

Notas de Clase: Estadística

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

Benjamín Oliva¹

Draft Agosto 2025

¹benjov@ciencias.unam.mx y <https://github.com/benjov>

Documento siempre proceso de mejora.
Comentarios, siempre serán bienvenidos...
Este documento servirá para el curso de
Estadística de la Licenciatura en Economía.

Índice general

1. Estadística descriptiva	5
1.1. Variables Aleatorias Discretas y Continuas	5
1.2. Funciones de Variables Aleatorias	5
2. Probabilidad	7
2.1. Espacios muestrales y eventos	7
2.1.1. Expresando eventos graficamente: Diagramas de Venn .	12
2.2. Definición de Probabilidad	15
2.2.1. Propiedades de las Probabilidades	15
2.2.2. Probabilidad en espacios muestrales finitos	18
2.3. Independencia de eventos	25
2.4. Probabilidad condicional	29
2.5. Teorema de Bayes	32
3. Variables Aleatorias	37
3.1. Introducción	37
3.2. Funciones de distribución de probabilidad y Funciones de pro- babilidad	40
3.3. Variables Aleatorias Discretas y Continuas	42
3.4. Algunas variables aleatorias importantes discretas	48
3.5. Algunas variables aleatorias importantes continuas	52
3.6. Distribuciones bivariadas	53
3.6.1. Distribuciones discretas	54
3.6.2. Distribuciones continuas	55
3.6.3. Distribuciones marginales	56
3.6.4. Variables Aleatorias Independientes	59
3.6.5. Distribuciones Condicionales	62
3.6.6. Transformaciones de Variables Aleatorias	63

4. Esperanza	65
4.1. Valor esperado de una variable aleatoria	65
4.2. Propiedades de la esperanza	67
4.3. Varianza y covarianza	69
4.4. Esperanza condicional	76
4.5. Funciones generadoras de momentos	79
4.5.1. Momentos	79
4.5.2. Funciones Generadoras de Momentos	81
5. Desigualdades y algunos resultados importantes	87
5.1. Desigualdades	87
5.2. Tipos de convergencia	91
5.2.1. Tipos de convergencia	92
5.2.2. La Ley de los Grandes Números	92
5.2.3. El Teorema del Límite Central	95
6. Modelos, Inferencia y Aprendizaje Estadísticos	97
6.1. Introducción	97
6.2. Modelos paramétricos y No paramétricos	98
6.3. Inferencia paramétrica	98
7. Estimación Puntual	101
7.1. Definiciones y ejemplos	101
7.2. Propiedades de los estimadores	104
7.2.1. Insesgamiento	104
7.2.2. Eficiencia	116
7.2.3. Consistencia	122
7.2.4. Suficiencia	129
7.3. Métodos de estimación: Método de momentos y Método de máxima verosimilitud	132
7.3.1. Método de Momentos	133
7.3.2. Método de Máxima Verosimilitud	136
8. Estimación por Intervalo	141
8.1. Introducción	141
8.2. Intervalos de confianza para medias	145
8.3. Intervalos de confianza para diferencias de medias	150

8.4.	Intervalos de confianza para proporciones y diferencia de proporciones	154
8.5.	Intervalos de confianza para varianzas y razón de varianzas . .	159
9.	Pruebas de Hipótesis	163
9.1.	Elementos de una prueba estadística	163
9.2.	Potencia de una prueba	165
9.3.	Nivel de significancia de una prueba de hipótesis	166
9.4.	Pruebas para medias, varianzas y proporciones	168
9.4.1.	Pruebas para medias	168
9.4.2.	Pruebas para diferencias de medias	172
9.4.3.	Pruebas para varianzas	174
9.4.4.	Pruebas para proporciones	177
9.5.	Pruebas de razón de verosimilitud	178
9.6.	Bondad de ajuste	181
10.	Análisis de varianza y diseño de experimentos	183
10.1.	Análisis de varianza: Introducción y motivación	183
10.2.	Planteamiento de la prueba de hipótesis	186
10.3.	Tablas ANOVA	188
10.4.	Diseño del Análisis de Varianza para Bloques	190
10.5.	La prueba F para diseño de Bloques aleatorizados	191
10.6.	Análisis de datos categóricos	198
10.6.1.	Introducción y motivación	198
10.6.2.	Prueba de Ji Cuadrado	199
10.7.	Consideraciones para el diseño de experimentos	201
10.7.1.	Elementos que se relaciona con el tamaño de la muestra	201
10.7.2.	Diseño de experimentos para aumentar la precisión . .	201
10.7.3.	El experimento de observaciones pareadas	201
11.	Regresión Lineal	205
11.1.	Introducción y motivación	205
11.2.	El concepto de regresión entre dos variables	205
11.3.	Modelo de regresión lineal	211
11.3.1.	Supuestos del modelo de regresión lineal	211
11.3.2.	Estimación de los parámetros del modelo lineal bivariado	214
11.4.	Estimación de parámetros por mínimos cuadrados	217
11.5.	Modelos no lineales	220

11.5.1. Regresión exponencial	221
11.5.2. Regresión logarítmica	222
11.5.3. Funciones logísticas	222
11.6. Modelo de regresión lineal múltiple	225
12. Estadística no paramétrica	237
12.1. Introducción y motivación	237
12.2. Estimaciones de funciones de densidad via Kernel (Núcleo) . .	237
12.3. La prueba de signo	238
12.4. Prueba de Wilcoxon	243
12.5. La prueba de Kolmogorov-Smirnov para dos variables inde- pendientes	246
13. C N	249
13.1. Introducción y motivación	249
14. Introducción al Aprendizaje Estadístico	251
14.1. Motivación e introducción	251
14.2. Modelos lineales y el procedimiento de mínimos cuadrados . .	253
14.3. Método de regresiones restringidas	253
14.3.1. Regresión Ridge	254
14.3.2. Regresión Lasso (Least Absolute Shrinkage and Selec- tion Operator)	255
14.4. Modelos lineales de clasificación	256
14.5. Aprendizaje no supervisado	258
14.6. Otros	258
Bibliografía	261
A. Algunos teoremas y resultados relevantes	263
A.1. Convergencia	263

Índice de figuras

1.	Esquema de la relación entre la Probabilidad y la Minería de Datos y la Inferencia Estadística, retomado de Wasserman (2004, p. ix) Wasserman 2004	2
2.1.	Ejemplos de Diagramas de Venn, retomado de Larsen (2012, p24) Larsen y Marx 2012	12
2.2.	Diagrama de Venn de $E = (A \cap B^C) \cup (B \cap A^C)$, retomado de Larsen (2012, p24) Larsen y Marx 2012	13
2.3.	Diagrama de Venn de $E = (A \cap B)^C$, retomado de Larsen (2012, p24) Larsen y Marx 2012	13
2.4.	Diagrama de Venn que ilustra que $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, retomado de Larsen (2012, p25) Larsen y Marx 2012	14
2.5.	Diagrama que ilustra el proceso de extraer al azar, retomado de Rincón (2012, p17) Rincón 2012	20
2.6.	Diagrama que ilustra el Triángulo de Pascal, retomado de Rincón (2012, p20) Rincón 2012	24
2.7.	Resumen de fórmulas de conteo, retomado de Rincón (2012, p22) Rincón 2012	25
2.8.	Ejemplo de un evento restringido, retomado de Larsen (2012, p32) Larsen y Marx 2012	29
3.1.	Ejemplo de una CDF, retomado de Wasserman (2004, p32) Wasserman 2004	41
3.2.	Ejemplo de una PDF, retomado de Wasserman (2004, p23) Wasserman 2004	43
3.3.	CDF de una Uniforme (0,10), retomado de Larsen (2012, p129) Larsen y Marx 2012	46
3.4.	CDF de una Uniforme (0,1), retomado de Wasserman (2004, p24) Wasserman 2004	47

7.1.	Gráfica de $p^2(1 - p)$	103
7.2.	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	106
9.1.	Ilustración 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	166
9.2.	Hipótesis de una cola (derecha) y su zona de rechazo. Retomado de Miller y Miller (2014; p. 360) I. Miller y M. Miller 2014	167
9.3.	Hipótesis de una cola (izquierda) y su zona de rechazo. Retomado de Miller y Miller (2014; p. 361) I. Miller y M. Miller 2014	168
9.4.	Hipótesis de dos colas y sus zonas de rechazo. Retomado de Miller y Miller (2014; p. 360) I. Miller y M. Miller 2014	169
10.1.	Dos diseños experimentales diferentes	191
11.1.	Relación entre Latitud y Velocidad del Viento	206
11.2.	Relación entre Latitud y Temperatura	206
11.3.	Relación entre Latitud y Nubosidad	207
11.4.	Relación entre Latitud y Humedad	207
11.5.	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 $y = \mu_{Y x} = \mathbb{E}[Y x]$	212
11.6.	Ilustración del error de estimación, retomado de Greene (2012, pp. 40) Greene 2012	218
11.7.	Funciones exponenciales, Retomado de Larsen y Marx (2012, p. 532) Larsen y Marx 2012	222
11.8.	Funciones logarítmicas, Retomado de Larsen y Marx (2012, p. 536) Larsen y Marx 2012	223
11.9.	Funciones logísticas, Retomado de Larsen y Marx (2012, p. 540) Larsen y Marx 2012	224
11.10.	Ilustración de la proyección de y en x_1 y x_2 y el significado del error e , retomado de Greene (2012, pp. 20) Greene 2012 . .	226

11.11 Ilustración del hiperplano generado de regresar X_1 y X_2 en Y , retomado de Hastie, Tibshirani y Friedman (2009, pp. 45) Hastie, Tibshirani y Friedman 2017	226
11.12 Ilustración de la proyección 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	227
11.13 Gráfica de Precios y Cantidades, y su recta de regresión	236
12.1. Kernel Epanechnikov	239
12.2. Kernel Gauss	239
14.1. División del conjunto de datos, retomado de Hastie, Tibshirani, y Friedman (2017, p. 222) Hastie, Tibshirani y Friedman 2017	258

Índice de cuadros

8.1. Uso de las transformaciones Z y T	150
8.2. Estadísticas de dos muestras aleatorias	152
8.3. Tabla de muestras aleatorias	153
8.4. Dos muestras aleatorias de poblaciones binomiales	158
8.5. Estadísticas para razón de varianzas	161
9.1. Tabla Tipo de Error	165
9.2. Estadísticas de dos muestras aleatorias	173
9.3. Estadísticas de dos muestras aleatorias	174
9.4. Table para razón de varianzas	177
10.1. Tabla ANOVA	188
10.2. Tabla Fumadores	189
10.3. Tabla ANOVA	190
10.4. Tabla de bloques y diseño experimental	194
10.5. Tabla ANOVA para Bloques	196
10.6. Medidas de concentración (mg / ml)	197
10.7. Tratamientos	197
10.8. Tabla ANOVA para Bloques	198
10.9. Número de accidentes por semana	200
10.10Número de accidentes por semana y calculos relacionados . . .	200
11.1. Tabla de Precios y Cantidades	235
11.2. Regresión resultados	236
12.1. Precios de Gasolina Regular (Pesos por Litro)	241
12.2. Calificaciones de Estadística	245
12.3. Calificaciones de Estadística	246
14.1. Sample table	259

Introducción, motivación y alcance del documento

Este documento servirá como guía para impartir el curso de Estadística en la Licenciatura en Economía.

En este documento buscamos implementar un enfoque moderno en el cual la estadística, la minería de datos y el aprendizaje de máquina (o aprendizaje estadístico) sean todas herramientas para coleccionar y analizar datos. Durante mucho tiempo se ha considerado a la estadística de forma separada de las técnicas de aprendizaje de máquina y minería de datos, ya que estas 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 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 la relación que guarda la estadística y la probabilidad.

En el curso queremos discutir los documentos de la bibliografía 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.

Conocimientos previos

Cálculo Diferencial e Integral y Álgebra Lineal.

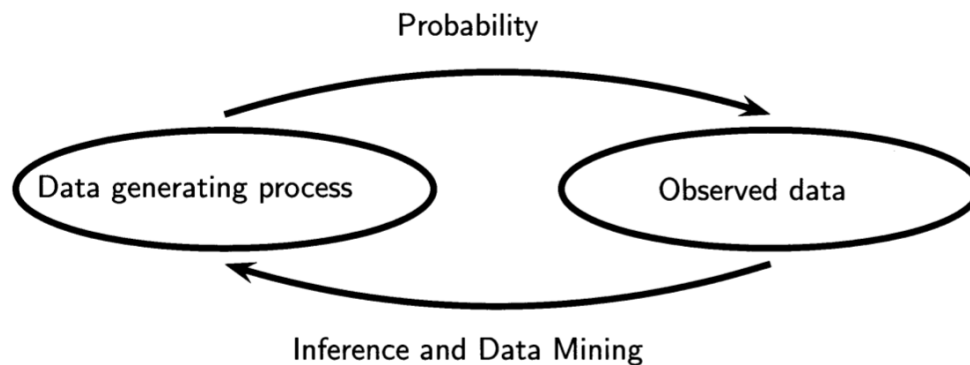


Figura 1: Esquema de la relación entre la Probabilidad y la Minería de Datos y la Inferencia Estadística, retomado de Wasserman (2004, p. ix) Wasserman 2004

Recursos en línea y otros materiales

Materiales recomendados durante el curso y autoaprendizaje.

Perspectiva histórica¹

Nadie sabe dónde o cuándo surgió por primera vez la noción de azar; se desvanece en nuestra prehistoria. Sin embargo, la evidencia que vincula a los primeros humanos con dispositivos para generar eventos aleatorios es abundante. La falta de registros históricos desdibuja la distinción inicialmente establecida entre ceremonias de adivinación y juegos recreativos. Sin embargo, entre las sociedades más recientes, el juego surgió como una entidad distinta, y su popularidad era irrefutable. Los griegos y los romanos eran jugadores consumados, al igual que los primeros cristianos.

Se han perdido las reglas de muchos de los juegos griegos y romanos, pero podemos reconocer el linaje de ciertas diversiones modernas en lo que se jugaba durante la Edad Media. El juego de dados más popular de ese período se llamaba azar, nombre que deriva del árabe *al zhar*, que significa “un dado”. Se cree que el azar fue traído a Europa por los soldados que regresaban de las Cruzadas; sus reglas son muy parecidas a las de los dados

¹Retomado de Larsen y Marx 2012, Cap. 1

de hoy en día. Por su parte, las cartas se introdujeron por primera vez en el siglo XIV e inmediatamente dieron origen a un juego que era una forma temprana del póquer.

El primer caso en el que alguien conceptualizó la probabilidad en términos de un modelo matemático ocurrió en el siglo XVI. Eso significa que pasaron más de dos mil años de juegos de dados, juegos de cartas y juegos de mesa antes de que alguien finalmente tuviera la idea de escribir incluso la más simple de las abstracciones probabilísticas. Los historiadores coinciden en general en que la probabilidad, como disciplina, tuvo un comienzo difícil debido a su incompatibilidad con dos de las fuerzas más dominantes en la evolución de nuestra cultura occidental: la filosofía griega y la teología cristiana primitiva.

Para empeorar las cosas, el antiempirismo que permeaba el pensamiento griego era tal que, para ellos, el conocimiento no era algo que se pudiera obtener por medio de la experimentación. Era mejor razonar una cuestión de manera lógica que buscar su explicación en un conjunto de observaciones numéricas. Juntas, estas dos actitudes tenían un efecto paralizante: los griegos no tenían motivación para pensar en la probabilidad en un sentido abstracto, ni se enfrentaban a los problemas de interpretación de datos que podrían haberlos encaminado hacia un cálculo de probabilidades.

Si las perspectivas para el estudio de la probabilidad eran sombrías bajo los griegos, empeoraron aún más cuando el cristianismo amplió su esfera de influencia. Los griegos y los romanos al menos aceptaban la existencia del azar. Sin embargo, creían que sus dioses no podían o no querían involucrarse en asuntos tan mundanos como el resultado de la tirada de un dado. Sin embargo, para los primeros cristianos, no existía tal cosa como la casualidad: cada evento que sucedía, sin importar cuán trivial fuera, se percibía como una manifestación directa de la intervención deliberada de Dios.

Fue en el siglo XVI cuando la probabilidad resurgió con Gerolamo Cardano. Cardano había aprovechado el principio más básico de la probabilidad. El modelo que descubrió puede parecer trivial en retrospectiva, pero representó un gran paso adelante: fue el primer caso registrado de alguien que calculó una probabilidad teórica, en lugar de una empírica. Aun así, el impacto real del trabajo de Cardano fue mínimo.

La fecha que muchos historiadores (aquellos que no son partidarios de Cardano) citan como el “inicio” de la probabilidad es 1654. En París, un jugador adinerado, el Chevalier de Mere, planteó a varios matemáticos destacados, entre ellos Blaise Pascal, una serie de preguntas, la más conocida de las cuales es el problema de los puntos. Dos personas, A y B, acuerdan jugar

una serie de juegos justos hasta que una de ellas haya ganado seis juegos. Cada una de ellas ha apostado la misma cantidad de dinero, con la intención de que el ganador se lleve todo el bote. Pero supongamos que, por la razón que sea, la serie termina prematuramente, momento en el que A ha ganado cinco juegos y B tres. ¿Cómo se deben dividir las apuestas?²

Pascal se sintió intrigado por las preguntas de De Mere y compartió sus pensamientos con Pierre Fermat, un funcionario de Toulouse y probablemente el matemático más brillante de Europa. Fermat respondió amablemente, y de la ahora famosa correspondencia Pascal-Fermat surgió no solo la solución al problema de los puntos, sino también la base para resultados más generales. Más importante aún, la noticia de lo que Pascal y Fermat estaban trabajando se difundió rápidamente.

²La respuesta correcta es que A debería recibir siete octavos de la cantidad total apostada. Para resolverlo: Supongamos que se reanudara la competición. ¿Qué escenarios llevarían a que A fuera el primero en ganar seis partidas?

1

Estadística descriptiva

1.1. Variables Aleatorias Discretas y Continuas

1.2. Funciones de Variables Aleatorias

2

Probabilidad

2.1. Espacios muestrales y eventos

La probabilidad es un lenguaje matemático para cuantificar la incertidumbre. En su definición clásica, debida a Cardano, aplicaba a casos en los que (1) hay un número finito de resultados, y (2) todos los resultados son igualmente posibles. Así, los ejemplos clásicos son el lanzamiento de una moneda o de un dado, por ejemplo. Ahora, iniciemos con algunas definiciones.

Definición 2.1 *Experimento*. *Un experimento es cualquier procedimiento que:*

1. *puede ser repetido, teóricamente un número infinito de veces, y*
2. *tiene un conjunto de posibles resultados bien definido.*

Definición 2.2 *Espacio Muestral*. *El espacio muestral Ω es el conjunto de los posibles resultados de un experimento.*

Definición 2.3 *Elementos del Espacio Muestral y Eventos*. *Los elementos $\omega \in \Omega$ se denominan resultados, realizaciones o elementos del espacio muestral, Ω . Los subconjuntos de Ω 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 águila en cada uno de los lanzamientos de las dos monedas: $E = \{AA, AS, SA\}$.

