mathbf{X}'\mathbf{X}$$

Donde esta ecuación es definida positiva. Ahora analicemos cada uno de los componentes de la ecuación (10.14)

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ x_{1K} & x_{2K} & x_{3K} & \dots & x_{nK} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & \dots & x_{2K} \\ 1 & x_{31} & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix} \\ &= \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \dots & \sum_{i=1}^n x_{iK} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \dots & \sum_{i=1}^n x_{i1}x_{iK} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \dots & \sum_{i=1}^n x_{i2}x_{iK} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ \sum_{i=1}^n x_{iK} & \sum_{i=1}^n x_{iK}x_{i1} & \sum_{i=1}^n x_{iK}x_{i2} & \dots & \sum_{i=1}^n x_{iK}^2 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
\mathbf{X'Y} &= \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ x_{1K} & x_{2K} & x_{3K} & \dots & x_{nK} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \\ \sum_{i=1}^n x_{i2} y_i \\ \vdots \\ \sum_{i=1}^n x_{iK} y_i \end{bmatrix}
\end{aligned}$$

Ahora verifiquemos las propiedades del estimador que teníamos pendientes de la regresión bivariada. Partiendo de la ecuación (10.14), tenemos :

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{x}] &= \mathbb{E}[(\mathbf{X'X})^{-1}\mathbf{X'Y}|\mathbf{x}] \\
&= (\mathbf{X'X})^{-1}\mathbf{X'}\mathbb{E}[\mathbf{Y}|\mathbf{x}] \\
&= (\mathbf{X'X})^{-1}\mathbf{X'X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

Por lo tanto el estimador de mínimos cuadrados ordinarios es insesgado. Analicemos ahora la varianza:

$$\begin{aligned}
Var[\hat{\boldsymbol{\beta}}|\mathbf{x}] &= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{x}])(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{x}])'|\mathbf{x}] \\
&= \mathbb{E}[((\mathbf{X'X})^{-1}\mathbf{X'Y} - \boldsymbol{\beta})((\mathbf{X'X})^{-1}\mathbf{X'Y} - \boldsymbol{\beta})'|\mathbf{x}] \\
&= \mathbb{E}[(\mathbf{X'X})^{-1}\mathbf{X'YY'X}(\mathbf{X'X})^{-1} - (\mathbf{X'X})^{-1}\mathbf{X'Y}\boldsymbol{\beta}' \\
&\quad -\boldsymbol{\beta}\mathbf{Y'X}(\mathbf{X'X})^{-1} + \boldsymbol{\beta}\boldsymbol{\beta}'|\mathbf{x}] \\
&= (\mathbf{X'X})^{-1}\mathbf{X'} \cdot \mathbb{E}[\mathbf{YY'}|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X'X})^{-1} \\
&\quad -(\mathbf{X'X})^{-1}\mathbf{X'} \cdot \mathbb{E}[\mathbf{Y}|\mathbf{x}] \cdot \boldsymbol{\beta}' \\
&\quad -\boldsymbol{\beta} \cdot \mathbb{E}[\mathbf{Y'}|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X'X})^{-1} + \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= (\mathbf{X'X})^{-1}\mathbf{X'} \cdot \mathbb{E}[\mathbf{YY'}|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X'X})^{-1} - (\mathbf{X'X})^{-1}\mathbf{X'X}\boldsymbol{\beta}\boldsymbol{\beta}' \\
&\quad -\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X'X}(\mathbf{X'X})^{-1} + \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= (\mathbf{X'X})^{-1}\mathbf{X'} \cdot \mathbb{E}[\mathbf{YY'}|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X'X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}' - \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= (\mathbf{X'X})^{-1}\mathbf{X'} \cdot \mathbb{E}[\mathbf{YY'}|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X'X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= (\mathbf{X'X})^{-1}\mathbf{X'}(\sigma^2\mathbb{I}_n - \mathbf{X'}\boldsymbol{\beta}'\boldsymbol{\beta}\mathbf{X})\mathbf{X}(\mathbf{X'X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= \sigma^2(\mathbf{X'X})^{-1}
\end{aligned}$$

De esta forma planteamos que:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Ahora, partiendo de que $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, requerimos de determinar:

$$\begin{aligned}
\mathbb{E}[\mathbf{e}'\mathbf{e}|\mathbf{x}] &= \mathbb{E}[(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})|\mathbf{x}] \\
&= \mathbb{E}[(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})'(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})|\mathbf{x}] \\
&= \mathbb{E}[(\mathbf{Y}' - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})|\mathbf{x}] \\
&= \mathbb{E}[\mathbf{Y}'(\mathbb{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbb{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}|\mathbf{x}] \\
&= \mathbb{E}[tr(\mathbf{e}'\mathbf{e})|\mathbf{x}] \\
&= \mathbb{E}[tr((\mathbb{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}\mathbf{Y}')|\mathbf{x}] \\
&= tr((\mathbb{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbb{E}[\mathbf{Y}\mathbf{Y}'|\mathbf{x}]) \\
&= tr((\mathbb{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\sigma^2\mathbf{I}_n - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}')) \\
&= tr(\sigma^2\mathbb{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= tr(\sigma^2\mathbb{I}_n) - tr(\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \sigma^2 tr(\mathbb{I}_n) - \sigma^2 tr(\mathbb{I}_{K+1}) \\
&= \sigma^2(n - (K+1))
\end{aligned}$$

De esta forma utilizaremos un estimador de la varianza σ^2 que es insesgado:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n - (K+1)}$$

De esta forma podemos plantear la distribución muesral de los parámetros $\boldsymbol{\beta}$ como:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &\sim N(\boldsymbol{\beta}, \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}) \\
\hat{\sigma}^2 &= \frac{\mathbf{e}'\mathbf{e}}{n - (K+1)}
\end{aligned}$$

Dado que desconocemos la varianza de la distribución de $\mathbf{Y}$ y $\mathbf{X}$ utilizaremos una transformación t-Student de la siguiente forma:

$$T = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t_{[n-(K+1)]}$$

Donde $(\mathbf{X}'\mathbf{X})^{-1}_{kk}$ es la posición (k,k) de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. De esta forma podemos establecer una pruba de hipótesis para cada una de las β_k's como:

$$\begin{aligned} H_0 &: \hat{\beta}_k = \beta_{k,0} \\ H_a &: \begin{cases} \hat{\beta}_k > \beta_{k,0} \\ \hat{\beta}_k < \beta_{k,0} \\ \hat{\beta}_k \neq \beta_{k,0} \end{cases} \end{aligned}$$

Para la cual podemos utilizar la misma regla de rechazo que ya hemos estudiado en la sección de prueba de hipótesis.