Ejemplo. Imaginemos el lanzamiento de un par de dados, uno rojo y otro verde. Cada resultado del lanzamiento es un par ordenado y el espacio muestral Ω puede ser representado como el siguiente:

$$\begin{aligned}\Omega = \{ & (1, 1), (1, 2), \dots, (1, 6), \\ & (2, 1), (2, 2), \dots, (2, 6), \\ & \vdots \quad \vdots \quad \vdots \quad \vdots \\ & (6, 1), (6, 2), \dots, (6, 6) \}\end{aligned}$$

Supongamos que estamos interesados en el evento ω que se define como la suma de las caras que da como resultado 7. Esos casos serían la diagonal que se compone de los casos: $(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)$.

Ejemplo. Supongamos que estamos ayudando a una televisora local a contratar 2 personas para su noticiero matutino. Asumamos que tenemos 3 candidatas femeninas (W_1, W_2, W_3) y 2 candidatos del sexo masculino (M_1, M_2) . Definamos que nuestro experimento es contratar a dos personas. Así, el espacio muestral tendrá 10 elementos:

$$\begin{aligned}\Omega = \{ & (W_1, W_2), (W_1, W_3), (W_2, W_3), (W_1, M_1), \\ & (W_1, M_2), (W_2, M_1), \dots, (M_1, M_2) \}\end{aligned}$$

¿Importa el orden en que son seleccionadas las personas? Sí.

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 infinitas veces. En este caso, el espacio muestral estará dado por: $\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \dots)\} : \omega_i \in \{A, S\}$. 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 Águila aparece hasta el tercer lanzamiento de la moneda: $E = \{(\omega_1, \omega_2, \omega_3, \dots) : \omega_1 = S, \omega_2 = S, \omega_3 = A, \omega_i \in \{A, S\}, \forall i > 3\}$.

Definición 2.4 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 $\neg A$.

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

Ejemplo. Sea A el conjunto de los pares (x, y) para los cuales se cumple:

$$x^2 + y^2 < 1$$

Dibuje y defina el complemento de A o A^c . Al respecto, queda claro que el complemento sería definido como:

$$A^c = x^2 + y^2 \geq 1$$

Definición 2.5 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 extender esta definición a casos en los que observamos una secuencia infinita de eventos A_1, A_2, A_3, \dots y obtener:

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

Definición 2.6 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 extender esta definición a casos en los que observamos una secuencia infinita de eventos A_1, A_2, A_3, \dots y obtener:

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

Ejemplo. Supongamos que extraemos una carta de un mazo de cartas de póker. Sea A el evento de que un As es seleccionado:

$$A = \{\text{As de corazones, As de diamantes, As de tréboles, As de espadas}\}$$

Por otro lado, sea B el evento de que una carta con palo de corazones es seleccionada:

$$B = \{2 \text{ de corazones, } 3 \text{ de corazones, } \dots, \text{As de corazones}\}$$

Entonces

$$A \cap B = \{\text{As de corazones}\}$$

y

$$A \cup B = \{2 \text{ de corazones}, 3 \text{ de corazones}, \dots, \text{As de corazones},$$

$$\text{As de diamantes}, \text{As de tréboles}, \text{As de espadas}, \}$$

Ejemplo. Sea A el conjunto definido como:

$$A = \{x : x^2 + 2x = 8\}$$

Por otro lado, sea B el conjunto definido como:

$$B = \{x : x^2 + x = 6\}$$

Determinemos los conjuntos dados por $A \cap B$ y $A \cup B$.

Partamos de que la primera de las ecuaciones se puede factorizar como $(x+4)(x-2) = 0$, por lo que su solución será $A = \{-4, 2\}$. De forma similar, la segunda de las ecuaciones puede ser escrita como $(x+3)(x-2) = 0$, lo que implica que $B = \{-3, 2\}$. De esta forma, la respuesta al cuestionamiento sería:

$$A \cap B = \{2\}$$

$$A \cup B = \{-4, -3, 2\}$$

Definición 2.7 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 2.8 Contención de Eventos. Definimos que A está contenido 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 2.9 Número de elementos de un evento. Sea A un conjunto finito, definimos $|A|$ como el número de elementos de A .

Definición 2.10 *Eventos Disjuntos o Eventos Mutuamente Excluyentes.* Decimos que una serie de conjuntos o eventos A_1, A_2, A_3, \dots son disjuntos o mutuamente excluyentes si:

$$A_i \cap A_j = \emptyset, \forall i \neq j$$

Definición 2.11 *Partición de Ω .* Sea una serie de conjuntos o eventos disjuntos dados por A_1, A_2, A_3, \dots , definimos la partición de Ω como:

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

Definición 2.12 *Función Indicadora de A .* Definimos a la función indicadora de A por:

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

Definición 2.13 *Función (Serie) monótona creciente.* Una serie de conjuntos o eventos A_1, A_2, A_3, \dots es monótona creciente si $A_1 \subset A_2 \subset A_3 \subset \dots$ y si:

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

Función (Serie) monótona decreciente. Una serie de conjuntos o eventos A_1, A_2, A_3, \dots es monótona decreciente si $A_1 \supset A_2 \supset A_3 \supset \dots$ 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 $i = 1, 2, 3, \dots$. Entonces, ¿cuál es $A_n \longrightarrow A$ considerando las dos definiciones anteriores?

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

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

$$\begin{aligned} \cup_{i=1}^{\infty} A_i &= (0, 1) \\ \cap_{i=1}^{\infty} A_i &= \emptyset \end{aligned}$$

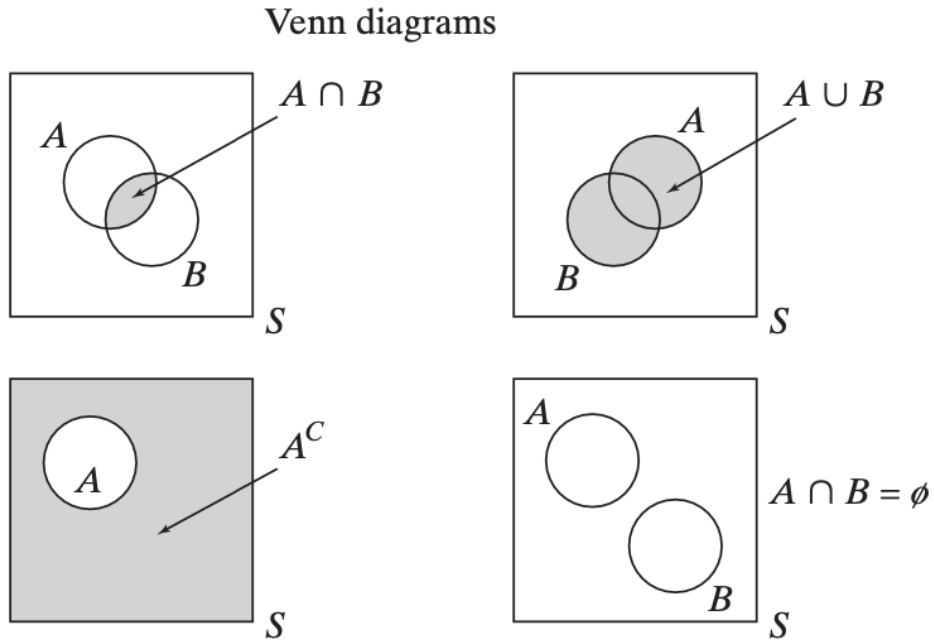


Figura 2.1: Ejemplos de Diagramas de Venn, retomado de Larsen (2012, p24)
Larsen y Marx 2012

2.1.1. Expresando eventos gráficamente: Diagramas de Venn

Las relaciones basadas en dos o más eventos a veces pueden ser difíciles de expresar utilizando únicamente ecuaciones o descripciones verbales. Un enfoque alternativo que puede ser muy eficaz es representar los eventos subyacentes gráficamente en un formato conocido como diagrama de Venn. La Figura 2.1 muestra diagramas de Venn para una intersección, una unión, un complemento y dos eventos que son mutuamente excluyentes. En cada caso, el interior sombreado de una región corresponde al evento deseado.

Ejemplo. Definamos dos eventos A y B en un espacio muestral; con frecuencia tendremos que considerar:

1. El evento de que ocurra exactamente uno de los dos.
2. El evento de que ocurra como máximo uno de los dos.

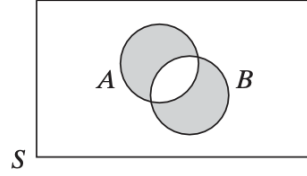


Figura 2.2: Diagrama de Venn de $E = (A \cap B^C) \cup (B \cap A^C)$, retomado de Larsen (2012, p24) Larsen y Marx 2012

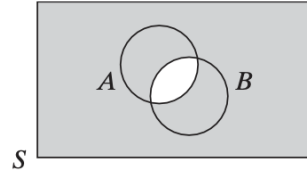


Figura 2.3: Diagrama de Venn de $E = (A \cap B)^C$, retomado de Larsen (2012, p24) Larsen y Marx 2012

Obtener expresiones para cada uno de estos es fácil si visualizamos los diagramas de Venn correspondientes.

El área sombreada en la Figura 2.2 representa el evento E de que A o B, pero no ambos, ocurran (es decir, ocurra exactamente uno de los dos).

Con solo observar el diagrama podemos formular una expresión para E . La parte de A , por ejemplo, incluida en E es $A \cap B^C$. De manera similar, la parte de B incluida en E es $B \cap A^C$. De ello se deduce que E puede escribirse como una unión:

$$E = (A \cap B^C) \cup (B \cap A^C)$$

La Figura 2.3 muestra el evento E de que ocurre como máximo uno de los dos eventos. Dado que este último incluye todos los resultados excepto los que pertenecen tanto a A como a B, podemos escribir:

$$E = (A \cap B)^C$$

Ejemplo. Sean A , B y C tres eventos cualesquiera definidos sobre un espacio muestral S . Muestre que:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

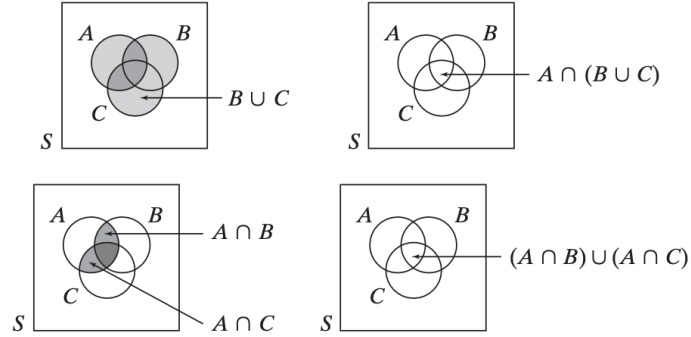


Figura 2.4: Diagrama de Venn que ilustra que $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, retomado de Larsen (2012, p25) Larsen y Marx 2012

La Figura 2.4 ilustra el proceso para mostrar la anterior afirmación. Siendo rigurosos, necesitamos probar la doble contención:

- $A \cap (B \cup C)$ está contenido en $(A \cap B) \cup (A \cap C)$
- $(A \cap B) \cup (A \cap C)$ está contenido en $A \cap (B \cup C)$

Partamos de que tomamos un elemento $x \in A \cap (B \cup C)$, entonces podemos afirmar que $x \in A$ y que $x \in (B \cup C)$. De esta forma sabemos que $x \in B$ o que $x \in C$ o en ambos.

En suma, sabemos que, por un lado, $x \in A$ y $x \in B$ o que, por otro lado, $x \in A$ y $x \in C$. Dicho de otra manera, $x \in (A \cap B)$ o $x \in (A \cap C)$. Así, $x \in (A \cap B) \cup (A \cap C)$. Por lo que hemos demostrado que $A \cap (B \cup C) \subset (A \cap B) \cup (A \cap C)$.

El caso restante es más o menos similar. Partamos de que $x \in (A \cap B) \cup (A \cap C)$. Esto significa que, por un lado, $x \in (A \cap B)$ o que, por otro lado, $x \in (A \cap C)$ (o ambos).

Así, existen dos casos. El primero, $x \in A$ y $x \in B$. El segundo, $x \in A$ y $x \in C$. Siendo la constante que, con seguridad, $x \in A$ y que, en consecuencia, $x \in B$ o $x \in C$.

Demostrándose así que, $x \in (B \cup C)$. Por lo que resulta inmediato que: $x \in A \cap (B \cup C)$ y que $(A \cap B) \cup (A \cap C) \subset A \cap (B \cup C)$.

2.2. Definición de Probabilidad

Asignemos un número real $\mathbb{P}(A)$ a todo evento A , al cual denominaremos **probabilidad** de A . También, llamaremos a \mathbb{P} como una **función de probabilidad** o una **medida de probabilidad**. Una función de probabilidad \mathbb{P} debe satisfacer 3 axiomas.

Definición 2.14 *Función de Probabilidad o Medida de Probabilidad.* Una función \mathbb{P} que asigna un número real $\mathbb{P}(A)$ a todo evento A es una **función de probabilidad** o una **medida de probabilidad** si satisface los siguientes tres axiomas:

- **Axioma 1:** $\mathbb{P}(A) \geq 0$ para cada evento A .
- **Axioma 2:** $\mathbb{P}(\Omega) = 1$
- **Axioma 3:** Si A_1, A_2, \dots son eventos disjuntos entonces:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Existen múltiples interpretaciones de $\mathbb{P}(A)$. Las dos interpretaciones más comunes son la frecuentista y la del grado de creencia. En la interpretación frecuentista, $\mathbb{P}(A)$ es la proporción de veces a largo plazo que A es cierto en repeticiones. Por ejemplo, si decimos que la probabilidad de que salga cara es $1/2$, queremos decir que si lanzamos la moneda muchas veces, la proporción de veces que obtenemos cara tiende a $1/2$ a medida que aumenta el número de lanzamientos.

Por su parte, la interpretación del grado de creencia es que $\mathbb{P}(A)$ mide la fuerza de creencia de un observador en que A es verdadero. En cualquier interpretación, requerimos que se cumplan los axiomas 1 a 3. La diferencia en la interpretación no importará mucho hasta que nos ocupemos de la inferencia estadística en el curso de Estadística II. Allí, las diferentes interpretaciones conducen a dos escuelas de inferencia: la frecuentista y la bayesiana.

2.2.1. Propiedades de las Probabilidades

Derivado de los axiomas, podemos listar una serie de propiedades:

1. $\mathbb{P}(\emptyset) = 0$

2. Si $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
3. $0 \leq \mathbb{P}(A) \leq 1$
4. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
5. Si $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

La demostración de cada una sería (cambiamos el orden para mantener consistencia lógica de los resultados):

1. Si $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Por el Axioma 3 resulta obvia la demostración.

2. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Sabemos por el axioma 2 que $\mathbb{P}(\Omega) = 1 = \mathbb{P}(A \cup A^c)$.

Debe ser claro que A y A^c son disjuntos o mutuamente excluyentes, así $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) = 1$

De donde podemos despejar para tener: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

3. $\mathbb{P}(\emptyset) = 0$.

Partamos de que $\emptyset = \Omega^C$, sabemos que $\mathbb{P}(\emptyset) = \mathbb{P}(\Omega^C) = 1 - \mathbb{P}(\Omega) = 0$

4. Si $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.

Notemos que podemos escribir a B como $B = A \cup (B \cap A^C)$. De donde podemos observar que A y $B \cap A^C$ son eventos disjuntos o mutuamente excluyentes. Entonces

$$\mathbb{P}(B) = \mathbb{P}(A \cup (B \cap A^C)) = \mathbb{P}(A) + \mathbb{P}(B \cap A^C)$$

Lo que implica claramente que $\mathbb{P}(B) \geq \mathbb{P}(A)$ a partir de que asumimos que $\mathbb{P}(B \cap A^C) \geq 0$.

5. $0 \leq \mathbb{P}(A) \leq 1$.

La demotración es inmediata de lo anterior, ya que $A \subset \Omega$ y que $\mathbb{P}(\Omega) = 1$.

Una propiedad menos obvia es la siguiente:

Lema. Para cualesquiera pares de eventos A y B , $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Demostración. Partamos de que $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$. Es decir, particionamos la unión de dos eventos A y B en subconjuntos que están disjuntos. Así, podemos usar algunas de las propiedades de \mathbb{P} para obtener:

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}((A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)) \\ &= \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) \\ &= \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) + \mathbb{P}(A \cap B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}((A \cap B^c) \cup (A \cap B)) + \mathbb{P}((A^c \cap B) \cup (A \cap B)) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \end{aligned}$$

Ejemplo. Muestre que $\mathbb{P}(A \cap B) \geq 1 - \mathbb{P}(A^c) + \mathbb{P}(B^c)$ para cualquier par de eventos A y B en Ω .

Partamos de que con el Lema anterior demostramos que $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, del cual podemos despejar para determinar que:

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

Adicionalmente, usando $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ y que, en consecuencia $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$. De esta forma podemos establecer que:

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) \\ &= 1 - \mathbb{P}(A^c) + 1 - \mathbb{P}(B^c) - \mathbb{P}(A \cup B) \\ &= 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c) + 1 - \mathbb{P}(A \cup B) \\ &\geq 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c) \end{aligned}$$

Notemos que hemos usado la condición de que para todo evento A y B en Ω , $1 \geq \mathbb{P}(A \cup B)$.

Ejemplo. Asumamos que lanzamos una moneda dos veces. Sea A_1 el evento de que caiga águila en el lanzamiento 1 y A_2 el evento de que resulte águila en el lanzamiento 2. Si ambas caras de las monedas son igualmente probables, entonces $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$.

Ejemplo. Se extraen dos cartas de una baraja de póker sin reemplazo. ¿Cuál es la probabilidad de que la segunda sea de mayor rango que la primera?

Para plantear la solución definamos a A_1 , A_2 y A_3 como los eventos “La primera carta sea de rango menor”, “La segunda carta es de rango mayor” y “Ambas cartas son del mismo rango”, respectivamente. Claramente, los tres eventos A_i 's son mutuamente excluyentes y consideran todos los posibles casos. De esta forma:

$$\mathbb{P}(A_1 \cup A_2 \cup A_3) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) = \mathbb{P}(\Omega) = 1$$

Una vez que la primera carta ha sido tomada, existen 3 posibles casos en los cuales la segunda puede ser del mismo rango, esto es:

$$\mathbb{P}(A_3) = \frac{3}{51}$$

También notemos que existe simetría entre los eventos A_1 y A_2 , de forma que: $\mathbb{P}(A_1) = \mathbb{P}(A_2)$. Así:

$$2\mathbb{P}(A_2) + \frac{3}{51} = 1$$

Es decir,

$$\mathbb{P}(A_2) = \frac{8}{17}$$

2.2.2. Probabilidad en espacios muestrales finitos

Supongamos un espacio muestral finito dado por $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$. Por ejemplo, si lanzamos un dado dos veces, entonces Ω tendrá 36 elementos: $\Omega = \{(i, j); i, j \in \{1, 2, \dots, 6\}\}$.

Si cada posible resultado es igualmente probable, entonces $\mathbb{P}(A) = |A|/36$, donde $|A|$ denota el número de elementos en A . Así, la probabilidad de que la suma de los números sea 11 es $2/36$, dado que sólo hay dos posibles resultados que corresponden a este evento.

Si Ω es finito y si cada resultado es igualmente probable, entonces:

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

La cual es conocida como la distribución de probabilidad uniforme. Para calcular probabilidades, necesitamos contar el número de elementos en un evento A . Los métodos para contar elementos son conocidos como métodos de combinatoria.

Conjunto Potencia

El conjunto potencia de Ω , denotado por 2^Ω , es aquel conjunto cuyos elementos son todos los subconjuntos posibles de Ω . En términos estrictos, esta nueva colección deja de ser un conjunto y se le llama clase de subconjuntos de Ω .

Ejemplo Si $\Omega = \{a, b, c\}$, entonces el conjunto 2^Ω consta de 8 elementos:

$$2^\Omega = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

Observe que los elementos del conjunto potencia son en sí mismos conjuntos, y que en esta colección están contenidos todos los eventos que podrían ser de interés en un experimento aleatorio. Al respecto, no es difícil demostrar que:

$$|2^\Omega| = 2^{|\Omega|}$$

De esta situación proviene la notación usada para el conjunto potencia: 2^Ω , ya que la expresión por sí misma no tiene sentido matemático y, por lo tanto, debe considerarse como un símbolo para denotar el conjunto potencia.

Análisis combinatorio

Principio de multiplicación. Si un procedimiento A_1 puede efectuarse de n formas distintas y un segundo procedimiento A_2 puede realizarse de m formas diferentes, entonces el total de formas en que puede efectuarse el primer procedimiento seguido del segundo es el producto $n \times m$. Es decir, $|A_1 \times A_2| = |A_1| \times |A_2|$, donde \times denota el producto cartesiano en el primer caso.

Ejemplo. Suponga que un cierto experimento aleatorio consiste en lanzar un dado y después seleccionar al azar una letra del alfabeto. ¿Cuál es la cardinalidad del correspondiente espacio muestral?

El experimento de lanzar un dado tiene 6 resultados posibles y consideremos que tenemos un alfabeto de 26 letras. El correspondiente espacio muestral tiene entonces cardinalidad $6 \times 26 = 156$.

El principio de multiplicación es válido no solamente para dos procedimientos sino que también vale para cualquier sucesión finita de procedimientos. Por ejemplo, si A_1, A_2, \dots, A_k denotan k procedimientos sucesivos, entonces este principio se puede enunciar en símbolos de la forma siguiente: $|A_1 \times A_2 \times \dots \times A_k| = |A_1| \times |A_2| \times \dots \times |A_k|$.

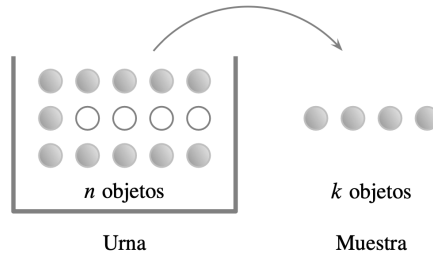


Figura 2.5: Diagrama que ilustra el proceso de extraer al azar, retomado de Rincón (2012, p17) Rincón 2012

Ejemplo. Suponga que una persona tiene 4 pantalones distintos, 6 camisas, y 2 pares de zapatos. ¿De cuántas formas distintas puede esta persona vestirse con estas prendas?

Solución: $4 \times 6 \times 2 = 48$. Es decir, se puede vestir de manera distinta durante 48 días sin repetir una combinación de prendas.

Vamos a considerar a continuación diferentes esquemas y contextos en donde es posible encontrar una fórmula matemática para ciertos problemas de conteo. En todos ellos aplicaremos el principio de multiplicación.

El esquema general es el de extraer al azar k objetos, uno a la vez, de una urna con n objetos distintos. Esto se muestra en la Figura 2.5.

Ordenaciones con repetición: muestras con orden y con reemplazo.

Suponga que tenemos una urna con n objetos distintos. Deseamos realizar k extracciones al azar de un objeto a la vez. Al efectuar una extracción, registramos el objeto escogido y lo regresamos a la urna, de esta forma el mismo objeto puede ser extraído varias veces (*muestreo con reemplazo*).

De esta forma, el total de arreglos que se pueden obtener de esta urna al hacer k extracciones es el número n^k , pues en cada extracción tenemos n objetos posibles para escoger y efectuamos k extracciones.

Esta fórmula es consecuencia del principio de multiplicación enunciado antes. A este número se le llama **ordenaciones con repetición**. Se dice que *la muestra es con orden pues es importante el orden en el que se van obteniendo los objetos, y es con reemplazo pues cada objeto seleccionado se reincorpora a la urna*.

Ejemplo. Suponga que tenemos un conjunto de 60 caracteres diferentes. Este conjunto contiene todas las letras minúsculas del alfabeto, las letras

mayúsculas, los diez dígitos y algunos caracteres especiales. ¿Cuántas passwords o palabras clave de longitud 4 se pueden construir usando este conjunto de 60 caracteres?

Este es un ejemplo de una ordenación de 60 caracteres en donde, dada la información, se permiten las repeticiones. Como cada carácter de los 60 disponibles puede ser escogido para ser colocado en cada una de las cuatro posiciones de la palabra clave, entonces se pueden construir $60 \times 60 \times 60 \times 60 = 60^4 = 12,960,000$ distintos passwords de longitud 4.

Ordenaciones sin repetición: muestras con orden y sin reemplazo

Suponga que se tiene la misma situación que antes, una urna con n objetos y de los cuales se deben extraer, uno a uno, k objetos. Suponga esta vez que el *muestreo es sin reemplazo*, es decir, una vez seleccionado un objeto, éste ya no se reincorpora a la urna.

El total de arreglos distintos que se pueden obtener de este modo es el número: $n(n-1)(n-2) \cdots (n-k+1)$.

Primeramente debemos observar que hay k factores en la expresión anterior. El primer factor es n y ello es debido a que tenemos cualesquiera de los n objetos para ser colocados en primera posición, para la segunda posición tenemos ahora $n-1$ objetos, para la tercera $n-2$ objetos, y así sucesivamente. Este razonamiento termina al escoger el k -ésimo objeto para el cual tenemos únicamente $n-k+1$ posibilidades.

Nuevamente por el principio de multiplicación, la respuesta es el producto indicado. La expresión encontrada puede escribirse como sigue:

$$P(n, k) = \frac{n!}{(n-k)!} \quad (2.1)$$

A esta expresión se le llama **permutaciones** de n en k .

En el caso particular cuando la muestra es exhaustiva, es decir, cuando $k = n$, o bien cuando todos los objetos son extraídos uno por uno, entonces se tienen las permutaciones o distintos órdenes en que se pueden colocar n objetos.

Ejemplo. ¿De cuántas formas distintas pueden asignarse los premios primero, segundo y tercero en una rifa de 10 boletos numerados del 1 al 10?

Claramente se trata de una ordenación sin repetición de 10 objetos en donde se deben extraer 3 de ellos. La respuesta es entonces que existen $10 \times 9 \times 8 = 720$ distintas asignaciones para los tres primeros lugares en la rifa.

Permutaciones: muestras exhaustivas con orden y sin reemplazo

La pregunta básica acerca del total de formas en que podemos poner en orden lineal (uno detrás de otro y, por lo tanto, no hay repetición) n objetos distintos tiene como respuesta el factorial de n , denotado por $n!$ y definido como sigue:

$$n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1$$

A este número también se le conoce como las permutaciones de n objetos, y se usa la notación $P(n) = n!$.

Adicionalmente y por conveniencia se define $0! = 1$.

Observe que las permutaciones de n objetos son un caso particular de la situación mencionada en la sección anterior sobre ordenaciones sin repetición cuando la muestra es exhaustiva, es decir, cuando se extraen uno a uno todos los objetos de la urna.

Ejemplo. Si deseamos conocer el total de formas distintas en que podemos colocar una enciclopedia de 5 volúmenes en un librero, la respuesta es claramente $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$.

El razonamiento es el siguiente: Cualquiera de los cinco libros puede ser colocado al principio, quedan cuatro libros por colocar en la segunda posición, restan entonces tres posibilidades para la tercera posición, etc. Por el principio de multiplicación, la respuesta es el producto de estos números.

En suma, dados n objetos, el número de formas en que podemos ordenar esos objetos es: $n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$. Por convención, definimos que $0! = 1$.

Combinaciones: muestras sin orden y sin reemplazo

Supongamos nuevamente que tenemos un conjunto de n objetos distinguibles y nos interesa obtener una muestra de tamaño k . Supongamos ahora que las muestras deben ser sin orden y sin reemplazo. Es decir, en la muestra no debe haber elementos repetidos, pues no hay reemplazo, y además la muestra debe verse como un conjunto, pues no debe haber orden entre sus elementos. ¿Cuántas diferentes muestras podemos obtener de estas características? Para responder a esta pregunta seguiremos el siguiente razonamiento: cuando el orden importa hemos encontrado antes la fórmula de permutaciones:

$$P(n, k) = \frac{n!}{(n-k)!}$$

Ahora que no nos interesa el orden en el que aparecen las extracciones, observamos que cada uno de los arreglos de la fórmula anterior está siendo

contado $k!$ veces, las veces en que los mismos k elementos pueden ser permutados unos con otros, siendo que el conjunto de elementos es el mismo. Para obtener arreglos en donde el orden no importa, debemos entonces dividir por $k!$. La fórmula a la que hemos llegado se llama combinaciones de n en k y la denotaremos como sigue:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (2.2)$$

La expresión en la ecuación (2.2) que se lee como “ n elegidos en grupos de k ”, el cual es el número de las distintas formas en que se puede elegir k objetos de un total de n . Finalmente, podemos establecer las siguientes propiedades:

$$\binom{n}{0} = \binom{n}{n} = 1 \text{ y } \binom{n}{k} = \binom{n}{n-k}$$

Ejemplo. Si asumimos un total de 20 personas en un salón de clases y queremos seleccionar un comité de 3 estudiantes, entonces existen:

$$\binom{20}{3} = \frac{20 \times 19 \times 18}{3 \times 2 \times 1} = 1140$$

Ejemplo. ¿Cuántos equipos distintos de tres personas pueden escogerse de un grupo de 5 personas? Observe que el orden de las tres personas escogidas no es importante, de modo que la respuesta es:

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = 10$$

A este número también se le conoce con el nombre de coeficiente binomial de n en k , pues aparece en el famoso teorema del binomio: para cualesquiera números reales a y b , y para cualquier número entero $n \geq 0$,

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k \quad (2.3)$$

Para los casos $n = 2$ y $n = 3$ el teorema del binomio se reduce a las siguientes fórmulas que muy seguramente el lector conoce:

$$\begin{aligned} (a+b)^2 &= a^2 + 2ab + b^2 \\ (a+b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3 \end{aligned}$$

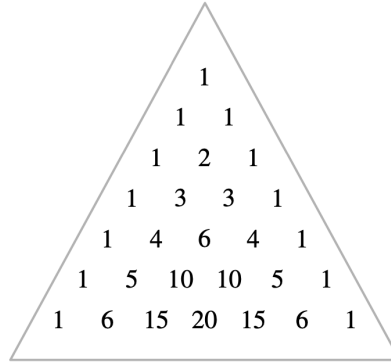


Figura 2.6: Diagrama que ilustra el Triángulo de Pascal, retomado de Rincón (2012, p20) Rincón 2012

El coeficiente binomial es también una forma de generar las entradas del así llamado triángulo de Pascal, que puede observarse en la Figura 2.6.

El n -ésimo renglón del triángulo de Pascal, iniciando desde cero, contiene los coeficientes del desarrollo de $(a + n)^n$. Existe una forma sencilla de construir este triángulo observando que cada uno de estos números, exceptuando los extremos, es la suma de los dos números inmediatos del renglón anterior.

Coeficiente multinomial

Ahora consideremos que tenemos n objetos no necesariamente distintos unos de otros, por ejemplo, supongamos que tenemos k_1 objetos de un primer tipo, k_2 objetos de un segundo tipo, y así sucesivamente, hasta k_m objetos del tipo m , en donde $k_1 + k_2 + \dots + k_m = n$. Estos n objetos pueden todos ordenarse uno detrás de otro de tantas formas distintas como indica el así llamado coeficiente multinomial:

$$\binom{n}{k_1 \ k_2 \ \dots \ k_{m-1} \ k_m} = \frac{n!}{k_1! k_2! \dots k_{m-1}! k_m!} \quad (2.4)$$

Un razonamiento para obtener la fórmula de la ecuación (2.4) es el siguiente: si consideramos que los n objetos son todos distintos, entonces claramente las distintas formas en que pueden escribirse todos estos objetos uno detrás de otro son $n!$. Pero para cada uno de estos arreglos, los k_1 objetos del primer tipo, supuestos inicialmente distintos cuando en realidad no lo son, pueden permutarse entre sí de $k_1!$ formas diferentes, siendo que el arreglo total es el mismo. De aquí que debemos dividir por $k_1!$. Lo mismo sucede con los

Muestras	con reemplazo	sin reemplazo
con orden	n^k	$\frac{n!}{(n-k)!}$
sin orden	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Figura 2.7: Resumen de fórmulas de conteo, retomado de Rincón (2012, p22) Rincón 2012

elementos del segundo tipo y así sucesivamente hasta los elementos del tipo m .

El coeficiente multinomial aparece en la siguiente fórmula:

$$(a_1 + a_2 + \cdots + a_m)^n = \sum \binom{n}{k_1 \ k_2 \ \cdots \ k_{m-1} \ k_m} a_1^{k_1} a_2^{k_2} \cdots a_m^{k_m} \quad (2.5)$$

En donde la suma se efectúa sobre todos los posibles valores enteros no negativos de k_1, k_2, \dots, k_m , tales que $k_1 + k_2 + \cdots + k_m = n$. A este resultado se le conoce como el teorema multinomial y es claramente una extensión del teorema del binomio.

Por ejemplo, compruebe el lector que la fórmula en la ecuación (2.5) produce la siguiente expresión:

$$(a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc$$

¿Puede usted desarrollar $(a + b + c)^3$? Es interesante observar que cuando hay únicamente dos tipos de objetos, el coeficiente multinomial se reduce al coeficiente binomial.

Resumen de fórmulas

La Figura 2.7 muestra el resumen de lo que hemos discutido en esta sección.

2.3. Independencia de eventos

Retomemos el ejemplo del lanzamiento de una moneda dos veces. Supongamos que queremos determinar la probabilidad de que ocurra que tanto

en el primer lanzamiento como en el segundo obtengamos un águila. Esta probabilidad sería $1/2 \times 1/2$. Multiplicamos las probabilidades dado que reconocemos que los dos lanzamientos son independientes. La definición más formal de esta situación sería la siguiente.

Definición 2.15 Independencia. Dos eventos A y B son *independientes* si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$$

Un conjunto de eventos $\{A_i : i \in I\}$ es independiente si

$$\mathbb{P}(\cap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i)$$

Para cada subconjunto finito J de I .

La independencia puede aparecer en dos vías distintas. Algunas veces, explícitamente **asumiremos** la independencia. Por ejemplo, en dos lanzamientos de una moneda solemos asumir que los lanzamientos son independientes, lo cual refleja el hecho de que la moneda no tiene memoria del primer lanzamiento.

En otros casos, **derivaremos** la independencia mediante la verificación de que $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Ejemplo. En el lanzamiento de un dado, sea $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$. Entonces, podemos derivar que $A \cap B = \{2, 4\}$ y que $\mathbb{P}(A \cap B) = 2/6 = \mathbb{P}(A)\mathbb{P}(B) = (1/2) \times (2/3)$. De esta forma, demostramos que A y B son independientes. Notemos que en este caso no asumimos que A y B eran independientes, esto lo notamos hasta que evaluamos que lo fueran.

Supongamos que A y B son eventos disjuntos, cada uno con una probabilidad positiva. Entonces, ¿pueden ser independientes? No. Esto es fácil de verificar ya que $\mathbb{P}(A)\mathbb{P}(B) > 0$, pero $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$. Con excepciones particulares, no hay forma de juzgar independencia mediante la visualización de diagramas de Venn.

Ejemplo. Sea A el evento de sacar un rey de una baraja de póquer estándar y B , el evento de sacar un diamante. En este caso los eventos A y B son independientes, ya que la probabilidad de su intersección (tomar un rey de diamantes) es igual a:

$$\mathbb{P}(A \cap B) = \frac{1}{52} = \frac{1}{4} \cdot \frac{1}{13} = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Ejemplo. Supongamos que una moneda es lanzada 10 veces. Sea el evento A = obtenemos por lo menos una aguilá. Sea T_j el resultado del lanzamiento j -ésimo en el que el resultado de lanzamiento de la moneda es un sol (cara). Entonces, dada la notación planteada y considerando la independencia en el resultado del lanzamiento de la moneda podemos obtener:

$$\begin{aligned}
 \mathbb{P}(A) &= 1 - \mathbb{P}(A^c) \\
 &= 1 - \mathbb{P}(\text{todos los lanzamientos resultaron un sol}) \\
 &= 1 - \mathbb{P}(T_1 \cap T_2 \cap \dots \cap T_{10}) \\
 &= 1 - \mathbb{P}(T_1)\mathbb{P}(T_2) \dots \mathbb{P}(T_{10}) \text{ usando la independencia de eventos} \\
 &= 1 - \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) \dots \left(\frac{1}{2}\right) \\
 &= 1 - \left(\frac{1}{2}\right)^{10} \\
 &= 0.999
 \end{aligned}$$

Ejemplo. Dos personas toman turnos para intentar encestar un balón de básquetbol. La persona 1 tiene éxito con una probabilidad de $1/3$ mientras que la persona 2 tiene éxito con una probabilidad de $1/4$. ¿Cuál es la probabilidad de que la persona 1 tenga éxito antes que la persona 2?

Sea E = que la persona 1 tenga éxito antes que la persona 2. Sea A_j el evento de que el primer éxito sea de la persona 1 y que ocurra en el ensayo número j . Notemos que A_1, A_2, \dots son eventos disjuntos y que por lo tanto:

$$E = \bigcup_{j=1}^{\infty} A_j$$

Por eso:

$$\mathbb{P}(E) = \sum_{j=1}^{\infty} \mathbb{P}(A_j)$$

Ahora analicemos cada evento A_j . Si $j = 1$, entonces $\mathbb{P}(A_1) = 1/3$. Partamos de entender que si A_2 ocurre, quiere decir que en la secuencia la persona 1 y la persona 2 perdieron, así: $\mathbb{P}(A_2) = (2/3)(3/4)(1/3) = (1/2)(1/3)$.