De forma complementaria podemos establecer hipótesis sobre combinaciones lineales de los estimadores de $\boldsymbol{\beta}$. Las cuales podemos platear como:

$$\begin{aligned} H_0 &: \mathbf{R}\hat{\boldsymbol{\beta}} = q \\ H_a &: \begin{cases} \mathbf{R}\hat{\boldsymbol{\beta}} > q \\ \mathbf{R}\hat{\boldsymbol{\beta}} < q \\ \mathbf{R}\hat{\boldsymbol{\beta}} \neq q \end{cases} \end{aligned}$$

Donde $\mathbf{R}$ es un vector de dimensión $1 \times (K+1)$ de restricciones que toma el valor de 0 y 1 de acuerdo con la restricción que queramos evaluar, y q es el valor que asumimos como cierto y queremos evaluar. De forma similar planteamos la transformación a evaluar como:

$$T = \frac{\mathbf{R}\hat{\boldsymbol{\beta}} - q}{\sqrt{\hat{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}} \sim t_{[n-(K+1)]}$$

Definamos a F de Fisher como la razón de dos pruebas t de Student, la primera para los valores de una combinación lineal de $\hat{\boldsymbol{\beta}}$, y la segunda para el estimador de la varianza $\hat{\sigma}^2$:

$$\begin{aligned} \mathbf{F} &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/J}{\hat{\sigma}^2} \\ &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \\ &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}Var(\hat{\boldsymbol{\beta}})\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \end{aligned}$$

Por lo tanto la estadística F de prueba será:

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \sim F_{J,n-K} \tag{10.15}$$

Así, la prueba de hipótesis que es la más común en el análisis de regresión y que se le conoce como prueba global. Dicha prueba asume una forma de la matriz $\mathbf{R}$ definida así:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Y un vector $\mathbf{r}$:

$$\mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

De esta forma la hipotésis nula de una prueba global se puede escribir como:

$$\begin{aligned} H_0 : &\quad \beta_1 = \beta_2 = \ldots = \beta_K = 0 \\ H_a : &\quad NoH_0 \end{aligned}$$

Esta prueba se le conoce como prueba global, ya que prueba si en conjunto todas las variables independientes tienen un efecto nulo en $\mathbf{Y}$.

Ejemplo. Suponga que tiene datos de precios y cantidades mostrados en el Cuadro 10.1. Con dichos datos estima la ecuación:

$$Q_i = \beta_0 + \beta_1 P_i$$

De esta forma la matriz $\mathbf{X}$ y vector $\mathbf{Y}$ serán como sigue:

$$\mathbf{X} = \begin{bmatrix} 1 & 3 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 18 \\ 16 \\ \vdots \\ 8 \end{bmatrix}$$

Cuadro 10.1: Tabla de Precios y Cantidades

P	Q
3	18
3	16
7	17
6	12
10	15
15	15
16	4
13	13
9	11
15	8
9	8

De esta forma los estimadores de $\boldsymbol{\beta}$ seran:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 18.3977 \\ -0.6356 \end{bmatrix}$$

Una vez estimados los parámetros podemos gráficar los puntos en la muestra la recta (ver Figura 10.13).

Ejemplo. Suponga que ha estimado una ecuación de regresión del precio de las casas. Su estimación se rigue por la siguiente ecuación:

$$ln(P) = \beta_0 + \beta_1 ln(AreaM^2) + \beta_2 \#Recamaras + \beta_3 \#Banios + \beta_4 Antiguedad$$

La estimación de dicha ecuación se reporta el en Cuadro 10.2, en el que se le pide que complete la columa de T-Estadística y determine si en cada caso acepta o rechaza la hipótesis:

$$\begin{aligned} H_0 &: \hat{\beta}_k = 0 \\ H_1 &: \hat{\beta}_k \neq 0 \end{aligned}$$

Asimismo, utilicemos la siguiente transformación:

$$T = \frac{\hat{\beta}_k - 0}{\sqrt{\hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}_{kk}}} = \frac{\hat{\beta}_k - 0}{\text{Errror Estándar}} \sim t_{[n-(K+1)]}$$

Donde $k = 0, 1, 2, 3, 4$ y $n = 100$.

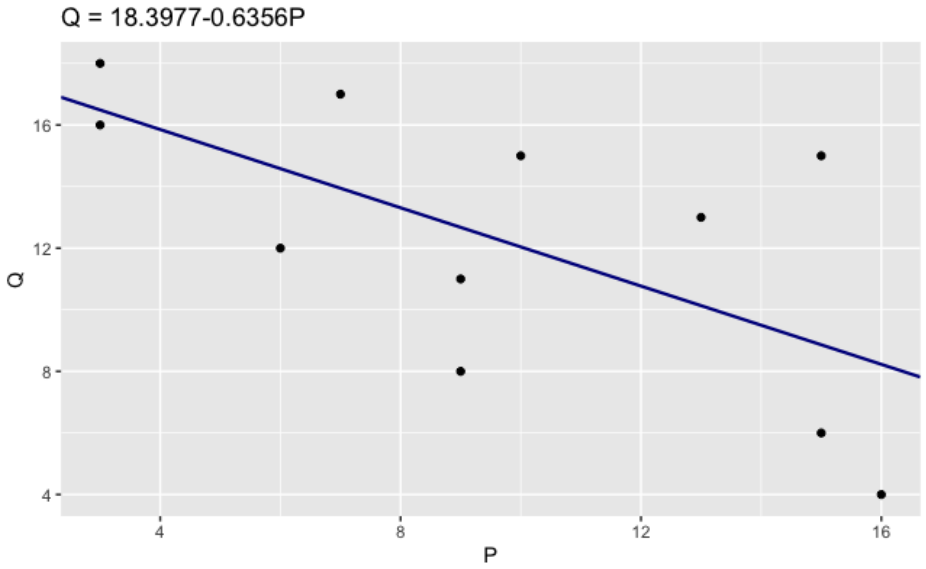

Figura 10.13: Gráfica de Precios y Cantidades, y su recta de regresión

Cuadro 10.2: Regresión resultados

$\hat{\beta}_k$	Coeficiente	Error Estándar	T-Estadística
β_0	7.094	0.232	
β_1	0.400	0.028	
β_2	0.078	0.015	
β_3	0.216	0.023	
β_4	0.212	0.024	

11

Estadística no paramétrica

11.1. Introducción y motivación

Es las secciones prrevias hemos hecho supuestos sobre la naturaleza de la función de densidad de probabilidad subyacente o que describe a la pobación de donde hemos extraido muestras de datos. Algunos de esos supuestos parecen obvios pero ¿cómo cambiarían nuestras conclusionnes si esos supuestos no se cumplierran?.

Ciertamente, las mismas estadísticas se podrían seguir calculando de la misma manera, así como establecer hipóttesis con la misma rergión de rechazo. No obstante, la relidad es que el modelo estadístico ya no se sostiene.

El término, no paramétrico se puede sustituir por 'distribución libre'. Es ese sentido nuestro objetivo será introducir algunas metodologías más importantes de la estadística no parámetrica para considerar situaciones en las que no hay restricciones de distribución.

11.2. Estimaciones de funciones de densidad via Kernel (Núcleo)

Este estimador fue propuesto por Rosenblatt (1956):

Definición 11.1 *Sea la muestra de n observaciones* $X_1, X_2, \ldots, X_n$*, donde cada* $X_i \in \mathbb{R}$*. Definimos la estimación de una fucnión de densidad de*

probabilidad f via Kernel como:

$$\hat{f}_n(x) = \frac{1}{n \cdot h_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Donde $K(x)$ es la función Kernel y h_n es el ancho de banda.

El ancho de banda (h_n), también llamado parámetro de suavizammienta es una secuencia de constantes positivas que determina el refinnamiento del soporte de la función estimada y el suavizamiento de la función. La elección del ancho de banda adecuado es un debate extenso:

- Si h es muy grande, la estimación resultante será muy suave
- Si es muy pequeña, prácticamente se interpolan los datos.

Cabe destacar que, al ser un parámetro fijo a lo largo de toda la muestra, la estimación kernel suele presentar distorsiones en las colas de la estimaciónn, tal y como veremos en los ejemplos de clase.

La función kerrnel $K(x)$ es la que define la forma y la importancia de los pesos que se asocian a cada observación para el cálculo de la estimación. Algunos ejemplos de funciones Kernel más utilizadas son las siguientes.

Epanechnikov: asigna los pesos de acuerdo con la siguiente función.

$$K(t) = \frac{3}{4}(1 - t^2) \text{ con } |t| < 1$$

La Figura 11.1 muestra la gráfica de la función kernel de Epanechnikov.

Gauss: Asigna los pesos siguiendo la distribución normal estándar.

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \text{ con } |t| < \infty$$

La Figura 11.2 muestra la gráfica de la función kernel de Gauss.

Existen otros tipos de kernel, pero los anteriores son los más comunes.

11.3. La prueba de signo

Esta sección es una continuación de la prueba de hipótesis diseñada en especial para datos no paramétricos. Para realizar estas pruebas no requerimos de hacer ninguna suposición acerca de la distribución de la población.

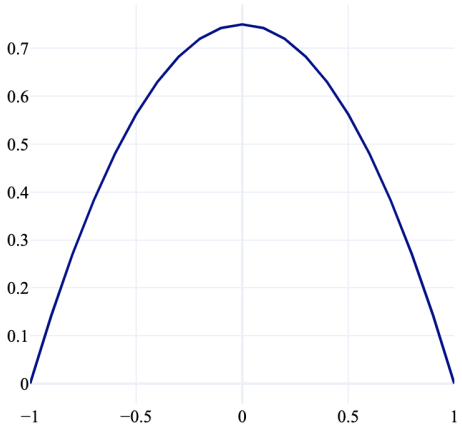

Figura 11.1: Kernel Epanechnikov

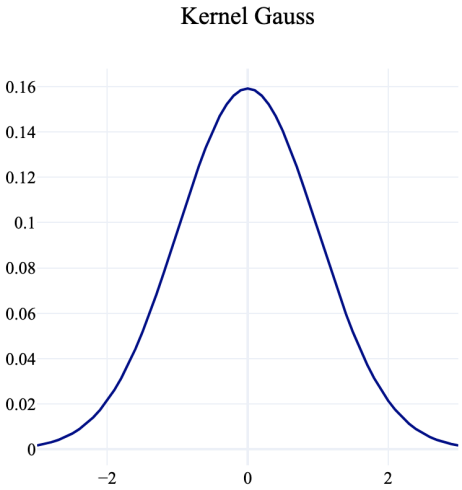

Figura 11.2: Kernel Gauss

En esta sección consideramos cinco pruebas sin distribución y coeficiente de correlación de los rangos de Spearman. Las pruebas son: de signo, de la mediana, de los rangos con signo de Wilcoxon, de la suma de los rangos de Wilcoxon y el análisis de la varianza por rangos de Kruskal-Wallis.

Sencillamente, la mediana se puede definir como el valor $\tilde{\mu}$ de una función de densidad, $f_Y(y)$, continua para el cual obsevamos que $P(Y \leq \tilde{\mu}) = P(Y \geq \tilde{\mu}) = 1/2$. Así, supongamos que tomamos una muestra aleatoria de n elementos tomados de $f_Y(y)$. Si la hipótesis nula $H_0 : \tilde{\mu} = \tilde{\mu}_0$ es cierta, entonces el número de observaciones, X, que excede a $\tilde{\mu}$ es una variable aleatoria binomial con $p = P(Y_i \geq \tilde{\mu}_0) = 1/2$. Así:

$$\begin{aligned} \mathbb{E}[X] &= \frac{n}{2} \\ Var(X) &= \frac{n}{4} \end{aligned}$$

De esta forma, podemos construir la siguiente transformación (siempre n sea suficientemente grande):

$$z = \frac{X - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \sim N(0,1)$$

Podemos darle formalidad a lo anterior.