Siguiendo esta idea, podemos encontrar que en el caso de j -ésimo caso tendríamos:

$$\mathbb{P}(A_j) = \left(\frac{1}{2}\right)^{j-1} \left(\frac{1}{3}\right)$$

De esta manera:

$$\begin{aligned}
\mathbb{P}(E) &= \sum_{j=1}^{\infty} \mathbb{P}(A_j) \\
&= \sum_{j=1}^{\infty} \left(\frac{1}{2}\right)^{j-1} \left(\frac{1}{3}\right) \\
&= \left(\frac{1}{3}\right) \sum_{j=1}^{\infty} \left(\frac{1}{2}\right)^{j-1} \\
&= \left(\frac{1}{3}\right) \lim_{j \rightarrow \infty} \frac{1 - \frac{1}{2}^{j+1}}{1 - \frac{1}{2}} \\
&= \left(\frac{2}{3}\right)
\end{aligned}$$

Para llegar a este resultado, retomemos cómo resolver la suma de los términos de a^i , donde $|a| < 1$, de tal forma que buscaremos dar una expresión más compresible a dicho término. Definamos la siguiente expresión de la suma de los primeros t términos como:

$$S_{t-1} = 1 + a + a^2 + \dots + a^{t-1} = \sum_{i=0}^{t-1} a^i \quad (2.6)$$

Por lo tanto, S_t estaría dado por la siguiente expresión de la suma de los primeros t términos:

$$\begin{aligned}
S_t &= a \sum_{i=0}^{t-1} a^i \\
&= a(1 + a + a^2 + \dots + a^{t-1}) \\
&= a + a^2 + a^3 + \dots + a^t \\
&= a \cdot S_{t-1}
\end{aligned} \quad (2.7)$$

Tomando los dos resultados de las ecuaciones (2.6) y (2.7) anteriores, podemos expresar que si a S_{t-1} le restamos S_t , y desarrollando ambos lados de la ecuación anterior podemos obtener:

$$\begin{aligned}
S_{t-1} - a \cdot S_{t-1} &= S_{t-1} - S_t \\
(1 - a)S_{t-1} &= (1 + a + a^2 + \dots + a^{t-1}) - (a + a^2 + a^3 + \dots + a^t) \\
(1 - a)S_{t-1} &= 1 - a^t
\end{aligned}$$

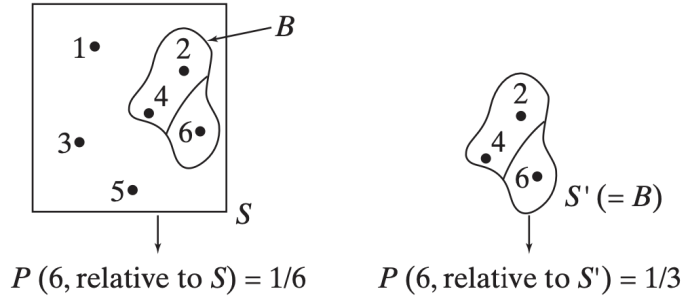


Figura 2.8: Ejemplo de un evento restringido, retomado de Larsen (2012, p32) Larsen y Marx 2012

Así, podemos concluir que:

$$S_{t-1} = \frac{1 - a^t}{1 - a} \quad (2.8)$$

Y que:

$$S_t = \frac{1 - a^{t+1}}{1 - a} \quad (2.9)$$

En resumen en esta sección hemos analizado que:

1. A y B son eventos independientes si y sólo si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
2. A veces se supone la independencia y otras se deriva.
3. Los eventos disjuntos con probabilidad positiva no son independientes.

2.4. Probabilidad condicional

La probabilidad condicional es una situación en la que tenemos conocimiento de que, por ejemplo, el evento B ha ocurrido, por lo que ahora el espacio muestral Ω ahora se ha reducido. Una manera de ilustrarlo es, por ejemplo, el caso del lanzamiento de un dado en el cual $\Omega = 1, 2, 3, 4, 5, 6$, pero $B = 2, 4, 6$, el evento de que el resultado del lanzamiento sea par. La Figura 2.8 ilustra esta situación y el cambio en la probabilidad asociada a un evento A definido como que el resultado del lanzamiento sea 6.

Asumamos que $\mathbb{P}(B) > 0$, definimos la probabilidad condicional del evento A dado el evento B como sigue:

Definición 2.16 Probabilidad Condicional. Si $\mathbb{P}(B) > 0$, entonces la **probabilidad condicional** de A dado B es:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Pensemos a $\mathbb{P}(A|B)$ como la fracción de veces que A ocurre entre aquellas en las que B ocurre. Para cualquier B tal que $\mathbb{P}(B) > 0$, $\mathbb{P}(\cdot|B)$ es una función de probabilidad. Es decir, satisface los tres axiomas de probabilidad discutidos previamente.

En particular, $\mathbb{P}(A|B) \geq 0$, $\mathbb{P}(\Omega|B) = 1$ y si A_1, A_2, \dots son una serie de eventos disjuntos, entonces:

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$$

Pero, en general, no es cierto que:

$$\mathbb{P}(A|B \cup C) = \mathbb{P}(A|B) + \mathbb{P}(A|C)$$

Las reglas de la probabilidad aplican a casos del lado izquierdo de la barra. En general, tampoco es cierto que: $\mathbb{P}(A|B) = \mathbb{P}(B|A)$. Por ejemplo, la probabilidad de que a una persona le salgan manchas si tiene sarampión es 1, pero la probabilidad de que tenga sarampión si tiene manchas no es 1. En este caso, la diferencia entre $\mathbb{P}(A|B)$ e $\mathbb{P}(B|A)$ es obvia, pero hay casos en los que es menos obvio. Según Wasserman (2004) Wasserman 2004, este error se comete con tanta frecuencia en casos legales que a veces se le llama falacia del fiscal.

Ejemplo. Se saca una carta de una baraja de póquer. ¿Cuál es la probabilidad de que la carta sea un trébol, dado que la carta es un rey?

Intuitivamente, la respuesta es $1/4$, ya que el rey tiene la misma probabilidad de ser un corazón, un diamante, un trébol o una pica. Más formalmente, sea C el evento “La carta es un trébol”; sea K el evento “La carta es un rey”. Por la definición de probabilidad condicional:

$$\mathbb{P}(C|K) = \frac{\mathbb{P}(C \cap K)}{\mathbb{P}(K)}$$

Pero $\mathbb{P}(K) = 4/52$ y $\mathbb{P}(C \cap K) = \mathbb{P}(\text{La carta es un rey de tréboles}) = 1/52$. Por lo tanto, confirmando nuestra intuición:

$$\mathbb{P}(C|K) = \frac{1/52}{4/52} = \frac{1}{4}$$

Ejemplo. Nuestras intuiciones pueden ser engañadas a menudo por problemas de probabilidad, incluso aquellos que parecen ser simples y directos. El problema de los “dos niños” descrito aquí es un caso que se cita a menudo.

Consideremos el conjunto de familias que tienen dos hijos. Supongamos que las cuatro posibles secuencias de nacimiento son: (el niño menor es un niño, el niño mayor es un niño), (el niño menor es un niño, el niño mayor es una niña), etcétera, son igualmente probables.

¿Cuál es la probabilidad de que ambos hijos sean niños dado que al menos uno es un niño?

La respuesta no es $\frac{1}{2}$.

La respuesta correcta se puede deducir de la definición de probabilidad condicional. Por suposición, cada una de las cuatro posibles secuencias de nacimiento (b, b), (b, g), (g, b) y (g, g) tiene una probabilidad de ocurrencia de $1/4$.

Sea A el evento de que ambos niños sean niños y sea B el evento de que al menos un niño sea un niño. Entonces:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$$

Lo anterior, dado que A es un subconjunto de B (por lo que la superposición entre A y B es solo A). Pero A tiene un resultado $\{(b, b)\}$ y B tiene tres resultados $\{(b, g), (g, b), (b, b)\}$.

Al aplicar la definición, se obtiene:

$$\mathbb{P}(A|B) = \frac{(1/4)}{(3/4)} = \frac{1}{3}$$

Otro enfoque correcto es volver al espacio muestral y deducir el valor de $P(A|B)$ a partir de los primeros principios.

Ejemplo. Una prueba médica para una enfermedad D tiene resultados *positivo* y *negativo*. Las probabilidades de estos son:

	D	D^c
<i>positivo</i>	0.009	0.099
<i>negativo</i>	0.001	0.891

A partir de la definición de probabilidad condicional, tenemos:

$$\begin{aligned}\mathbb{P}(\textit{positivo}|D) &= \frac{\mathbb{P}(\textit{positivo} \cap D)}{\mathbb{P}(D)} = \frac{0.009}{0.009 + 0.001} = 0.9 \\ \mathbb{P}(\textit{negativo}|D^c) &= \frac{\mathbb{P}(\textit{negativo} \cap D^c)}{\mathbb{P}(D^c)} = \frac{0.891}{0.891 + 0.099} = 0.9\end{aligned}$$

Aparentemente, la prueba es bastante precisa. Las personas enfermas dan positivo el 90 % de las veces y las personas sanas dan negativo aproximadamente el 90 % de las veces. Supongamos que se hace una prueba y da positivo. ¿Cuál es la probabilidad de que tenga la enfermedad? La mayoría de las personas responde 0.90. La respuesta correcta es:

$$\mathbb{P}(D|\text{positivo}) = \frac{\mathbb{P}(\text{positivo} \cap D)}{\mathbb{P}(\text{positivo})} = \frac{0.009}{0.009 + 0.099} = 0.8$$

La lección que se desprende de esto es que hay que calcular la respuesta numéricamente. No hay que confiar en la intuición.

Ejemplo. Saque dos cartas de una baraja, sin reemplazo. Sea A el evento de que la primera carta que saque sea el As de tréboles y sea B el evento de que la segunda carta que saque sea la reina de diamantes. Entonces:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) = (1/52) \times (1/51)$$

Ejemplo. Retomemos el ejemplo inicial. Supongamos que extraemos una carta de un mazo de póker. ¿Cuál es la probabilidad de que la carta sea un trébol, dado que la carta es un rey?

Intuitivamente, la respuesta es $1/4$. El rey tiene la misma probabilidad de ser un corazón, un diamante, un trébol o un diamante. Más formalmente, sea A el evento “La carta es un trébol”; sea B el evento “La carta es un rey”. Según la definición de probabilidad condicional:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\text{La carta es un rey de trébol})}{\mathbb{P}(\text{La carta es un rey})} = \frac{1/52}{4/52} = \frac{1}{4}$$

Observe que en este ejemplo la probabilidad condicional $\mathbb{P}(A|B)$ es numéricamente la misma que la probabilidad incondicional $\mathbb{P}(B)$. Ambas iguales a $1/4$. Esto significa que nuestro conocimiento de que B ha ocurrido no nos da información adicional sobre las probabilidades de que A ocurra. **Se dice que dos eventos que tienen esta propiedad son independientes.**

2.5. Teorema de Bayes

El resultado interesante que estudiaremos a continuación involucra nuevamente probabilidades condicionales. Fue publicado por primera vez en 1763,

dos años después de la muerte de su creador: el matemático y teólogo inglés Thomas Bayes.

El teorema de Bayes es la base de los “sistemas expertos” y las “redes de Bayes”, que trataremos de analizar más adelante. Primero, necesitamos un resultado preliminar.

Teorema 2.1 *La Ley de la Probabilidad Total.* Sea A_1, A_2, \dots, A_k una partición de Ω . Entonces, para cualquier evento B

$$\mathbb{P}(B) = \sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i) \quad (2.10)$$

Demostración. Definamos a $C_j = B \cap A_j$, podemos notar que la secuencia de C_1, C_2, \dots, C_k son disjuntos y que por lo tanto:

$$B = \cup_{j=1}^k C_j$$

Entonces,

$$\mathbb{P}(B) = \sum_j \mathbb{P}(C_j) = \sum_j \mathbb{P}(B \cap A_j) = \sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)$$

Lo anterior, dada que de la definición de probabilidad condicional: $\mathbb{P}(B \cap A_j) = \mathbb{P}(B|A_j)\mathbb{P}(A_j)$.

Teorema 2.2 *Teorema de Bayes.* Sea A_1, \dots, A_k una partición de Ω tal que $\mathbb{P}(A_i) > 0$ para cada i . Si $\mathbb{P}(B) > 0$, entonces, para cada $i = 1, \dots, k$,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)} \quad (2.11)$$

Observación. Usualmente llamaremos a $\mathbb{P}(A_i)$ como la probabilidad *prior* de A y a $\mathbb{P}(A_i|B)$ la probabilidad *posterior*.

Demostración. Aplicamos la definición de probabilidad condicional dos veces, seguida de la ley de probabilidad total:

$$\begin{aligned} \mathbb{P}(A_i|B) &= \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)} \end{aligned}$$

Ejemplo. Asumamos que dividimos nuestro correo electrónico en tres categorías: $A_1 = \text{“spam”}$, $A_2 = \text{“baja prioridad”}$ y $A_3 = \text{“alta prioridad”}$. De nuestra experiencia previa hemos descubierto que:

$$\begin{aligned}\mathbb{P}(A_1) &= 0.7 \\ \mathbb{P}(A_2) &= 0.2 \\ \mathbb{P}(A_3) &= 0.1\end{aligned}$$

Por supuesto, $\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) = 1$.

Ahora, sea B el evento de que el correo electrónico contenga la palabra “gratis”. Según nuestra experiencia previa:

$$\begin{aligned}\mathbb{P}(B|A_1) &= 0.9 \\ \mathbb{P}(B|A_2) &= 0.01 \\ \mathbb{P}(B|A_3) &= 0.01\end{aligned}$$

Notemos que: $\mathbb{P}(B|A_1) + \mathbb{P}(B|A_2) + \mathbb{P}(B|A_3) \neq 1$.

Si recibimos un correo electrónico con la palabra “gratis”. ¿Cuál es la probabilidad de que sea spam? El teorema de Bayes de la ecuación 2.11 arroja:

$$\mathbb{P}(A_1|B) = \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = 0.995 \quad (2.12)$$

Ejemplo. En una fábrica hay dos máquinas. La máquina 1 realiza el 60 % de la producción total y la máquina 2 el 40 %. De su producción total, la máquina 1 produce 3 % de material defectuoso, la 2 el 5 %. El asunto es que se ha encontrado un material defectuoso, ¿cuál es la probabilidad de que este material defectuoso provenga de la máquina 2?

Sea M_1 el evento “La máquina 1 produjo el material escogido”, M_2 en evento “La máquina 2 produjo el material escogido” y finalmente sea D el evento “El material escogido es defectuoso”. El problema es encontrar $\mathbb{P}(M_2|D)$ y observamos que la información que tenemos es $\mathbb{P}(D|M_2)$. Por el teorema de Bayes tenemos que:

$$\begin{aligned}\mathbb{P}(M_2|D) &= \frac{\mathbb{P}(D|M_2)\mathbb{P}(M_2)}{\mathbb{P}(D|M_1)\mathbb{P}(M_1) + \mathbb{P}(D|M_2)\mathbb{P}(M_2)} \\ &= \frac{\frac{5}{100} \times \frac{40}{100}}{\frac{3}{100} \times \frac{60}{100} + \frac{5}{100} \times \frac{40}{100}} \\ &= \frac{10}{19}\end{aligned}$$

Como un ejercicio al lector se deja comprobar que $\mathbb{P}(M_1|D) = 9/19$.

Ejemplo. Durante un apagón, cien personas son detenidas bajo sospecha de saqueo. A cada una de ellas se le aplica una prueba de polígrafo. Por experiencia se sabe que la fiabilidad del polígrafo es del 90 % cuando se aplica a un sospechoso culpable y del 98 % cuando se aplica a alguien inocente. Supongamos que de las cien personas detenidas, sólo doce (12) estuvieron realmente implicadas en algún delito. ¿Cuál es la probabilidad de que un sospechoso determinado sea inocente si el polígrafo dice que es culpable?

Sea B el evento “El polígrafo dice que el sospechoso es culpable” sean A_1 y A_2 los eventos “El sospechoso es culpable” “El sospechoso no es culpable”, respectivamente.

Decir que el polígrafo es “90 % confiable cuando se aplica a un sospechoso culpable” significa que $\mathbb{P}(B|A_1) = 0.90$. De manera similar, la confiabilidad del 98 % para sospechosos inocentes implica que $\mathbb{P}(B^C|A_2) = 0.98$ o, equivalentemente, $\mathbb{P}(B|A_2) = 0.02$.

Sabemos que $\mathbb{P}(A_1) = 12/100$ y $\mathbb{P}(A_2) = 88/100$. Sabiendo eso, podemos sustituir en la Ecuación (2.11).

De esta forma podemos concluir que la probabilidad de que un sospechoso sea inocente dado que el polígrafo dice que es culpable es de 0.14:

$$\begin{aligned}\mathbb{P}(A_2|B) &= \frac{\mathbb{P}(B|A_2)\mathbb{P}(A_2)}{\mathbb{P}(B|A_1)\mathbb{P}(A_1) + \mathbb{P}(B|A_2)\mathbb{P}(A_2)} \\ &= \frac{(0.02)(88/100)}{(0.90)(12/100) + (0.02)(88/100)} \\ &= 0.14\end{aligned}$$

3

Variables Aleatorias

3.1. Introducción

Hasta ahora se asignaron probabilidades a eventos, es decir, a conjuntos de resultados de muestra. Los eventos con los que trabajamos estaban compuestos por un número finito o infinito numerable de resultados de muestra, en cuyo caso la probabilidad del evento era simplemente la suma de las probabilidades asignadas a sus resultados.

El primer objetivo de esta sección es incorporar otras formas de asignar probabilidades a los resultados de muestra. Al hacerlo, “redefiniremos” los espacios muestrales utilizando funciones conocidas como **variables aleatorias**.

En general, las variables aleatorias son funciones que asocian números con algún atributo de un resultado de muestra que se considera especialmente importante. Si X denota la variable aleatoria y ω denota un resultado de muestra, entonces $X(\omega) = x$, donde x es un número real. Las variables aleatorias a menudo pueden crear un espacio muestral drásticamente más simple.

La estadística y la minería de datos se ocupan de los datos. ¿Cómo vinculamos los espacios muestrales y los eventos con los datos? El vínculo lo proporciona el concepto de variable aleatoria.

Definición 3.1 *Variable aleatoria.* Una variable aleatoria es un mapeo:

$$X : \Omega \rightarrow \mathbb{R}$$

que asigna un número real $X(\omega)$ para cada resultado ω .

Ejemplo. Supongamos el lanzamiento de una moneda 10 veces. Sea $X(\omega)$ el número de caras en la secuencia de ω . En este caso, podemos observar que si $\omega = HHTHHTHHTT$, entonces $X(\omega) = 6$.