Teorema 11.1 *Sean $y_1, y_2, \cdots, y_n$ una muestra aleatoria de tamaño n de una función de distribución continua con mediana $\tilde{\mu}$, donde $n \geq 10$. Sea k el número de y_i's más grandes que una valor $\tilde{\mu}_0$, y sea:*

$$z = \frac{k - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$$

1. *Para probar $H_0 : \tilde{\mu} = \tilde{\mu}_0$ versus $H_1 : \tilde{\mu} > \tilde{\mu}_0$ a un nivel de significancia de α, rechazamos H_0 si $z \geq z_\alpha$*

2. *Para probar $H_0 : \tilde{\mu} = \tilde{\mu}_0$ versus $H_1 : \tilde{\mu} < \tilde{\mu}_0$ a un nivel de significancia de α, rechazamos H_0 si $z \leq -z_\alpha$*

3. *Para probar $H_0 : \tilde{\mu} = \tilde{\mu}_0$ versus $H_1 : \tilde{\mu} \neq \tilde{\mu}_0$ a un nivel de significancia de α, rechazamos H_0 si $|z| \geq z_\alpha$*

Obs: La prueba de signos está diseñada para hacer inferencias sobre las medianas. Si el muestreo es realizado sobre una pdf simetrica, la mediana es la misma que la media, así probar si $\tilde{\mu} \neq \tilde{\mu}_0$ es equivalente a probar $\mu \neq \tilde{\mu}_0$. Otra observación es que cuando $n < 10$ z no se aproxima a una normal, por lo que de forma alternativa se emplea una aproximación binomial.

Ejemplo. El Cuadro 11.1 reporta una muestra del precio de la gasolina regular de 20 estaciones de servicio cobrado durante el 8 de mayo de 2021. Suponga que sabe que la mediana pobacional es de \$20.30. Considerando la muestra y un nivel de confianza del 5 %, verifique:

$$\begin{aligned} H_0 &: \tilde{\mu} = 20.30 \\ H_1 &: \tilde{\mu} \neq 20.30 \end{aligned}$$

Cuadro 11.1: Precios de Gasolina Regular (Pesos por Litro)

Razón Social	Precio
ESTACION DE SERVICIO CALAFIA, S.A. DE C.V.	19.99
LAS MEJORES ESTACIONES, S.A DE C.V	20.35
DIAZ GAS, S.A. DE C.V.	15.43
COMBU-EXPRESS, S.A. DE C.V.	20.19
PETROMAX, S.A. DE C.V.	17.09
ESTACION RAEL, S. DE R.L. DE C.V.	20.19
MULTISERVICIOS LA PILARICA SA DE CV	21.7
PETROMAX, S.A. DE C.V.	21.39
ESTACION RAEL, S. DE R.L. DE C.V.	20.59
ESTACION PIRU, S.A. DE C.V.	19.69
PETROMAX, S.A. DE C.V.	21.09
AUTOSERVICIO BERLANGA S.A. DE C.V.	21.35
PETROMAX, S.A. DE C.V.	16.09
GRUPO OCTANO, S.A. DE C.V.	21.99
CIRCULO DOS, S.A. DE C.V.	20.49
SERVICIO SEVILLA, S.A. de C.V.	20.29
PETROMAX, S.A. DE C.V.	20.39
GASOLINERA HUITEPEC, S.A. DE C.V.	20.6
PETROMAX, S.A. DE C.V.	17.09
ENERGIA Y SERVICIOS COORDINADOS, S.A. DE C.V.	20.94

Ahora establecemos la transformación:

$$z = \frac{k - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{11 - \frac{20}{2}}{\sqrt{\frac{20}{4}}} = \frac{11 - 10}{\sqrt{5}} = 0.4472$$

Así, podemos aceptar la hipótesis nula, ya que $|z| = 0.4472 < 1.96 = z_{0.05}$.

La prueba de signo toma su nombre de situaciones en las que son muestras pareadas del tipo $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, y para las cuales queremos evaluar una prueba respecto de, por ejemplo, la efectividad de un medicamento, las preferencias sobre un nuevo producto, etc. En estas situaciones dificilmente podríamos aplicar una prueba como la t, puesto que los datos no necesariamente provienen de una normal o simplemente no tenemos elementos para sostenerlo.

Sea:

$$p = P(X_i > Y_i),\ i = 1, 2, \ldots, b$$

Este es el un caso particular de la prueba de signo; asumiento $p = 1/2$. Sea U una variable aleatoria binomisla que cintabiliza cuantas observaciones pareadas cumplen con $d_i = x_i - y_i > 0$. Esta prueba requiere que $b > 10$ para que se aproxime a una normal estándar.

Ejemplo. El departamento de investigación de mercado de Cola, Inc., tiene la tarea de probar una nueva bebida. Se consideran dos versiones: un refresco más bien dulce y uno un tanto amargo. La prueba de preferencia que se realizará consiste en una muestra de 64 consumidores. Cada uno de éstos degustará las dos bebidas de cola, la dulce (con la etiqueta A) y la amarga (con la etiqueta B), e indicará su preferencia.

Después de realizar el experimento determinó que 42 personas prefirieron la versión dulce del producto. Realice una prueba de hipótesis para determinar si hay una diferencia entre las preferencias por el refresco dulce o por el amargo. Utilice un nivel de significancia de 0.05.

Formulamos la hipótesis:

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p \neq 0.5 \end{aligned}$$

$$z = \frac{u - \frac{b}{2}}{\sqrt{\frac{b}{4}}} = \frac{42 - \frac{64}{2}}{\sqrt{\frac{64}{4}}} = \frac{42 - 32}{\sqrt{16}} = 2.5$$

Rechazamos la hipótesis nula.

11.4. Prueba de Wilcoxon

Aunque la prueba de signos es un procedimiento no paramétrico genuino, su extrema simplicidad lo hace algo atípico. La prueba de rango con signo de Wilcoxon presentada en esta sección es más representativa de los procedimientos no paramétricos en su conjunto. Al igual que la prueba de signos, se puede adaptar a varias estructuras de datos diferentes. Se puede utilizar, por ejemplo, como una prueba de una muestra para la ubicación, donde se convierte en una alternativa a la prueba t. También se puede aplicar a datos emparejados y, con solo modificaciones menores, puede convertirse en una prueba de dos muestras para la ubicación y una prueba de dos muestras para la dispersión (siempre que las dos poblaciones tengan ubicaciones iguales).