Ejemplo. Sea $\Omega = \{(x, y); x^2 + y^2 \leq 1\}$ el disco unitario. Consideremos extraer un punto al azar de Ω . Más adelante precisaremos esta idea. Un resultado típico tiene la forma $\omega = (x, y)$. Algunos ejemplos de variables aleatorias en este caso serían:

- $X(\omega) = x$, proyección sobre el eje horizontal.
- $Y(\omega) = y$, proyección sobre el eje vertical.
- $Z(\omega) = |x| + |y|$, distancia del taxista.
- $W(\omega) = \sqrt{x^2 + y^2}$, distancia al centro del círculo.

Ejemplo. Ahora incorporemos otros conceptos usando conceptos conocidos. Dada una variable aleatoria X y un subconjunto A de la recta real, definamos $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$, y sea:

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(X^{-1}(A)) \\ &= \mathbb{P}(\{\omega \in \Omega; X(\omega) \in A\}) \\ \mathbb{P}(X = x) &= \mathbb{P}(X^{-1}(x)) \\ &= \mathbb{P}(\{\omega \in \Omega; X(\omega) = x\})\end{aligned}$$

Notemos que X denota la variable aleatoria y x denota un valor particular de X .

Ejemplo. Supongamos el lanzamiento de una moneda dos veces y sea X el número de caras. Entonces,

$$\begin{aligned}\mathbb{P}(X = 0) &= \mathbb{P}(\{TT\}) = \frac{1}{4} \\ \mathbb{P}(X = 1) &= \mathbb{P}(\{HT, TH\}) = \frac{1}{2} \\ \mathbb{P}(X = 2) &= \mathbb{P}(\{HH\}) = \frac{1}{4}\end{aligned}$$

La variable aleatoria y su distribución se pueden resumir de la siguiente

manera:

ω	$\mathbb{P}(\{\omega\})$	$X(\omega)$	x	$\mathbb{P}(X = x)$
TT	$1/4$	0	0	$1/4$
TH	$1/4$	1	1	$1/2$
HT	$1/4$	1	2	$1/4$
HH	$1/4$	2		

¿Cómo se podría generalizar a n lanzamientos.

Ejemplo. Considere el siguiente experimento: todos los días durante un mes usted copia cada número que aparece en las portadas del periódico. Esos números necesariamente serían extremadamente diversos: uno podría ser la edad de una celebridad que acaba de morir, otro podría informar la tasa de interés que se paga actualmente por los bonos del Tesoro del gobierno, y otro podría dar el número de metros cuadrados de espacio comercial recientemente inaugurado.

Supongamos que luego calcula la proporción de esos números cuyo primer dígito es un 1, la proporción cuyo primer dígito es un 2, y así sucesivamente.

¿Qué relación esperaría que tuvieran esas proporciones? ¿Los números que comienzan con un 2, por ejemplo, aparecerían con tanta frecuencia como los números que comienzan con un 6?

Sea $\mathbb{P}(d)$ la probabilidad de que el primer dígito de un “número del periódico” sea d , $d = 1, 2, \dots, 9$.

Es probable que nuestra intuición nos diga que los nueve primeros dígitos deberían ser igualmente probables, es decir, $\mathbb{P}(1) = \mathbb{P}(2) = \dots = \mathbb{P}(9) = 1/9$.

Dada la aleatoriedad de los números, no hay ninguna razón obvia por la que un dígito deba ser más común que otro. Sin embargo, nuestra intuición estaría equivocada: los primeros dígitos no son igualmente probables. De hecho, ¡ni siquiera están cerca de ser igualmente probables!

El mérito de haber hecho este notable descubrimiento corresponde a Simon Newcomb, un matemático que observó hace más de cien años que algunas partes de las tablas de logaritmos se utilizan más que otras. En concreto, las páginas al principio de dichas tablas están más dobladas que las páginas al final, lo que sugiere que los usuarios tienen más ocasiones de buscar logaritmos de números que comienzan con dígitos pequeños que de números que comienzan con dígitos grandes. Casi cincuenta años después, un físico, Frank Benford, reexaminó la afirmación de Newcomb con más detalle y buscó una explicación matemática. Lo que ahora se conoce como la **ley de Benford** afirma que los primeros dígitos de muchos tipos diferentes de mediciones, o

combinaciones de mediciones, a menudo siguen el modelo de probabilidad discreta:

$$\mathbb{P}(d) = \log \left(1 + \frac{1}{d} \right), \text{ para } d = 1, 2, \dots, 9$$

La siguiente tabla muestra la comparación entre una distribución uniforme de probabilidad y la Ley de Benford.

d	Ley Uniforme	Ley de Benford
1	0.111	0.301
2	0.111	0.176
3	0.111	0.125
4	0.111	0.097
5	0.111	0.079
6	0.111	0.067
7	0.111	0.058
8	0.111	0.051
9	0.111	0.046

3.2. Funciones de distribución de probabilidad y Funciones de probabilidad

Dada una variable aleatoria X , definimos la función de distribución acumulativa (o función de distribución de probabilidad) de la siguiente manera.

Definición 3.2 *Función de Distribución de Probabilidad (Función de Distribución Acumulada) o CDF.* La función de distribución de probabilidad, o *CDF*, es la función $F_X : \mathbb{R} \rightarrow [0, 1]$ definido como:

$$F_X(x) = \mathbb{P}(X \leq x)$$

Más adelante veremos que la función de distribución de probabilidad contiene efectivamente toda la información sobre la variable aleatoria. A veces escribimos la función de distribución de probabilidad como F en lugar de F_x .

Ejemplo. Supongamos el lanzamiento de una moneda dos veces y sea X

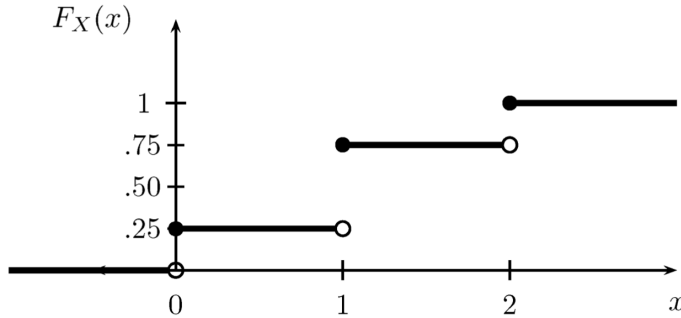


Figura 3.1: Ejemplo de una CDF, retomado de Wasserman (2004, p32) Wasserman 2004

el número de caras. Entonces,

$$\begin{aligned}\mathbb{P}(X = 0) &= \mathbb{P}(\{TT\}) = \frac{1}{4} \\ \mathbb{P}(X = 1) &= \mathbb{P}(\{HT, TH\}) = \frac{1}{2} \\ \mathbb{P}(X = 2) &= \mathbb{P}(\{HH\}) = \frac{1}{4}\end{aligned}$$

La función de distribución de probabilidad es:

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{4} & \text{si } 0 \leq x < 1 \\ \frac{3}{4} & \text{si } 1 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

La CDF se muestra en la Figura 3.1. Aunque este ejemplo es simple, estúdielo con atención. Las CDF pueden ser muy confusas. Observe que la función es continua hacia la derecha, no decreciente y que está definida para todo x , aunque la variable aleatoria solo toma los valores 0, 1 y 2. ¿Queda claro por qué $F_x(1.4) = .75$?

El siguiente resultado muestra que la CDF determina completamente la distribución de una variable aleatoria.

Teorema 3.1 *Sea X que tiene una CDF F y sea Y con una CDF G . Si $F(x) = G(x)$ para todo x , entonces $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ para todo evento (medible) A .*

Teorema 3.2 Una función F que mapea la recta real a $[0,1]$ es una CDF para alguna probabilidad \mathbb{P} si y solo si F satisface las siguientes tres condiciones:

1. F no es decreciente: $x_1 < x_2$ implica que $F(x_1) \leq F(x_2)$

2. F está normalizada:

$$\lim_{x \rightarrow -\infty} F(x) = 0$$

y

$$\lim_{x \rightarrow \infty} F(x) = 1$$

3. F es continua por la derecha: $F(x) = F(x^+)$ para todo x , donde

$$F(x^+) = \lim_{y \rightarrow x, y > x} F(y)$$

Demostración. (Para el curso de análisis real 1)

3.3. Variables Aleatorias Discretas y Continuas

Definición 3.3 *Función de Densidad de Probabilidad o PDF o Función de Masa de Probabilidad de una variable aleatoria discreta.*

Una variable aleatoria X es **discreta** si toma una cantidad contable de valores $\{x_1, x_2, \dots\}$. Definimos la función de densidad de probabilidad para X mediante $f_X(x) = \mathbb{P}(X = x)$.

Así, $f_X(x) > 0$ para todo $x \in \mathbb{R}$ y

$$\sum_i f_X(x_i) = 1$$

Algunas veces escribiremos f en lugar de f_X . La CDF de X está relacionada a f_X por

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

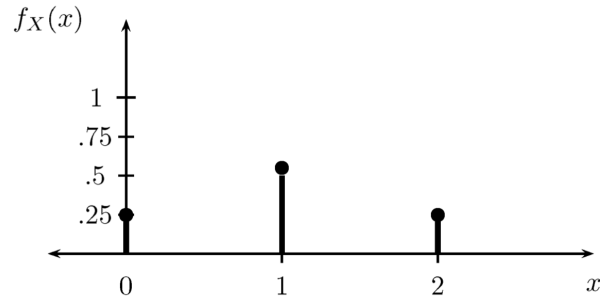


Figura 3.2: Ejemplo de una PDF, retomado de Wasserman (2004, p23) Wasserman 2004

Ejemplo. Supongamos el lanzamiento de una moneda dos veces y sea X el número de caras. Entonces,

$$\begin{aligned}\mathbb{P}(X = 0) &= \mathbb{P}(\{TT\}) = \frac{1}{4} \\ \mathbb{P}(X = 1) &= \mathbb{P}(\{HT, TH\}) = \frac{1}{2} \\ \mathbb{P}(X = 2) &= \mathbb{P}(\{HH\}) = \frac{1}{4}\end{aligned}$$

La función de probabilidad es:

$$f_X(x) = \begin{cases} \frac{1}{4} & \text{si } x = 0 \\ \frac{1}{2} & \text{si } x = 1 \\ \frac{1}{4} & \text{si } x = 2 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

La Figura 3.2 ilustra esta función de probabilidad.

Ejemplo. Supongamos el lanzamiento de un dado dos veces. Sean i y j la caras observadas en el primer y segundo dado, respectivamente. Definamos la variable aleatoria X como la suma de ambas caras: $X(i, j) = i + j$. Econtremos la $f_X(x)$.

Recordemos que la definición indica que cada $f_X(x)$ es la suma de las probabilidades de los resultados que son mapeados por X en cada valor de

x . Por ejemplo:

$$\begin{aligned}
 \mathbb{P}(X = 5) &= \mathbb{P}(\{\omega \in \Omega : X(\omega) = 5\}) \\
 &= \mathbb{P}(\{(1, 4), (4, 1), (2, 3), (3, 2)\}) \\
 &= \mathbb{P}(1, 4) + \mathbb{P}(4, 1) + \mathbb{P}(2, 3) + \mathbb{P}(3, 2) \\
 &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\
 &= \frac{4}{36}
 \end{aligned}$$

Lo anterior asumiendo que el dado es justo. Podemos calcular los otros valores de x de forma similar. Los cuales mostramos acontinuación:

x	$f_X(x) = \mathbb{P}(X = x)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Así, la función de probabilidad es:

$$f_X(x) = \begin{cases} \frac{1}{36} & \text{si } x = 2 \\ \frac{2}{36} & \text{si } x = 2 \\ \frac{3}{36} & \text{si } x = 4 \\ \frac{4}{36} & \text{si } x = 5 \\ \frac{5}{36} & \text{si } x = 6 \\ \frac{6}{36} & \text{si } x = 7 \\ \frac{5}{36} & \text{si } x = 8 \\ \frac{4}{36} & \text{si } x = 9 \\ \frac{3}{36} & \text{si } x = 10 \\ \frac{2}{36} & \text{si } x = 11 \\ \frac{1}{36} & \text{si } x = 12 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Definición 3.4 *Función de Densidad de Probabilidad o PDF o Función de Masa de Probabilidad de una variable aleatoria continua.*

Una variable aleatoria X es **continua** si existe una función f_X tal que $f_X(x) \geq 0$ para todo x ,

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

y para cada $a \leq b$,

$$\mathbb{P}(a < x < b) = \int_a^b f_X(x) dx$$

La función f_X es llamada la función de densidad de probabilidad. De la cual tenemos que:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

y que $f_X(x) = F'_X(x)$ para todos los puntos x en los que F_X es diferenciable.

Ejemplo. El modelo equiprobable equivalente en el caso continuo es la función $f(t)$ definida por:

$$f(t) = \begin{cases} 1/(b-a) & \text{para todo } t \in [a, b] \\ 0 & \text{en cualquier otro caso} \end{cases}$$

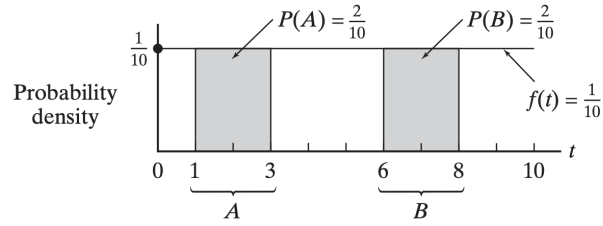


Figura 3.3: CDF de una Uniforme (0,10), retomado de Larsen (2012, p129)
Larsen y Marx 2012

Supongamos $a = 0$ y $b = 10$, y sean $A = [1, 3]$ y $B = [6, 8]$. Entonces:

$$\mathbb{P}(A) = \int_1^3 \left(\frac{1}{10} \right) dt = \frac{2}{10} = \mathbb{P}(B) = \int_6^8 \left(\frac{1}{10} \right) dt$$

La Figura 3.3 ilustra esta CDF.

Ejemplo. Supongamos que X tiene una PDF:

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

Claramente $f_X(x) \geq 0$ y

$$\int_0^1 f_X(x) dx = 1$$

Se dice que una variable aleatoria con esta densidad tiene una distribución Uniforme (0, 1). Esto pretende captar la idea de elegir un punto al azar entre 0 y 1. La CDF se da por

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x > 1 \end{cases}$$

La Figura 3.4 ilustra esta CDF.

Ejemplo. Sea:

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

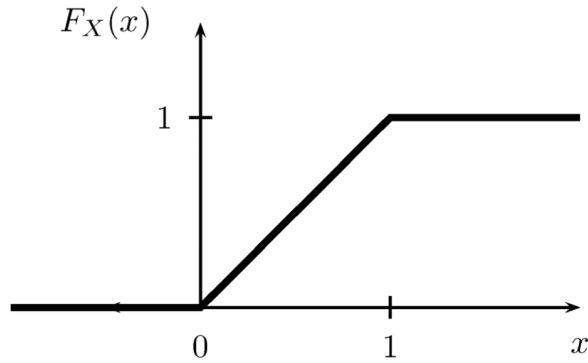


Figura 3.4: CDF de una Uniforme (0,1), retomado de Wasserman (2004, p24)
Wasserman 2004

Notemos que esta no es un PDF, ya que:

$$\int_D f(x)dx = \int_0^\infty \frac{1}{1+x}dx = \log(\infty) = \infty$$

Ejemplo. Determinemos si:

$$f(t) = \begin{cases} 3t^2 & \text{para } 0 \leq t \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

puede ser definida como función de probabilidad para un espacio continuo.
Para ello requerimos probar:

1. $f(t) \geq 0$ para todo t , lo cual es claro.
2. Verifiquemos que:

$$\int_0^1 f(t)dt = \int_0^1 3t^2 dt = t^3 \Big|_0^1 = 1$$

Lema. Sea F la CDF de una variable aleatoria X (tanto continua como discreta). Entonces:

1. $\mathbb{P}(X = x) = F(x) - F(x^-)$ donde $F(x^-) = \lim_{y \uparrow x} F(y)$
2. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$

3. $\mathbb{P}(X > x) = 1 - F(x)$

4. Si X es continua, entonces:

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) \\ &= \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) \\ &= \mathbb{P}(a \leq X \leq b) \end{aligned}$$

Como último concepto de esta sección, introduzcamos el concepto de función cuantil.

Definición 3.5 *Función Cuantil.* Sea X una variable aleatoria con CDF F . La CDF inversa o función cuantil se define por

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

para $q \in [0, 1]$. Si F es estrictamente creciente y continua, entonces $F^{-1}(q)$ es el único número real x tal que $F(x) = q$.

Ejemplo. Llamamos $F^{-1}(1/4)$ al primer cuartil, $F^{-1}(1/2)$ a la mediana (o segundo cuartil) y $F^{-1}(3/4)$ al tercer cuartil.

3.4. Algunas variables aleatorias importantes discretas

Notación: es tradicional escribir $X \sim F$ para indicar que X tiene una distribución F . Esta notación es desafortunada, ya que el símbolo \sim también se utiliza para denotar una aproximación. La notación $X \sim F$ es tan generalizada que no podemos evitarla. Lea $X \sim F$ como “ X tiene una distribución F ”, no como “ X es aproximadamente F ”.

Ejemplo. Distribución Uniforme Discreta. Sea $k > 1$ un entero. Supongamos que X tiene una función de probabilidad dada por:

$$f(x) = \begin{cases} \frac{1}{k} & \text{para } x = 1, 2, \dots, k \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Decimos que X tiene una distribución uniforme en $\{1, 2, \dots, k\}$.

Es claro que $f(x) \geq 0$, para todo x .

Ahora comprobemos que

$$\begin{aligned}\sum_i f(x_i) &= 1 \\ \sum_i f(x_i) &= \sum_{x=1}^k \frac{1}{k} \\ &= \frac{k}{k} = 1\end{aligned}$$

Ejemplo. Distribución Bernoulli. Sea X una representación del lanzamiento de una moneda. Entonces, $\mathbb{P}(X = 0) = p$ y $\mathbb{P}(X = 1) = 1 - p$, para algún $p \in [0, 1]$. Decimos que X tiene una distribución de Bernoulli escrita como $X \sim \text{Bernoulli}(p)$. De esta forma, la función de densidad de probabilidad estará dada por:

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

Es claro que $f(x) \geq 0$, para todo x .

Ahora comprobemos que

$$\begin{aligned}\sum_x f(x) &= 1 \\ \sum_x f(x) &= \sum_x p^x(1-p)^x \\ &= (1-p) + p = 1\end{aligned}$$

Ejemplo. Distribución Binomial. Supongamos que tenemos una moneda que cae cara con probabilidad p para algún $0 \leq p \leq 1$. Lanza la moneda n veces y sea X el número de caras. Supóngase que los lanzamientos son independientes.

Sea $f(x) = \mathbb{P}(X = x)$ la función de masa. Se puede demostrar que:

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

Donde $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

Una variable aleatoria con esta PDF es denominada variable aleatoria Binomial y se escribe $X \sim \text{Binomial}(n, p)$.