La prueba de Wilcoxon es aplicable en los mismos supuestos que en el caso anterior, para variables continuas relacionadas. Esta prueba es más potente que la de los signos, pues tiene en cuenta el signo, el aumento o disminución de la variable y la magnitud del cambio.

La técnica, consiste en:

1. Calcular las diferencias entre las variables de cada elemento y ordenarlas por valor absoluto, de menor a mayor; esto significa que una diferencia negativa de 5 puntos se ordena posteriormente a una diferencia positiva de 4 puntos

2. Una vez ordenadas las diferencias, las numeramos de 1 a n, siendo n el número de individuos de la muestra; a este número le llamaremos rango. El rango 1 lo asignaremos a la mÌnima diferencia observada en valor absoluto, y así sucesivamente hasta n, cuyo rango corresponderá a la máxima diferencia. Si hay empate, se asigna a cada diferencia empatada la media de los rangos implicados en el empate; por ejemplo, si hay 3 elementos empatados a los que les corresponderían los rangos 4, 5 y 6, se asigna a los tres el rango medio que, en este caso, es la suma de los tres, que es 15, dividida por 3; por lo tanto, asignaríamos a los tres el rango 5 y al siguiente elemento le asignarÌamos el rango 7.

3. Una vez ordenados los datos, sumamos los rangos de las diferencias positivas (W+) también sumamos los rangos de las diferencias negativas (W-) y elegimos el menor de los dos. En la mayoría de las tablas y estadÌsticas, se usa la suma de rangos menor. Los casos en los que la diferencia sea cero se ignorarán.

La prueba se basa en que, si no hay efecto entre las dos variables relacionadas, los rangos estarán repartidos de forma homogénea, y tan probable será encontrar un rango grande positivo como negativo. Por lo tanto, si sumamos los rangos correspondientes a diferencias positivas (W+) y los rangos correspondientes a diferencias negativas (W-), deben ser similares y se encontrará entre ellos pequeñas diferencias debidas al azar. Si las diferencias entre la suma de rangos, son significativamente mayores en un sentido, rechazaremos la hipótesis nula y concluiremos que el efecto de la causa diferenciadora es significativo.

La hipótesis serán:

$$\begin{aligned} H_0 &: W(+) = W(-) \\ H_1 &: W(+) \neq W(-) \end{aligned}$$

El contraste se resuelve, para muestras pequeñas, consultando las tablas de Wilcoxon, en las que se representan las máximas o mínimas sumas de rangos consideradas aceptables y se rechaza la hipótesis nula, en caso de que la suma de rangos observada sea superior o inferior a estos valores.

Para muestras grandes ($n > 10$), podemos hacer una aproximación a la normal, con la media y desviación tÌpica definidas por las siguientes expresiones:

$$\begin{aligned} \mu &= \frac{n(n+1)}{4} \\ \sigma^2 &= \frac{n(n+1)(2n+1)}{24} \end{aligned}$$

A partir de las expresiones anteriores podemos establecer la estadística de prueba:

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

En la expresión anterior, W es la mínima suma de rangos entre la suma de rangos de las diferencias positivas y la suma de rangos de las diferencias negativas.

Ejemplo. Supongamos que tenemos 15 alumnos cuyas puntuaciones en estadística conocemos; después de realizar un curso especial, volvemos a evaluarlos y los resultados son los que se muestran en el Cuadro 11.2. Pruebe la hipótesis de que no hay efecto entre las dos variables relacionadas.

Cuadro 11.2: Calificaciones de Estadística

Antes	Después	Signo
5	6	+
6	6	0
6	7	+
8	9	+
7	6	-
5	4	-
4	6	+
3	3	0
7	8	+
5	8	+
6	4	-
6	7	+
3	2	-
5	7	+
5	8	+

A continuación, resolveremos mediante la prueba de Wilcoxon. En el Cuadro 11.3 las diferencias entre las variables antes y después están calculadas en valor absoluto. En la columna RANGOS, se han ordenado las diferencias según los siguientes criterios: hay dos diferencias iguales a cero, las cuales se han ignorado; como veremos, a todos los efectos trataremos los datos anteriores como si hubiera 13 datos en lugar de 15, la diferencia mínima observada en valor absoluto, pero hay 8. Por lo tanto, a este grupo de diferencias le corresponderían los rangos del 1 al 8. Lo que hacemos es asignar el rango medio de estos 8 a todos ellos; el rango medio es 4.5, a continuación hay tres diferencias con valor 2, a las que corresponderían los rangos 9, 10 y 11. Asignamos el rango medio de las tres a cada una de ellas, que es 10, y por último tenemos dos diferencias iguales a tres, a las que corresponderían los rangos 12 y 13; asignamos a cada uno de ellos el rango medio, que es 12.5.

En la columna rangos con signo, asignamos el signo menos a las diferencias negativas y el signo más a las diferencias positivas. El signo (-), en este caso, significa que la puntuación ha aumentado, puesto que al restar ANTES-DESPUÉS las puntuaciones que han aumentado tienen diferencia negativa. El signo en esta prueba es un símbolo diferenciador y debe tenerse cuidado

con su interpretación. Sumamos los rangos con signo positivo W+ = 23.5 y los negativos W- = 67.5.

Cuadro 11.3: Calificaciones de Estadística

Antes	Después	Signo	Diferencias	Rangos	Rangos con signo
5	6	+	1	4.5	-4.5
6	6	0			
6	7	+	1	4.5	-4.5
8	9	+	1	4.5	-4.5
7	6	-	1	4.5	+4.5
5	4	-	1	4.5	+4.5
4	6	+	2	10	-10
3	3	0			
7	8	+	1	4.5	-4.5
5	8	+	3	12.5	-12.5
6	4	-	2	10	+10
6	7	+	1	4.5	-4.5
3	2	-	1	4.5	+4.5
5	7	+	2	10	-10
5	8	+	3	12.5	-12.5

Aplicando la estadística z:

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{23.5 - \frac{13(13+1)}{4}}{\sqrt{\frac{13(13+1)(226+1)}{24}}} = -1.5375$$

Por lo que no podemos rechazar la hipótesis nula y no tenemos evidencia que el curso mejore las puntuaciones de estadística.

11.5. La prueba de Kolmogorov-Smirnov para dos variables independientes

Kolmogorov-Smimov idearon una prueba, válida para comparar dos variables independientes. Las variables que comparar deben ser numéricas. En el caso de dos variables independientes, la prueba pretende comprobar si las distribuciones poblacionales de las dos variables son iguales o distintas. La

prueba de dos colas es sensible a diferencias en tendencia central, dispersión y colocación.

Sean $y_1, y_2, ..., y_{n_1}$ y $z_1, z_2, ..., z_{n_2}$ dos muestras que presumimos provienen de una misma distribución. Así, la hipótesis serán:

$$\begin{aligned} H_0 &: \text{Las distribuciones son iguales} \\ H_1 &: \text{Las distribuciones son distintas} \end{aligned}$$

La estadística de contraste es D, que es la máxima diferencia entre las proporciones escalonadascalculadas para cada valor. El parámetro D vienes definido según la siguientes expresión:

$$D = max[S_1(x) - S_2(x)]$$

Donde S_1 es la proporción de valores de la primera muestra, que son iguales o menores que x, S_2 es la proprción de valores de la segunda muestra que son iguales o menores que x. La diferencia anterior se calcula para todos los valores y el valor de la diferencia máxima es el par·metro D.

El parámetro D está tabulado y, consultando las correspondientes tablas, podremos comprobar si las diferencias son o no significativas. Cuando las dos muestras son mayores que 40 casos, podemos utilizar el siguiente estadístico para resolver el contraste de hipótesis:

$$\chi^2 = 2 \cdot D^2 \left(\frac{n_1 n_2}{n_1 + n_2} \right) \sim \chi^2_{[2]}$$

12

C N

12.1. Introducción y motivación

13

Introducción al Aprendizaje Estadístico

13.1. Motivación e introducción

El aprendizaje estadístico juega un rol essencial en muchas áreas de la ciencia, finanzas y la industtria. Algunos ejemplos son:

1. Predecir si un paciente–que se encuetra hospitalizado debido a un ataque al corazón–tenndrá un segundo ataque. La predicción estará basada en métricas demográficas, de la dieta y de regitros clínicos.

2. Predecir el precio de una acción en los siguienntes 6 meses; considerando la base de las medidas de desempeño de la compañía y de otros datos económicos.

3. Identificar los números en la digitalización de formas escritas a mano.

4. Identificar los factores de riesgo para el cáncer de próstata, basados en datos clínicos y de otras variables demográficas.

El aprendizaje estadístico comprende a un conjunto de herramientas para modelar y entender conjuntos de datos complejos. También se le conoce como Machine Learning (ML), el cual conjuga el desarrollo reciente en el área de la estadística junto con el crecimiento en paralelo de la computación.

El aprendizaje estadístico considera muchos métodos convencionales y de uso amplio como análisis de regresión, clasificación, árboles de decisión,

etc. También se refiere a una amplia gama de herramientas para entender o interpretar datos clasificadas como supervisadas y no supervisadas.

- El aprendizaje estadístico supervisado involucra la construcción de un modelo estadístico para predecir o estimar un resultado (o variable que se pede supervisar) basado en uno o más variables explicativas.

- El aprendizaje no supervisado considera variables explicativas, pero los resultados observados no son una variable explicada que sea suceptible de supervisión.

Ahora introduzcamos un poco de notación. Denotaremos a una variable independiente o explicativa con x_{ij}, si dicha variable es un conjunto de variables acomodadas en un vector utilizaremos $\mathbf{X}_i$; en estos casos denotaremos a un elemento o variable del vector $\mathbf{X}_i$ como x_{ik}, donde $i = 1, 2, \ldots, n$ denota a los individuos en la muestra y $k = 1, 2, \ldots, K$ denota al n úmero de variables. Por convención diremos que $x_{i1} = 1$ para todo $i = 1, 2, \ldots, n$, ya que en dicha variable consideraremos al término constante en la regresión.

Por su parte, los resultados, variables dependientes o variables de respuesta se denotarán como:

- y_i denotará una respuesta que es una cantidad continua

- g_i denotará una respuesta cualitativa, discreta o de grupo

Así, con $\mathbf{X}$, y $\mathbf{Y}$ y $\mathbf{G}$ denotaremos al a la matriz y vector columna que contiene a todos los valores de las variables dependientes y de respuesta apiladas para cada uno de los elementos en la muestra indexados con i.

Usaremos mayúsculas como X, Y o G para representar a los aspectos genéricos de las variables. De esta forma, debe ser claro que las letras minúsculas seran empleadas para representar a valores observados de las variables, así el valor observado de la variable k-ésima para el elemento de la muestra i-ésimo en $\mathbf{X}$ será representado como x_{ij}. Finalmente, con la notación $\hat{Y}$ o $\hat{G}$ representaremos a los valores estimados o predecidos.

El aprendizaje estadístico parte del establecimiento de que una variable dependiente Y es una función de un conjunto de variables explicativas $\mathbf{X} = [X_1, X_2, \ldots, X_p]$. De esta forma platearemos:

$$Y = f(\mathbf{X}) + \varepsilon \tag{13.1}$$

Así, el aprendizaje estadístico se trata de un conjunto de aproximaciones para f. ¿Por qué estimar f? La respuesta es por predicción y por inferencia. La más común de ambas razones es la predicción, con una predicción de Y podríamos establecer:

$$\hat{Y} = \hat{f}(\mathbf{X}) \tag{13.2}$$

13.2. Modelos linealles y el procedimiento de mínimos cuadrados

Los modelos lineales han sido utilizados de forma recurrente en estadística en por lo menos los últimos 30 años y sin duda se mantendrá como una de las herramientas más utilizadas.

Sean un conjunto de observaciones indexadas por $i = 1, 2, \ldots, n$ y un vector de variables independientes o explicativas X descrita de la forma:

$$\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$$

Así, podemos describir a la variable Y (o y_i para una observación especifica) con una ecuación:

$$\hat{y}_i = \mathbf{X}_i\boldsymbol{\beta}$$

Donde $i = 1, 2, \ldots, n$ y el vector $\boldsymbol{\beta}$ incluye a un término constante.