Para motivar la forma en que llegamos a una expresión así, partamos de un ejemplo. Supongamos el caso del lanzamiento de una moneda, en el cual la probabilidad p de que salga aguilá es desconocida. Ahora supongamos que hacemos 3 lanzamientos de la moneda.

Imaginemos que los resultados fueron: $\{A, A, S\}$, donde A denota aguilá y S denota sol. Basados en esos resultados, ¿cuál es el valor que considera apropiado 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 aguilá en el lanzamiento de una moneda, es decir:

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

Así, el modelo general puede ser escrito como:

$$f(x) = p^x(1-p)^{1-x} = \begin{cases} p & \text{si } x = 1 \\ (1-p) & \text{si } x = 0 \end{cases}$$

Ahora, supongamos una sucesión de variables aleatorias independientes $\{X_i : i = 1, 2, \dots, n\}$ de forma que podemos describir el fenómeno del lanzamiento de una moneda tres veces como:

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

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

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

De forma general, el lanzamiento de una moneda n veces estará dado por la sucesión de v.a.'s:

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

Donde x_i , para $i = 1, 2, \dots, n$. Entonces, el problema general estará dado por:

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) &= p^{x_1}(1-p)^{1-x_1} \dots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

En este caso, podemos definir a $x = \sum_{i=1}^n x_i$ el número de veces que cae aguilas en n lanzamientos. Pero, ¿de cuántas formas posibles podemos obtener a x ? La respuesta nos lleva a:

$$\binom{n}{x} p^x (1-p)^{n-x}$$

Finalmente, es claro que $f(x) \geq 0$, para todo x .

Ahora comprobemos que

$$\begin{aligned} \sum_x f(x) &= 1 \\ \sum_{x=0}^n f(x) &= \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} \\ &= ((1-p) + p)^n = 1 \end{aligned}$$

Ejemplo. En comunicaciones de diversos tipos, un mensaje enviado puede no recibirse correctamente porque el canal de comunicaciones es “ruidoso”. En particular, un bit transmitido puede cambiarse con una probabilidad p . Un método para resolver este problema es enviar el bit cinco veces y luego usar la mayoría de estos cinco bits recibidos como el mensaje deseado. Según este esquema, ¿cuál es la probabilidad de que el mensaje se reciba correctamente?

Los cinco bits recibidos forman una secuencia de cinco ensayos de Bernoulli con una probabilidad p de que el bit se cambie (éxito). Entonces, el mensaje se recibe correctamente si el número de bits cambiados (éxitos) es 0, 1 o 2. La probabilidad de que esto ocurra es:

$$\mathbb{P}(X \geq 2) = \sum_{x=0}^2 \binom{5}{x} p^x (1-p)^{5-x}$$

Ejemplo. X tiene una distribución Poisson con parámetro λ , escrita como $X \sim \text{Poisson}(\lambda)$ si:

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

Es claro que $f(x) \geq 0$, para todo x .
 Ahora comprobemos que

$$\begin{aligned}\sum_x f(x) &= 1 \\ \sum_{x=0}^{\infty} f(x) &= \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} e^{\lambda} = 1\end{aligned}$$

λ se denomina parámetro de la distribución. La distribución de Poisson es usada como un modelo de conteo y es adecuada para predecir la cantidad de llamadas telefónicas que llegan a una central telefónica determinada en un período de tiempo determinado, la cantidad de partículas emitidas por una fuente radiactiva en un período de tiempo determinado, el número de accidentes de tráfico, etc.

3.5. Algunas variables aleatorias importantes continuas

Ejemplo. La Distribución Uniforme. X tiene una distribución *Uniforme*(a, b), escrita como $X \sim \text{Uniforme}(a, b)$, para $a < b$ si:

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

La función de distribución es:

$$F(x) = \int_a^x \frac{1}{b-a} dt = \begin{cases} 0 & \text{para } x < a \\ \frac{x-a}{b-a} & \text{para } x \in [a, b] \\ 1 & \text{para } x > b \end{cases}$$

Ejemplo. Normal (Gaussiana). X tiene una distribución normal con parámetros μ y σ , denotada por $X \sim N(\mu, \sigma^2)$, si:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{1}{2\sigma^2}(x-\mu)^2)}, \text{ para } x \in \mathbb{R}$$

Es claro que $f(x) \geq 0$, para todo x .

Probar:

$$\begin{aligned}\int_{-\infty}^{\infty} f(x) &= 1 \\ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{1}{2\sigma^2}(x-\mu)^2)} dx &= 1\end{aligned}$$

Un caso particular es conocida como la distribución normal estándar. Esta situación se observa cuando $\mu = 0$ y $\sigma = 1$. Tradicionalmente, esta se denota como Z . Estos son algunos resultados:

1. Si $X \sim N(\mu, \sigma^2)$, entonces $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$
2. Si $Z \sim N(0, 1)$, entonces $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$

Ejemplo. Distribución Exponencial. X tiene una distribución exponencial con parámetro β , denotada por $X \sim \text{Exp}(\beta)$, si:

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

3.6. Distribuciones bivariadas

Hasta ahora, hemos introducido la terminología básica para describir el comportamiento probabilístico de una sola variable aleatoria. Esta información, aunque adecuada para muchos problemas, es insuficiente cuando más de una variable interesa al experimentador. Los investigadores médicos, por ejemplo, siguen explorando la relación entre el colesterol en sangre y las enfermedades cardíacas y, más recientemente, entre el colesterol “bueno” y el colesterol “malo”.

El punto es que hay muchas situaciones en las que dos variables aleatorias relevantes, digamos, X e Y , están definidas en el mismo espacio muestral. Sin embargo, conocer solo $f_X(x)$ y $f_Y(y)$ no proporciona necesariamente suficiente información para caracterizar el importantísimo comportamiento simultáneo de X e Y . El propósito de esta sección es presentar los conceptos, definiciones y técnicas matemáticas asociadas con distribuciones basadas en dos (o más) variables aleatorias.

3.6.1. Distribuciones discretas

Como vimos en el caso de una sola variable, la función de densidad de probabilidad se define de manera diferente según si la variable aleatoria es discreta o continua. La misma distinción se aplica a las funciones de densidad de probabilidad conjuntas. Comenzaremos con un análisis de las funciones de densidad de probabilidad conjuntas tal como se aplican a dos variables aleatorias discretas.

Dado un par de variables aleatorias discretas X e Y , definiremos la función de masa conjunta o función de densidad de probabilidad mediante $f(x, y) = \mathbb{P}(X = x \text{ e } Y = y)$. De ahora en adelante, escribiremos $\mathbb{P}(X = x \text{ e } Y = y)$ como $\mathbb{P}(X = x, Y = y)$. Escribimos f como $f_{x,y}$ cuando queremos ser más explícitos.

Ejemplo. A continuación, presentamos una distribución bivariada para dos variables aleatorias X e Y , cada una de las cuales toma valores 0 o 1:

	$Y = 0$	$Y = 1$	
$X = 0$	1/9	2/9	1/3
$X = 1$	2/9	4/9	2/3
	1/3	2/3	1

Por lo tanto, $f(1, 1) = \mathbb{P}(X = 1, Y = 1) = 4/9$.

Definición 3.6 Supongamos que Ω es un espacio muestral discreto en el que se definen dos variables aleatorias, X e Y . La función de densidad de probabilidad conjunta de X e Y (o función de densidad de probabilidad conjunta) se denota $p_{X,Y}(x, y)$ o $f_{X,Y}(x, y)$, donde:

$$f_{X,Y}(x, y) = \mathbb{P}(\{\omega | X(\omega) = x \text{ y } Y(\omega) = y\}) = \mathbb{P}(X = x, Y = y)$$

Ejemplo. Un supermercado tiene dos líneas exprés. Sean X e Y el número de clientes en la primera y en la segunda, respectivamente, en un momento dado. Fuera de las horas pico, la función de densidad de flujo conjunta de X e Y se resume en la siguiente tabla:

	$X = 0$	$X = 1$	$X = 2$	$X = 3$	
$Y = 0$	0.1	0.2	0	0	0.3
$Y = 1$	0.2	0.25	0.05	0	0.5
$Y = 2$	0	0.05	0.05	0.025	0.125
$Y = 3$	0	0	0.025	0.05	0.075
	0.3	0.5	0.125	0.075	1

Determinemos $\mathbb{P}(|X - Y| = 1)$, es decir, la probabilidad de que X y Y difieran exactamente 1.

$$\begin{aligned}
 \mathbb{P}(|X - Y| = 1) &= \sum_{|X-Y|=1} \sum f_{X,Y}(x, y) \\
 &= f_{X,Y}(0, 1) + f_{X,Y}(1, 0) + f_{X,Y}(1, 2) + f_{X,Y}(2, 1) \\
 &\quad + f_{X,Y}(2, 3) + f_{X,Y}(3, 2) \\
 &= 0.2 + 0.2 + 0.05 + 0.05 + 0.025 + 0.025 \\
 &= 0.55
 \end{aligned}$$

¿Qué piensan de que sea simétrica $f_{X,Y}(x, y)$? ¿Cuál es $\mathbb{P}(|X - Y| \geq 2)$?

3.6.2. Distribuciones continuas

Definición 3.7 En el caso continuo, llamamos a una función $f(x, y)$ una PDF para las variables aleatorias (X, Y) si:

1. $f(x, y) \geq 0$ para todo (x, y) ;
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$, y
3. Para cualquier conjunto $A \subset \mathbb{R} \times \mathbb{R}$, $\mathbb{P}((X, Y) \in A) = \int \int_A f(x, y) dx dy$.

En el caso discreto o continuo definimos la CDF conjunta como $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$.

Ejemplo. Supongamos que la variación de dos variables aleatorias continuas, X e Y , se puede modelar mediante la función de densidad de probabilidad conjunta $f_{X,Y}(x, y) = cxy$, para $0 < y < x < 1$. Halle c .

Por inspección, $f_{X,Y}(x, y)$ será no negativa siempre que $c \geq 0$. Sin embargo, la c particular que califica a $f_{X,Y}(x, y)$ como una función de densidad de probabilidad conjunta es la que hace que el volumen bajo $f_{X,Y}(x, y)$ sea

igual a 1. Pero:

$$\begin{aligned}
\int \int_D cxy dx dy &= c \int_0^1 \int_0^x xy dy dx \\
&= c \int_0^1 x \left(\frac{y^2}{2} \Big|_0^x \right) dx \\
&= c \int_0^1 \left(\frac{x^3}{2} \right) dx \\
&= c \left(\frac{x^4}{8} \Big|_0^1 \right) \\
&= c \left(\frac{1}{8} \right) = 1
\end{aligned}$$

Por lo tanto $c = 8$.

3.6.3. Distribuciones marginales

Empecemos por una definición.

Definición 3.8 *Función de densidad marginal para X discreta.* Si (X, Y) tienen una distribución conjunta con función de densidad $f_{X,Y}$, entonces, la función de densidad marginal de X está definida por:

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f(x, y)$$

Función de densidad marginal para Y discreta. Si (X, Y) tienen una distribución conjunta con función de densidad $f_{X,Y}$, entonces, la función de densidad marginal de Y está definida por:

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f(x, y)$$

Las funciones de distribución marginal correspondientes se denotan por F_X y F_Y .

Ejemplo. Supongamos que $f_{X,Y}$ está dada en la tabla que sigue. La distribución marginal de X corresponde a los totales de las filas y la distribución

marginal de Y corresponde a los totales de las columnas.

	$Y = 0$	$Y = 1$	
$X = 0$	1/10	2/10	3/10
$X = 1$	3/10	4/10	7/10
	4/10	6/10	1

En este caso, podemos encontrar que $f_X(0) = 3/10$ y $f_X(1) = 7/10$.

Definición 3.9 *Función de densidad marginal el caso continuo.* Para variables aleatorias continuas, la función de densidad marginal será:

$$f_X(x) = \int f(x, y) dy$$

y

$$f_Y(y) = \int f(x, y) dx$$

Las funciones de distribución marginal correspondientes se denotan por F_X y F_Y .

Ejemplo. Supongamos que:

$$f_{X,Y}(x, y) = \begin{cases} e^{-(x+y)} & \text{si } x, y \geq 0 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos f_X .

$$f_X(x) = e^{-x} \int_0^{\infty} e^{-y} dy = e^{-x}$$

$$f_X(x) = \begin{cases} e^{-x} & \text{si } x \geq 0 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Ejemplo. Supongamos que:

$$f(x, y) = \begin{cases} x + y & \text{si } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos f_Y .

$$f_Y(y) = \int_0^1 (x + y) dx = \int_0^1 x dx + \int_0^1 y dx = \frac{1}{2} + y$$

$$f_Y(y) = \begin{cases} \frac{1}{2} + y & \text{si } 0 \leq y \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Ejemplo. Sea (X, Y) con una densidad:

$$f(x, y) = \begin{cases} \frac{21}{4}x^2y & \text{si } x^2 \leq y \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Determinemos f_X .

$$f_X(x) = \int f(x, y)dy = \frac{21}{4}x^2 \int_{x^2}^1 ydy = \frac{21}{8}x^2(1 - x^4)$$

$$f_X(x) = \begin{cases} \frac{21}{8}x^2(1 - x^4) & \text{si } -1 \leq x \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Definición 3.10 *Función de Distribución Conjunta.* Sea (X, Y) un vector continuo con función de distribución $F(x, y)$. Se dice que (X, Y) es absolutamente continuo si existe una función no negativa e integrable $f(x, y) : \mathbb{R}^2 \rightarrow [0, \infty)$, tal que para todo $(x, y) \in \mathbb{R}^2$, se cumple la igualdad:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v)dudv$$

Notemos que, cuando $f(x, y)$ es continua:

$$f(x, y) = \frac{\partial^2}{\partial y \partial x} F(x, y)$$

Definición 3.11 *Función de Distribución Marginal.* Sea (X, Y) un vector con función de distribución $F(x, y)$. A la función:

$$F(x) = \lim_{y \rightarrow \infty} F(x, y)$$

se le conoce como la función de distribución marginal de X . Análogamente, se define la función de distribución marginal de Y como:

$$F(y) = \lim_{x \rightarrow \infty} F(x, y)$$

Ejemplo. Supongamos que:

$$F_{X,Y}(x, y) = \begin{cases} 0 & \text{si } x < 0 \text{ o } y < 0 \\ xy/4 & \text{si } 0 \leq x, y < 2 \\ x/2 & \text{si } 0 \leq x < 2, y \geq 2 \\ y/2 & \text{si } 0 \leq y < 2, x \geq 2 \\ 1 & \text{si } x \geq 2, y \geq 2 \end{cases}$$

Esta función está definida de manera distinta en cada una de las cinco regiones disjuntas y exhaustivas del plano Cartesiano dadas por las condiciones anteriores.

Para encontrar, por ejemplo, la función de distribución marginal $F_X(x)$ simplemente tenemos que hacer la variable y tender a infinito en las regiones donde ello sea posible. Ello puede hacerse en las regiones dadas por las condiciones del primer, tercer y quinto renglón de la lista anterior. Esto da como resultado:

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ x/2 & \text{si } 0 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

3.6.4. Variables Aleatorias Independientes

Definición 3.12 *Variables aleatorias independientes.* Dos variables aleatorias X y Y son **independientes** si, para cada evento A y B ,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \quad (3.1)$$

En principio, para comprobar si X e Y son independientes necesitamos comprobar la ecuación (3.1) para todos los subconjuntos A y B . Afortunadamente, tenemos el siguiente resultado que enunciamos para variables aleatorias continuas aunque también es cierto para variables aleatorias discretas.

Teorema 3.3 Sean X e Y variables aleatorias con función de densidad de probabilidad conjunta (PDF) $f_{X,Y}$. Entonces X es independiente de Y si y solo si $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ para todos los valores x e y .

Ejemplo. Sean X e Y variables aleatorias que tienen la siguiente distribución:

	$Y = 0$	$Y = 1$	
$X = 0$	1/4	1/4	1/2
$X = 1$	1/4	1/4	1/2
	1/2	1/2	1

Entonces, $f_X(0) = f_X(1) = 1/2$ y $f_Y(0) = f_Y(1) = 1/2$. X y Y son independientes ya que:

$$f_X(0)f_Y(0) = 1/2 \cdot 1/2 = f_{X,Y}(0,0) = 1/4$$

$$f_X(0)f_Y(1) = 1/2 \cdot 1/2 = f_{X,Y}(0,1) = 1/4$$

$$f_X(1)f_Y(0) = 1/2 \cdot 1/2 = f_{X,Y}(1,0) = 1/4$$

$$f_X(1)f_Y(1) = 1/2 \cdot 1/2 = f_{X,Y}(1,1) = 1/4$$

Por lo tanto, X y Y son independientes.

Ejemplo. Supongamos que X y Y son independientes y que ambas tienen la misma función de densidad:

$$f(x) = \begin{cases} 2x & \text{si } 0 \leq x \leq 1 \\ 0 & \text{e.c.o.c.} \end{cases}$$

$$f(y) = \begin{cases} 2y & \text{si } 0 \leq y \leq 1 \\ 0 & \text{e.c.o.c.} \end{cases}$$

De esta forma, la función de densidad conjunta sería:

$$f(x, y) = f_X(x)f_Y(y) = \begin{cases} 4xy & \text{si } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{e.c.o.c.} \end{cases}$$

Ahora, determinemos $\mathbb{P}(X + Y \leq 1)$. Para ello:

$$\begin{aligned} \mathbb{P}(X + Y \leq 1) &= \int \int_{x+y \leq 1} f(x, y) dy dx \\ &= 4 \int_0^1 x \int_0^{1-x} y dy dx \\ &= 4 \int_0^1 x \frac{(1-x)^2}{2} dx \\ &= \frac{1}{6} \end{aligned}$$

El siguiente resultado nos ayuda a encontrar o verificar independencia.

Teorema 3.4 *Supongamos que el rango o dominio de X y Y es un rectángulo. $f(x, y) = g(x)h(y)$ para algunas funciones g y h , si y solo si X y Y son independientes.*

Demostración. Supongamos que X y Y son independientes. Entonces, $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y) = F_X(x)F_Y(y)$, y podemos escribir:

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) \\ &= \frac{\partial^2}{\partial x \partial y} F_X(x)F_Y(y) \\ &= \frac{d}{dx} F_X(x) \frac{d}{dy} F_Y(y) \\ &= f_X(x)f_Y(y) \end{aligned}$$

Ahora necesitamos mostrar que $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ implica que X y Y son independientes. Partamos:

$$\begin{aligned} \mathbb{P}(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x, y) dx dy \\ &= \int_A \int_B g(x)h(y) dx dy \\ &= \int_A g(x) dx \int_B h(y) dy \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \end{aligned}$$

Ejemplo. Sean X y Y con una función de densidad:

$$f(x, y) = \begin{cases} 2e^{-(x+2y)} & \text{si } x > 0, y > 0 \\ 0 & \text{e.c.o.c.} \end{cases}$$

El dominio de X y Y es el rectángulo $(0, \infty) \times (0, \infty)$. Podemos escribir: $f(x, y) = g(x)h(y)$, donde:

$$\begin{aligned} g(x) &= 2e^{-x} \\ h(x) &= e^{-2y} \end{aligned}$$

Así, X y Y son independientes.