Visto de esta forma, a partir de la función lineal $f(X) = \mathbf{X}'\boldsymbol{\beta}$ podemos determinar el vector gradiente $f'(X) = \boldsymbol{\beta}$, que es la ruta por la que la función $f(X)$ crece más rápidamente.

¿cómo estimamos a $\boldsymbol{\beta}$?, con un procedimiento de mínimos cuadrados visto en la sección previa:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

De igual forma, son validos los análisis de pruebas de hipótesis t y F descritos anteriormente.

13.3. Método de regresiones restringidas

Estte tipo de métodos permite restringir los posibles valores de las estimacionnes del parámetro $\boldsymbol{\beta}$ a un subconjunto seleccionado. Veámos dos casos particulaes i) una regresión de cresta destringida (ridge) y ii) la regresión Lasso.

13.3.1. Regresión Ridge

La regresión de ridge restringe los coeficientes de la regreesión mediante la imposición de una penalización en su magnitud. Los coeficientes esttimados por este método resultan de resolvel el problema:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i \boldsymbol{\beta})^2 + \lambda \sum_{k=2}^{K} \beta_k^2 \right] \tag{13.3}$$

Donde $\lambda \geq 0$ y suponemos que β_1 es el término constante de la regresión.

La ideea de la penalización de los paramétros se deriva de la forma equivalente del problema en la ecuación (13.3) es:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i \boldsymbol{\beta})^2 \right]$$

Sujeto a:

$$\sum_{k=2}^{K} \beta_k^2 \leq t$$

Note que la penalización no aplica al término constante, ¿por qué?, para garentizar que la estimación del hiperplano asociado pasa por la media de Y y no por el cero (0). ¿Qué implicaciones tiene para la estimación?

El término constante se estima considerando que este toma el valor de la media de Y dado por:

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Para los restantes coeficientes los determinaremos mediante un procedimiento dado por:

$$\min_{\boldsymbol{\beta}_R, \lambda} [(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R)'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R) + \lambda \boldsymbol{\beta}_R' \boldsymbol{\beta}_R] \tag{13.4}$$

Donde $\boldsymbol{\beta}_R$ contiene sólo las pendientes. Pare determinar un valor estimado debemos resolver el problema descrito en la ecuación (13.4):

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_R} S(\boldsymbol{\beta}_R) &= \frac{\partial}{\partial \boldsymbol{\beta}_R} ((\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R)'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R) + \lambda \boldsymbol{\beta}_R' \boldsymbol{\beta}_R) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}_R} (\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}_R + \boldsymbol{\beta}_R' \mathbf{X}'\mathbf{X}\boldsymbol{\beta}_R + \lambda \boldsymbol{\beta}_R' \boldsymbol{\beta}_R) \\
&= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_R + 2\lambda \boldsymbol{\beta}_R
\end{aligned}$$

Determinando el mínimo:

$$\begin{aligned} -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_R^{Ridge} + 2\lambda\hat{\boldsymbol{\beta}}_R^{Ridge} &= 0 \\ -\mathbf{X}'\mathbf{Y} + (\mathbf{X}'\mathbf{X} + \lambda\mathbb{I}_{K-1})\hat{\boldsymbol{\beta}}_R^{Ridge} &= 0 \end{aligned}$$

Por lo tanto, el estimador estará dado por:

$$\hat{\boldsymbol{\beta}}_R^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbb{I}_{K-1})^{-1}\mathbf{X}'\mathbf{Y} \tag{13.5}$$

Así, el problema de la estimación es que tiene 1 grado de libertad, λ, que es un valor de penalización y que resultta arbitrario.

13.3.2. Regresión Lasso (Least Absolute Shrinkage and Selection Operator)

La regresión Lasso se define por la solución al problema:

$$\hat{\boldsymbol{\beta}}^{Lasso} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right] \tag{13.6}$$

Sujeto a:

$$\sum_{k=2}^{K} |\beta_k| \leq t$$

Donde $\lambda \geq 0$ y suponemos que β_1 es el término constante de la regresión. El término constante se estima considerando que este toma el valor de la media de Y dado por:

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

La ideea de la penalización de los paramétros se deriva de la forma equivalente del problema en la ecuación (13.6) es:

$$\hat{\boldsymbol{\beta}}^{Lasso} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda \sum_{k=2}^{K} |\beta_k| \right] \tag{13.7}$$

13.4. Modelos lineales de clasificación

A continuación, analizamos el problema de clasificación y nos enfocamos en métodos lineales para casificación. Supongamos que nuestro predictor está dado por una función $G(X)$ que toma valores en un conjunto de clases o tipos indexado por $j = \{0, 1, 2, \ldots, J\}$. Este tipo de modelos se les conoce como discriminante por probabilidades:

$$\mathbb{P}(G = j|X = x) \tag{13.8}$$

De esta forma, supongamos un problema de determinar la probabilidad de dos clases $j = \{1, 2\}$, para cada un tendríamos:

$$\begin{aligned} \mathbb{P}(G = 1|X = x) &= \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1 + e^{\mathbf{X}\boldsymbol{\beta}}} \\ \mathbb{P}(G = 2|X = x) &= \frac{1}{1 + e^{\mathbf{X}\boldsymbol{\beta}}} \end{aligned}$$

De esta forma garantizamos que $1 = \mathbb{P}(G = 1|X = x) + \mathbb{P}(G = 2|X = x)$. Así, podemos tomar una transformación *Logit* de $ln(\theta/(1-\theta))$, ya que estamos ante el caso de un modelo binomial. En nuestro caso concreto tendríamos el caso:

$$ln\left\{\frac{\mathbb{P}(G = 1|X = x)}{\mathbb{P}(G = 2|X = x)}\right\} = \mathbf{X}\boldsymbol{\beta} \tag{13.9}$$

A este modelo se le conoce como modelo *Logit* o regresión logística.

Regresión Logística

La regresión logística aparece por la necesidad de un modelo para determinar las probabilidades de ocurrencia de $J + 1$ clases en función de una relación lineal de X y al mismo tiempo garantizar que la estimación respetará que la estimación se ubique en el intervalo $[0, 1]$.

Supongamos un matriz $\mathbf{X}$ que contiene información de las variables explicativas y en su primera conlumna incluye al término constante. Consideremos un caso generalizado en el que la clases o valores que puede tomar $G(X)$ son $j = 0, 1, 2, \ldots, J$. Considerando esto, el modelo a continuación plateada será conocido como el modelo multinomial *logit*, para lo cual las probabilidades para cada clase estarán dadas por:

$$\mathbb{P}(G = j|\mathbf{X} = \mathbf{x}) = \frac{e^{\mathbf{X}\boldsymbol{\beta}_j}}{1 + \sum_{h=1}^{J} e^{\mathbf{X}\boldsymbol{\beta}_h}} \tag{13.10}$$

Donde $j = 1, 2, \ldots, J$ y $\boldsymbol{\beta}_j$ es un vector de dimensión $K \times 1$. Por su parte la probabilidad de el caso en que la clase es 0, de forma complementaria tendríamos:

$$\mathbb{P}(G = 0|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + \sum_{h=1}^{J} e^{\mathbf{X}\boldsymbol{\beta}_h}} \tag{13.11}$$

Estimación

Los modelos de regresión logítstica soos estimados usualmente por métodos de Máxima Verosimilitud mediante el uso de $\mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i)$, para $j = 0, 1, \ldots, J$ e $i = 1, 2, \ldots, n$. Así, plateamos la función de verosimilitud como:

$$L(\theta) = \Pi_{i=1}^{n} \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \tag{13.12}$$

Utilizando una transformación logarítmica:

$$lnL(\theta) = l(\theta) = \sum_{i=1}^{n} ln\mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \tag{13.13}$$

Por simplicidad continuaremos con el caso dee una estimación para un modelo de respuesta binomial en el cual $G = \{0, 1\}$. De esta forma establezcamos la función de verosimilitud como:

$$\begin{aligned} lnL(\theta) = l(\theta) &= \sum_{i=1}^{n} G_i ln\mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \\ &\quad +(1 - G_i)ln(1 - \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i)) \\ &= \sum_{i=1}^{n} G_i \mathbf{X}_i \boldsymbol{\beta} - ln(1 + e^{\mathbf{X}_i\boldsymbol{\beta}}) \end{aligned}$$

Las condiciones de primer orden son:

$$\frac{\partial l(\theta)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(G_i \mathbf{X}_i - \mathbf{X}_i \frac{e^{\mathbf{X}_i\boldsymbol{\beta}}}{1 + \mathbf{X}_i\boldsymbol{\beta}} \right) = (G_i - \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i))\mathbf{X}_i \tag{13.14}$$

Así, la solucción estará dada por aquellas $\boldsymbol{\beta}$ que satisfacen:

$$\sum_{i=1}^{n} \left(G_i - \frac{e^{\mathbf{X}_i\boldsymbol{\beta}}}{1 + \mathbf{X}_i\boldsymbol{\beta}} \right) \mathbf{X}_i = 0 \tag{13.15}$$

Figura 13.1: División del conjunto de datos, retomado de Hastie, Tibshirani, y Friedman (2017, p. 222) Hastie, Tibshirani y Friedman 2017

Ajuste y separación del conjunto de datos

Supomgamos la variable objetivo Y, un vector de variables explicativas o variables 'input' X y un modelo predictivo $\hat{f}(X)$ que es estimado a partir de un conjunto de entrenamiento τ.

Definiremos la función de pérdida derivada de la estimación y capturada por los errores entre Y y $\hat{f}(X)$ estará dada por:

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{error cuadrático} \\ |Y - \hat{f}(X)| & \text{error absoluto} \end{cases} \tag{13.16}$$

De esta forma podemos establecer un error cuadrático como:

$$\begin{aligned} \mathbb{E}[Y - \hat{Y}]^2 &= \mathbb{E}[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= \mathbb{E}[f(X) - \hat{f}(X)]^2 + Var(\varepsilon) \end{aligned}$$

Donde $\mathbb{E}[f(X) - \hat{f}(X)]^2$ es el único componente reducible. Para hacer predicciones requerimos de un connjunto de datos de entrenamiento y otro más de prueba – en el primero estimamos $f(\cdot)$ y enn el segundo hacemos predicciones –.

Finalmente, la condición de inferencia nos permite construir pruebas de hipótesis, estimadores que cumple ciertas propiedades, intervalos de confianza, etc. Así, dividiremos al conjunto de datos conforme se describe en la Figura 13.1.

13.5. Aprendizaje no supervisado

13.6. Otros

El Cuadro (13.1) muestra los resultados para este nuevo modelo. No lo motramos en esta sección, pero ambos modelos reportados tienen raices de sus respectivos polinomios característicos menores a 1 en valor absoluto.

Cuadro 13.1: Sample table

S. No.	Column#1	Column#2	Column#3
1	50	837	970
2	47	877	230
3	31	25	415
4	35	144	2356
5	45	300	556

Bibliografía

[AFK17] Alan Agresti, Christine A. Franklin y Bernhard Klingenberg. *The Art and Science of Learning from Data, Global Edition*. Mexico City: Pearson, 2017.

[Gre12] William H. Greene. *Econometric Analysis*. Mexico City: Prentice Hall, 2012.

[HTF17] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. LLC: Springer, 2017.

[LM12] Richard J. Larsen y Morris L. Marx. *An Introduction to Mathematical Statistics and its Applications*. Estados Unidos: Prentice Hall, 2012.

[MM14] Irwin Miller y Marylees Miller. *John E. Freud's Mathematical Statistics with Applications*. Estados Unidos: Pearson, 2014.

[MGB73] Alexander M. Mood, Franklin A. Graybill y Duane C. Boes. *Introduction to the Theory of Statistics*. Estados Unidos: McGraw-Hill, 1973.

[WMS16] Dennis Wackerly, William Mendenhall y Richard L. Scheaffer. *Mathematical Statistics with Applications*. Estados Unidos: McGraw-Hill, 2016.

[Was04] Larry Wasserman. *All of Statistics. A Concise Course in Statistical Inference*. Estados Unidos: Springer Link, 2004.

Apéndice A

Algunos teoremas y resultados relevantes

A.1. Convergencia