Ejemplo. Supongamos que el comportamiento probabilístico de dos variables aleatorias X y Y es descrita por la PDF conjunta:

$$f_{X,Y}(x, y) = \begin{cases} 12xy(1-y) & \text{si } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{e.c.o.c.} \end{cases}$$

¿ X y Y son independientes?

Sí. Si asumimos que $g(x) = 12x$ y $h(y) = y(1-y)$.

3.6.5. Distribuciones Condicionales

Si X y Y son dicretas, entonces podemos calcular la distribución condicional de X dado que hemos observado a $Y = y$, como:

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

Definición 3.13 La función de densidad de probabilidad condicional en el caso discreto es:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Si $f_Y(y) > 0$.

Para el caso de funciones de distribución continuas la definición es la mismas, salvo por unos detalles.

Definición 3.14 La función de densidad de probabilidad condicional en el caso continuo es:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Asumiendo que $f_Y(y) > 0$. Entonces,

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y) dx$$

Ejemplo. Sea

$$f(x, y) = \begin{cases} x + y & \text{si } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{e.c.o.c} \end{cases}$$

Determinemos $\mathbb{P}(X < 1/4 | Y = 1/3)$. Antes hemos mostrado que:

$$f_Y(y) = \int_0^1 (x + y)dx = \int_0^1 xdx + \int_0^1 ydx = \frac{1}{2} + y$$

$$f_Y(y) = \begin{cases} \frac{1}{2} + y & \text{si } 0 \leq y \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Así,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x+y}{y+\frac{1}{2}}$$

Así,

$$\begin{aligned} \mathbb{P}\left(X < \frac{1}{4} \mid Y = \frac{1}{3}\right) &= \int_0^{1/4} f_{X|Y}\left(X \mid \frac{1}{3}\right) dx \\ &= \int_0^{1/4} \frac{x + \frac{1}{3}}{\frac{1}{3} + \frac{1}{2}} dx \\ &= \frac{11}{80} \end{aligned}$$

3.6.6. Transformaciones de Variables Aleatorias

Supongamos que X es una variable aleatoria con PDF f_X y CDF F_X . Sea $Y = r(X)$ una función de X , por ejemplo, $Y = X^2$ o $Y = e^X$. Llamamos $Y = r(X)$ una transformación de X . ¿Cómo calculamos la PDF y la CDF de Y ? En el caso discreto, la respuesta es fácil. La función de masa de Y está dada por:

$$\begin{aligned} f_Y(y) &= \mathbb{P}(Y = y) \\ &= \mathbb{P}(r(X) = y) \\ &= \mathbb{P}(\{x : r(x) = y\}) \\ &= \mathbb{P}(X \in r^{-1}(y)) \end{aligned}$$

Veámos un ejemplo.

Ejemplo. Supongamos que $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = 1/4$ y $\mathbb{P}(X = 0) = 1/2$. Sea $Y = X^2$, entonces,

$$\begin{aligned} \mathbb{P}(Y = 0) &= \mathbb{P}(X = 0) = \frac{1}{2} \\ \mathbb{P}(Y = 1) &= \mathbb{P}(X = 1) + \mathbb{P}(X = -1) = \frac{1}{2} \end{aligned}$$

En resumen, lo que tenemos es

x	$f_X(x)$	y	$f_Y(y)$
-1	1/4	0	1/2
0	1/2	1	1/2
1	1/4		

Y toma algunos valores que X dado que la transformación no es uno a uno.

4

Esperanza

4.1. Valor esperado de una variable aleatoria

La media, o el valor esperado, de una variable aleatoria X es el valor promedio de X .

Definición 4.1 Valor Esperado. *El valor esperado o media o primer momento de una variable aleatoria X está definido por:*

$$\mathbb{E}[X] = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{si } X \text{ es discreta} \\ \int x f(x) dx & \text{si } X \text{ es continua} \end{cases} \quad (4.1)$$

Asumiendo que la suma (o la integral) está bien definida. Asumimos la siguiente notación para denotar a el valor esperado de X :

$$\mathbb{E}[X] = \int x dF(x) = \mu = \mu_X$$

Obs.: Para asegurar que $\mathbb{E}[X]$ está bien definida, decimos que $\mathbb{E}[X]$ existe si:

$$\int_x |x| dF_X(x) < \infty$$

En cualquier otro caso decidimos que la esperanza no existe.

Ejemplo. Sea $X \sim \text{Bernoulli}(p)$. Entonces:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{X=0}^1 xf(x) \\ &= 0 \times (1-p) + 1 \times (p) \\ &= p\end{aligned}$$

Ejemplo. Sea $X \sim \text{Binomial}(n, p)$. Entonces:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=0}^n x \cdot p_X(x) \\ &= \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^n x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^n \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\ &= np \sum_{x=1}^n \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\ &= np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} \\ &= np\end{aligned}$$

Ejemplo. Sea $X \sim \text{Uniforme}(-1, 3)$. Entonces,

$$\begin{aligned}\mathbb{E}[X] &= \int x dF_X(x) \\ &= \int x f_X(x) dx \\ &= \frac{1}{4} \int x dx \\ &= 1\end{aligned}$$

Ejemplo. Sea $X \sim \text{exp}(\mu)$. Entonces,

$$f_Y(y) = \begin{cases} \frac{1}{\mu} e^{-y/\mu} & \text{si } y \geq 0 \\ 0 & \text{e.c.o.c} \end{cases}$$

Determinemos:

$$\begin{aligned}
 \mathbb{E}[Y] &= \int_0^\infty y \frac{1}{\mu} e^{-y/\mu} dy \\
 &= \mu \int_0^\infty \frac{y}{\mu} e^{-y/\mu} \frac{dy}{\mu} \\
 &= \mu \int_0^\infty w e^{-w} dw \\
 &= \mu \left[-w e^{-w} + \int_0^\infty e^{-w} dw \right] \\
 &= \mu \left[-w e^{-w} - e^{-w} \right]_0^\infty \\
 &= \mu
 \end{aligned}$$

4.2. Propiedades de la esperanza

Teorema 4.1 Sean X y Y con esperanza finita y c una constante. Entonces:

1. $\mathbb{E}[c] = c$
2. $\mathbb{E}[cX] = c\mathbb{E}[X]$
3. Si $X \geq 0$, entonces, $\mathbb{E}[X] \geq 0$
4. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

En general,

Teorema 4.2 Si X_1, X_2, \dots, X_n son variables aleatorias y a_1, a_2, \dots, a_n son constantes, entonces:

$$\mathbb{E} \left(\sum_i a_i X_i \right) = \sum_i a_i \mathbb{E}[X_i] \quad (4.2)$$

Ejemplo. Sea $X \sim \text{Binomial}(n, p)$. Sabemos que en este caso el valor

esperado está dado por:

$$\begin{aligned}
 \mathbb{E}[X] &= \int x dF_X(x) \\
 &= \sum_x x f_X(x) \\
 &= \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} \\
 &= np
 \end{aligned}$$

Pero esta suma no fue facil de evaluar. En su lugar, podemos retomar que:

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

Donde, $X_i = \{0, 1\}$. Entonces, $\mathbb{E}[X_i] = 1 \times p + 0 \times (1-p) = p$.

De esta forma:

$$\mathbb{E}[X] = \mathbb{E}\left(\sum_i X_i\right) = \sum_i \mathbb{E}[X_i] = np$$

Teorema 4.3 Sea X_1, \dots, X_n variables aleatorias independientes. Entonces,

$$\mathbb{E}(\Pi_{i=1}^n X_i) = \Pi_{i=1}^n \mathbb{E}[X_i] \quad (4.3)$$

Obs: Notemos que la suma no requiere de la independencia, pero la multiplicación sí.

Ahora, es posible determinar la esperanza de una función de una variable aleatoria. En algunos casos es necesario calcular la esperanza de una función de una variable aleatoria, por ejemplo, si X es una variable aleatoria, entonces es claro que $Y = X^2$ es una función de X y es también una variable aleatoria. Si quisiéramos calcular la esperanza de Y según la definición tendríamos que calcular:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy$$

Para lo cual se necesita encontrar primero la función de densidad de Y y ello en general no es fácil. El siguiente resultado es muy útil y nos dice la forma de calcular esta esperanza conociendo únicamente la función de densidad de X . A este resultado a veces se le refiere como el teorema del estadístico inconsciente.

Teorema 4.4 Sea X una variable aleatoria continua y sea $g : \mathbb{R} \rightarrow \mathbb{R}$ una función tal que $g(X)$ es una variable con esperanza finita. Entonces:

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

Ejemplo. Calculemos $\mathbb{E}[Y]$ donde $Y = X^2$ y X es la variable aleatoria continua con función de densidad dada por:

$$f(x) = \begin{cases} 2x & \text{si } 0 < x < 1 \\ 0 & \text{e.c.o.c.} \end{cases}$$

Por la proposición anterior tenemos que:

$$\mathbb{E}[Y] = \mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 2x^3 dx = 1/2$$

4.3. Varianza y covarianza

La varianza cuantifica la dispersión de una función de densidad.

Definición 4.2 Sea una variable aleatoria con media μ . La varianza de X , denotada por σ^2 o σ_X^2 , está definida como:

$$Var(X) = \sigma^2 = \mathbb{E}[(X - \mu)^2] = \begin{cases} \int_{-\infty}^{\infty} (X - \mu)^2 f(x) dx \\ \sum_x (X - \mu)^2 f(x) \end{cases}$$

Ejemplo. Calcularemos la varianza de la variable aleatoria discreta X con función de densidad dada por la siguiente tabla.

x	$f(x)$
-1	1/8
0	4/8
1	1/8
2	2/8

Recordemos primeramente que por cálculos previos, $\mu = 1/2$. Aplicando

la definición de varianza tenemos que:

$$\begin{aligned} \text{Var}(X) &= \sum_x (x - \mu)^2 f(x) \\ &= (-1 - 1/2)^2 1/8 + (0 - 1/2)^2 4/8 \\ &\quad + (1 - 1/2)^2 1/8 + (2 - 1/2)^2 2/8 \\ &= 1 \end{aligned}$$

Ejemplo. Calcularemos la varianza de la variable aleatoria continua X con función de densidad

$$f(x) = \begin{cases} 2x & \text{si } 0 < x < 1 \\ 0 & \text{e.c.o.c} \end{cases}$$

En un cálculo previo habíamos encontrado que $\mu = 2/3$. Por lo tanto,

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_0^1 (x - \mu)^2 2x dx = 1/18$$

Teorema 4.5 *Asumiendo que la varianza está bien definida, esta tendrá las siguientes propiedades.*

1. $\text{Var}(X) \geq 0$
2. $\text{Var}(c) = 0$
3. Si a y b son constantes, entonces $\text{Var}(aX + b) = a^2 \text{Var}(X)$
4. $\text{Var}(X) = \mathbb{E}[X^2] - \mu^2$
5. En general, $\text{Var}(X+Y) \neq \text{Var}(X) + \text{Var}(Y)$ Podemos dar un ejemplo. Tomemos $X = Y$, entonces,

$$\text{Var}(2X) \neq 2\text{Var}(X)$$

6. Si X_1, \dots, X_n son independientes y a_1, \dots, a_n son constantes, entonces,

$$\text{Var} \left(\sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$$

Demostración.

1. $Var(X) \geq 0$

Este es evidente a partir de la definición de varianza pues en ella aparece una suma o integral de términos no negativos.

2. $Var(c) = 0$

La constante c es una v.a. con un único valor, de modo que $\mathbb{E}[c] = c$ y, entonces, $Var(X) = \mathbb{E}[(c - c)^2] = 0$.

3. Si a y b son constantes, entonces $Var(aX + b) = a^2 Var(X)$

$$\begin{aligned} Var(aX + b) &= \mathbb{E}[(aX + b) - \mathbb{E}[aX + b]]^2 \\ &= \mathbb{E}[(aX - a\mathbb{E}[X])^2] \\ &= a^2 \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= a^2 Var(X) \end{aligned}$$

4. $Var(X) = \mathbb{E}[X^2] - \mu^2$

$$\begin{aligned} Var(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mu + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2 \end{aligned}$$

5. En general, $Var(X + Y) \neq Var(X) + Var(Y)$

6. Si X_1, \dots, X_n son independientes y a_1, \dots, a_n son constantes, entonces,

$$Var\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 Var(X_i)$$

Para demostrarla, hagamos lo siguiente:

$$\begin{aligned}
Var\left(\sum_{i=1}^n a_i X_i\right) &= \mathbb{E}\left[\left(\sum_{i=1}^n a_i X_i\right)^2\right] - \left(\mathbb{E}\left[\sum_{i=1}^n a_i X_i\right]\right)^2 \\
&= \mathbb{E}\left[a_i^2 \left(\sum_{i=1}^n X_i\right)^2\right] - \left(a_i \sum_{i=1}^n \mathbb{E}[X_i]\right)^2 \\
&= a_i^2 \mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] - a_i^2 n^2 \mu^2 \\
&= a_i^2 \mathbb{E}[(X_1 + X_2 + \dots + X_n)(X_1 + X_2 + \dots + X_n)] \\
&\quad - a_i^2 n^2 \mu^2 \\
&= a_i^2 \mathbb{E}[X_1^2 + X_1 X_2 + \dots + X_1 X_n + \dots + \\
&\quad X_n X_1 + X_n X_2 + \dots + X_n^2] - a_i^2 n^2 \mu^2 \\
&= a_i^2 \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \dots + \mathbb{E}[X_n^2] + \mathbb{E}[X_1 X_2] + \\
&\quad \dots + \mathbb{E}[X_1 X_n] \\
&\quad + \dots + \mathbb{E}[X_{n-1} X_n] - n^2 \mu^2 \\
&= a_i^2 (\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \dots + \mathbb{E}[X_n^2]) + \mu^2 + \\
&\quad \dots + a_i^2 \mu^2 + \dots + a_i^2 \mu^2 - a_i^2 n^2 \mu^2 \\
&= a_i^2 (\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \dots + \mathbb{E}[X_n^2]) \\
&\quad + a_i^2 n(n-1) \mu^2 - a_i^2 n^2 \mu^2 \\
&= a_i^2 (\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] + \dots + \mathbb{E}[X_n^2]) - n a_i^2 \mu^2 \\
&= \mathbb{E}[X_1^2] - \mu^2 + \mathbb{E}[X_2^2] - \mu^2 + \dots + \mathbb{E}[X_n^2] - \mu^2 \\
&= \sum_{i=1}^n a_i^2 Var(X_i) [Tiene errores]
\end{aligned}$$

Para la demostración anterior hemos utilizado que:

$$Cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mu_i \mu_j = 0$$

Por lo tanto: $\mathbb{E}[X_i X_j] = \mu_i \mu_j$, para todo $i \neq j$.

Obs.: La desviación estándar de una v.a. estará definida como $\sigma = \sqrt{Var(X)}$

Ejemplo. Sea Y una v.a. descrita por la PDF:

$$f_Y(y) = \begin{cases} 2y & \text{si } 0 \leq y \leq 1 \\ 0 & \text{e.c.o.c.} \end{cases} \quad (4.4)$$

¿Cuál es la desviación estándar de $3Y + 2$?

Primero:

$$\mathbb{E}[Y] = \int_0^1 y \cdot 2y dy = \frac{2}{3}$$

$$\mathbb{E}[Y^2] = \int_0^1 y^2 \cdot 2y dy = \frac{1}{2}$$

Segundo:

$$Var(Y) = \mathbb{E}[Y^2] - \mu^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$$

De esta forma,

$$Var(3Y + 2) = 3^2 Var(Y) = \frac{1}{2}$$

Ejemplo. Sea $X \sim \text{Binomial}(n, p)$. Podmeos escribir a $X = \sum_i X_i$, donde $X_i \sim \text{Bernoulli}(p)$.

Entonces:

$$\begin{aligned} \mathbb{P}(X_i = 1) &= p \\ \mathbb{P}(X_i = 0) &= 1 - p \end{aligned}$$

De esta forma,

$$\begin{aligned} \mathbb{E}[X_i] &= p \times 1 + (1 - p) \times 0 = p \\ \mathbb{E}[X_i^2] &= p \times 1^2 + (1 - p) \times 0^2 = p \\ Var(X_i) &= \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 \\ &= p - p^2 \\ &= p(1 - p) \end{aligned}$$

Finalmente,

$$\begin{aligned}
 Var(X) &= Var\left(\sum_i X_i\right) \\
 &= \sum_i Var(X_i) \\
 &= \sum_i p(1-p) \\
 &= np(1-p)
 \end{aligned}$$

Si X_1, \dots, X_n son variables aleatorias, podemos definir la **media muestral** como:

$$\bar{X} = \frac{1}{n} \sum_i^n X_i \quad (4.5)$$

Y a la **varianza muestral**:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (4.6)$$

Teorema 4.6 Sea X_1, \dots, X_n variables aleatorias IID y sea $\mu = \mathbb{E}(X_i)$, $\sigma^2 = Var(X_i)$. Entonces,

$$\begin{aligned}
 \mathbb{E}[\bar{X}] &= \mu \\
 Var(\bar{X}) &= \frac{\sigma^2}{n} \\
 \mathbb{E}(S^2) &= \sigma^2
 \end{aligned}$$

Si X y Y son variables aleatorias, entonces la correlación entre X y Y cuantifica que tan fuerte es la relación lineal entre X y Y .

Definición 4.3 Covarianza. Sean X y Y variables aleatorias con medias μ_X y μ_Y , y desviaciones estándar σ_X y σ_Y , respectivamente. La covarianza entre X y Y estará dada por:

$$Cov(X, Y) = \mathbb{E}((X - \mathbb{E}[X])(Y - \mathbb{E}[Y])) \quad (4.7)$$

La correlación:

$$\rho = \rho_{X,Y} = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad (4.8)$$

Teorema 4.7 *La covarianza satisface:*

$$Cov(X, Y) = \mathbb{E}[X, Y] - \mathbb{E}[X]\mathbb{E}[Y]$$

La correlación satisface:

$$-1 \leq \rho(X, Y) \leq 1$$

Si $Y = aX + b$ para algunas constantes a y b , entonces $\rho(X, Y) = 1$ si $a > 0$ y $\rho(X, Y) = -1$ si $a < 0$. Si X y Y son independientes, entonces $Cov(X, Y) = \rho = 0$. El caso inverso no es necesariamente cierto.

Demostración.

$$\begin{aligned} Cov(X, y) &= \mathbb{E}((X - \mathbb{E}[X])(Y - \mathbb{E}[Y])) \\ &= \mathbb{E}(XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]) \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

$$\begin{aligned} \rho(X, Y) &= \frac{Cov(X, y)}{\sigma_X \sigma_Y} \\ &= \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2} \sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}} \\ &= \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{(\mathbb{E}[X^2] - (\mathbb{E}[X])^2)(\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2)}} \end{aligned}$$

Teorema 4.8 *Sean X y Y , variables aleatorias, entonces:*

$$\begin{aligned} Var(X + Y) &= Var(X) + Var(Y) + 2Cov(X, Y) \\ Var(X - Y) &= Var(X) + Var(Y) - 2Cov(X, Y) \end{aligned}$$

En general, para cualesquiera variables aleatorias X_1, X_2, \dots, X_n ,

$$Var\left(\sum_i a_i X_i\right) = \sum_i a_i^2 Var(X_i) + 2 \sum_{i < j} a_i a_j Cov(X_i, X_j)$$

Demostración.

$$\begin{aligned}
Var(X + Y) &= \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2 \\
&= \mathbb{E}[X^2 + Y^2 + 2XY] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] - (\mathbb{E}[X])^2 - (\mathbb{E}[Y])^2 - 2\mathbb{E}[X]\mathbb{E}[Y] \\
&= (\mathbb{E}[X^2] - (\mathbb{E}[X])^2) + (\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2) \\
&\quad + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \\
&= Var(X) + Var(Y) + 2Cov(X, Y)
\end{aligned}$$

El resto se deja para el lector.

4.4. Esperanza condicional

Supongamos que X e Y son variables aleatorias. ¿Cuál es la media de X entre los momentos en que $Y = y$?

Definición 4.4 *Esperanza condicional.* La esperanza condicional de X dado $Y = y$ está dada por:

$$\mathbb{E}[X|Y = y] = \begin{cases} \sum x f_{X|Y}(x|y) & \text{para el caso discreto} \\ \int x f_{X|Y}(x|y) dx & \text{para el caso continuo} \end{cases} \quad (4.9)$$

Si $g(x, y)$ es una función de x y y , entonces:

$$\mathbb{E}[g(X, Y)|Y = y] = \begin{cases} \sum g(x, y) f_{X|Y}(x|y) & \text{para el caso discreto} \\ \int g(x, y) f_{X|Y}(x|y) dx & \text{para el caso continuo} \end{cases} \quad (4.10)$$

Ejemplo. Supongamos $X \sim Uniforme(0, 1)$. Una vez que hemos observado a $X = x$, podemos obtener $Y|X = x \sim Uniforme(x, 1)$. De esta forma, consideremos que:

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x} & \text{si } x < y < 1 \\ 0 & \text{e.c.o.c} \end{cases}$$

Entonces, determinemos $\mathbb{E}[Y|X = x]$.

$$\begin{aligned}\mathbb{E}[Y|X = x] &= \int_x^1 y f_{Y|X}(y|x) dy \\ &= \int_x^1 y \frac{1}{1-x} dy \\ &= \frac{1+x}{2}\end{aligned}$$

Teorema 4.9 Regla de la Esperanza Iterada. Para las variables aleatorias X y Y , asumiendo que los valores esperados existen, tenemos que:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y|X]] &= \mathbb{E}[Y] \\ \mathbb{E}[\mathbb{E}[X|Y]] &= \mathbb{E}[X]\end{aligned}$$

En general, para cualquier función $g(x, y)$,

$$\mathbb{E}[\mathbb{E}[g(X, Y)|X]] = \mathbb{E}[g(x, y)]$$

Demostración. Recordemos que:

$$\begin{aligned}f_{Y|X}(y|x) &= \frac{f(x, y)}{f_X(x)} \\ f(x, y) &= f_X(x) \cdot f_{Y|X}(y|x)\end{aligned}$$

Entonces,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y|X]] &= \int \mathbb{E}[Y|X = x] f_X(x) dx \\ &= \int \int y f_{Y|X}(y|x) dy f_X(x) dx \\ &= \int \int y f_{Y|X}(y|x) f_X(x) dx dy \\ &= \int \int y f(x, y) dx dy \\ &= \int y \int f(x, y) dx dy \\ &= \int y f_Y(y) dy \\ &= \mathbb{E}[Y]\end{aligned}$$

El caso de $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ es inmediato. El caso restante es un poco más complicado.

Ejemplo. Retomemos el ejemplo en el que $Y|X = x \sim \text{Uniforme}(x, 1)$. Supongamos $X \sim \text{Uniforme}(0, 1)$. De esta forma, consideremos que:

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x} & \text{si } x < y < 1 \\ 0 & \text{e.c.o.c} \end{cases}$$

Entonces, determinemos $\mathbb{E}[Y|X = x]$.

$$\begin{aligned} \mathbb{E}[Y|X = x] &= \int_x^1 y f_{Y|X}(y|x) dy \\ &= \int_x^1 y \frac{1}{1-x} dy \\ &= \frac{y^2}{2} \Big|_x^1 \frac{1}{1-x} \\ &= \frac{1}{2} \frac{1-x^2}{1-x} \\ &= \frac{1}{2} \frac{(1-x)(1+x)}{1-x} \\ &= \frac{1+x}{2} \end{aligned}$$

Ahora, determinemos: $\mathbb{E}[Y]$. Usando el $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$, tenemos:

$$\begin{aligned} \mathbb{E}[\mathbb{E}[Y|X = x]] &= \mathbb{E}\left[\frac{1+X}{2}\right] \\ &= \frac{1 + \mathbb{E}[X]}{2} \\ &= \frac{1 + \frac{1}{2}}{2} \\ &= \frac{3}{4} \end{aligned}$$

Donde:

$$\mathbb{E}[X] = \int_0^1 x f(x) dx = \int_0^1 x \frac{1}{b-a} dx = \int_0^1 x dx = \frac{x^2}{2} \Big|_0^1 = \frac{1}{2}$$

4.5. Funciones generadoras de momentos

4.5.1. Momentos

Definimos el n -ésimo momento de una variable aleatoria X , cuando existe, como el número $\mathbb{E}[X^n]$ para cada $n = 1, 2, \dots$ suponiendo que tal esperanza exista. Así, los momentos de una variable aleatoria X son la colección de números:

$\mathbb{E}[X]$	Primer Momento
$\mathbb{E}[X^2]$	Segundo Momento
$\mathbb{E}[X^3]$	Tercer Momento
\vdots	
$\mathbb{E}[X^n]$	n -ésimo Momento

Para variables aleatorias discretas el n -ésimo momento se calcula como sigue:

$$\mathbb{E}[X^n] = \sum_{i=0}^{\infty} x_i^n f(x_i)$$

Mientras que para variables aleatorias continuas la fórmula es la siguiente:

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

Observe que el primer momento es simplemente la esperanza de la variable aleatoria y recordando la fórmula $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, puede decirse ahora que la varianza es el segundo momento menos el primer momento al cuadrado.

Existen interpretaciones conocidas para los primeros momentos pero es difícil encontrar alguna interpretación para todos ellos. Los momentos indican, sin embargo, alguna característica numérica de la variable aleatoria: el primer momento es el valor promedio, el segundo momento está relacionado con la dispersión de los valores de la variable aleatoria, el tercer momento está relacionado con la simetría de la correspondiente función de densidad.

Ejemplo. Considere una variable aleatoria continua con función de den-

sidad:

$$f(x) = \begin{cases} x+1 & \text{si } -1 < x < 0 \\ 1-x & \text{si } 0 \leq x < 1 \\ 0 & \text{e.c.o.c.} \end{cases}$$

Determinemos el primer y segundo momento.

$$\begin{aligned} \mathbb{E}[X] &= \int_{\mathbb{D}} x f(x) dx \\ &= \int_{-1}^0 x(x+1) dx + \int_0^1 x(1-x) dx \\ &= \int_{-1}^0 (x^2 + x) dx + \int_0^1 (x - x^2) dx \\ &= \left. \frac{x^3}{3} \right|_{-1}^0 + \left. \frac{x^2}{2} \right|_{-1}^0 + \left. \frac{x^2}{2} \right|_0^1 - \left. \frac{x^3}{3} \right|_0^1 \\ &= \frac{1}{3} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} \\ &= 0 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X^2] &= \int_{\mathbb{D}} x^2 f(x) dx \\ &= \int_{-1}^0 x^2(x+1) dx + \int_0^1 x^2(1-x) dx \\ &= \int_{-1}^0 (x^3 + x^2) dx + \int_0^1 (x^2 - x^3) dx \\ &= \left. \frac{x^4}{4} \right|_{-1}^0 + \left. \frac{x^3}{3} \right|_{-1}^0 + \left. \frac{x^3}{3} \right|_0^1 - \left. \frac{x^4}{4} \right|_0^1 \\ &= -\frac{1}{4} + \frac{1}{3} + \frac{1}{3} - \frac{1}{4} \\ &= \frac{2}{3} \end{aligned}$$

Ejemplo. El n -ésimo momento de una variable aleatoria continua X con función de densidad:

$$f(x) = \begin{cases} \frac{x}{2} & \text{si } 0 < x < 2 \\ 0 & \text{e.c.o.c.} \end{cases}$$

Será:

$$\begin{aligned}
\mathbb{E}[X^n] &= \int_{\mathbb{D}} x^n f(x) dx \\
&= \int_0^2 x^n \frac{x}{2} dx \\
&= \int_0^2 \frac{x^{n+1}}{2} dx \\
&= \left. \frac{x^{n+2}}{2(n+2)} \right|_0^2 \\
&= \frac{2^{n+1}}{n+2}
\end{aligned}$$

Ahora, definamos también el n -ésimo momento central de X , cuando existe, como el número:

$$\mathbb{E}[(X - \mathbb{E}[X])^n], \text{ para } n = 1, 2, \dots$$

Para variables aleatorias discretas el n -ésimo momento central se calcula como sigue:

$$\mathbb{E}[(X - \mathbb{E}[X])^n] = \sum_{i=0}^{\infty} (X_i - \mathbb{E}[X])^n f(x_i)$$

Mientras que para variables aleatorias continuas la fórmula es la siguiente:

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

4.5.2. Funciones Generadoras de Momentos

Ahora definiremos la función generadora de momentos que se utiliza para encontrar momentos, para encontrar la distribución de sumas de variables aleatorias y que también se utiliza en las demostraciones de algunos teoremas.

Definición 4.5 *Función Generadora de Momentos.* La función generadora de momentos, FGM o transformada de Laplace, de una variable aleatoria X está definida por:

$$M_X(t) = \mathbb{E}[e^{tX}] = \begin{cases} \sum_x e^{tX} f(x) dx & \text{caso discreto} \\ \int_{\mathbb{D}} e^{tX} f(x) dx & \text{caso continuo} \end{cases} \quad (4.11)$$

Donde t varia sobre los numeros reales y para los cuales el valor esperado existe.

Cuando la función generadora de momentos está bien definida, lo siguiente es cierto:

$$\begin{aligned} M'_X(0) &= \left[\frac{d}{dt} \mathbb{E} [e^{tX}] \right]_{t=0} \\ &= \mathbb{E} \left[\frac{d}{dt} e^{tX} \right]_{t=0} \\ &= \mathbb{E} [X e^{tX}]_{t=0} \\ &= \mathbb{E}[X] \end{aligned}$$

Al tomar k derivadas concluimos que:

$$\begin{aligned} M_X^{(k)}(0) &= \left[\frac{d^k}{dt^k} \mathbb{E} [e^{tX}] \right]_{t=0} \\ &= \mathbb{E} \left[\frac{d^k}{dt^k} e^{tX} \right]_{t=0} \\ &= \mathbb{E} [X^k e^{tX}]_{t=0} \\ &= \mathbb{E}[X^k] \end{aligned}$$

Esto nos proporciona un método para calcular los momentos de una distribución.

Ejemplo. Supongamos una variable aleatoria X con función de densidad geométrica:

$$f(x) = \begin{cases} (1-p)^{x-1}p & \text{si } x = 1, 2, \dots \\ 0 & \text{e.c.o.c.} \end{cases}$$

En la práctica, esta es la función de densidad de probabilidad que modela la ocurrencia del primer éxito en una serie de ensayos independientes, donde cada ensayo tiene una probabilidad p de terminar en éxito.

Encontremos la FGM. Entonces,

$$\begin{aligned}
 M_X(t) &= \mathbb{E}[e^{tX}] \\
 &= \sum_{x=1}^{\infty} e^{tx} (1-p)^{x-1} p \\
 &= \frac{p}{1-p} \sum_{x=1}^{\infty} e^{tx} (1-p)^x \\
 &= \frac{p}{1-p} \sum_{x=1}^{\infty} (e^t(1-p))^x
 \end{aligned}$$

Notemos que los valores de t para los que es convergente la suma, $M_X(t) < \infty$, son aquellos en los que:

$$\begin{aligned}
 e^t(1-p) &< 1 \\
 e^t &< \frac{1}{1-p} \\
 t &< \ln\left(\frac{1}{1-p}\right)
 \end{aligned}$$

De hecho:

$$0 < t < \ln\left(\frac{1}{1-p}\right)$$

Recordemos el caso de la suma:

$$S_n = \sum_{k=0}^n r^k \tag{4.12}$$

Por lo tanto, podemos definir S_{n+1} que estaría dado por la siguiente expresión:

$$\begin{aligned}
 S_{n+1} &= r \sum_{k=0}^n r^k \\
 &= r(1 + r + r^2 + \dots + r^n) \\
 &= r + r^2 + r^3 + \dots + r^{n+1} \\
 &= rS_n
 \end{aligned} \tag{4.13}$$

Tomando los dos resultados de las ecuaciones (4.12) y (4.13) anteriores, podemos expresar que si a S_n le restamos S_{n+1} , y desarrollando ambos lados de la ecuación anterior podemos obtener:

$$\begin{aligned} S_n - rS_n &= (1 - r)S_n \\ &= (1 + r + r^2 + \dots + r^n) - (r + r^2 + r^3 + \dots + r^{n+1}) \\ (1 - r)S_n &= 1 - r^{n+1} \end{aligned}$$

Así, podemos concluir que:

$$S_n = \frac{1 - r^{n+1}}{1 - r} \quad (4.14)$$

Podemos calcular:

$$\lim_{n \rightarrow \infty} S_n = \frac{1}{1 - r}$$

Esto, siempre que $|r| < 1$

Encontremos la FGM. Entonces,

$$\begin{aligned} M_X(t) &= \frac{p}{1 - p} \sum_{x=1}^{\infty} (e^t(1 - p))^x \\ &= \frac{p}{1 - p} \left[\sum_{x=0}^{\infty} (e^t(1 - p))^x - (e^t(1 - p))^0 \right] \\ &= \frac{p}{1 - p} \left[\frac{1}{1 - (1 - p)e^t} - 1 \right] \\ &= \frac{pe^t}{1 - (1 - p)e^t} \end{aligned}$$

Determinemos el valor esperado de X :

$$\begin{aligned} M_X^{(1)}(t) &= \frac{d}{dt} \frac{pe^t}{1 - (1 - p)e^t} \\ M_X^{(1)}(0) &= \frac{1}{p} \end{aligned}$$

Ejemplo. Suponga que Y tiene un función de densidad exponencial:

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y} & \text{si } y > 0 \\ 0 & \text{e.c.o.c.} \end{cases}$$

Determinemos $M_Y(t)$.

$$\begin{aligned}
 M_Y(t) &= \mathbb{E} [e^{ty}] \\
 &= \int_0^\infty e^{ty} \lambda e^{-\lambda y} dy \\
 &= \int_0^\infty \lambda e^{-(\lambda-t)y} dy \\
 &= \int_0^\infty \frac{\lambda}{\lambda-t} e^{-u} du \\
 &= \frac{\lambda}{\lambda-t} [-e^{-u}]_0^\infty \\
 &= \frac{\lambda}{\lambda-t} \left[-\lim_{u \rightarrow \infty} e^{-u} + 1 \right] \\
 &= \frac{\lambda}{\lambda-t}
 \end{aligned}$$

Esto es válido para $t \geq \lambda$.

Determinemos el valor esperado de Y :

$$\begin{aligned}
 M_Y^{(1)}(t) &= \frac{d}{dt} \frac{\lambda}{\lambda-t} \\
 M_Y^{(1)}(0) &= \frac{1}{\lambda}
 \end{aligned}$$

Finalmente,

Teorema 4.10 Sean X y Y variables aleatorias independientes, y cuyas FGM existen para una vecindad no trivial al rededor del cero. Entonces para cualquier $t \in (-s, s)$, para algún $s > 0$,

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

Demostración.

$$\begin{aligned}
 M_{X+Y}(t) &= \mathbb{E} [e^{X+Y}] \\
 &= \mathbb{E} [e^X e^Y] \\
 &= \mathbb{E} [e^X] \mathbb{E} [e^Y] \\
 &= M_X(t) M_Y(t)
 \end{aligned}$$

5

Desigualdades y algunos resultados importantes

5.1. Desigualdades

Las desigualdades son útiles para limitar cantidades que de otro modo serían difíciles de calcular. También se utilizarán en la teoría de convergencia que analizamos más adelante. Nuestra primera desigualdad es la desigualdad de Markov.

Teorema 5.1 *Desigualdad de Markov.* *Sea X una variable aleatoria no negativa y supongamos que $\mathbb{E}[X]$ existe. Para cualquier $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Demostración. Dado $X > 0$,

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty xf(x)dx \\ &= \int_0^t xf(x)dx + \int_t^\infty xf(x)dx \\ &\geq \int_t^\infty xf(x)dx \\ &\geq t \int_t^\infty f(x)dx \\ &= t\mathbb{P}(X \geq t)\end{aligned}$$

En palabras, este resultado establece que la probabilidad de que X exceda un valor t positivo está acotada superiormente por la media entre t . La siguiente desigualdad será usada en la siguiente sección para demostrar la ley débil de los grandes números.

Teorema 5.2 *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 $t > 0$:

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

O equivalentemente:

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

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: $\sigma^2 = \text{Var}(X)$ y que

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

Usando los resultados anteriores:

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

De esta forma:

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

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

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

Ejemplo. Sea X una variable aleatoria con función de densidad dada por:

$$f(x) = \begin{cases} 2e^{-2x} & \text{si } x \geq 0 \\ 0 & \text{e.c.o.c.} \end{cases}$$

Determinemos $\mathbb{P}(|X - \mu| > 1)$.

La función generadora de momentos está dada por:

$$\begin{aligned}
 M_X(t) &= \mathbb{E}[e^{tX}] \\
 &= \int_{-\infty}^{\infty} e^{tx} f(x) dx \\
 &= \int_0^{\infty} e^{tx} 2e^{-2x} dx \\
 &= 2 \int_0^{\infty} e^{(t-2)x} dx \\
 &= \left. \frac{2e^{(t-2)x}}{t-2} \right|_0^{\infty} \\
 &= \frac{2}{2-t}
 \end{aligned}$$

Siempre que $t < 2$.

De esta forma, podemos obtener:

$$\begin{aligned}
 M'_X(0) &= -\left. \frac{2(-1)}{(2-t)^2} \right|_{t=0} \\
 &= \frac{1}{2}
 \end{aligned}$$

De esta forma, partamos de:

$$\begin{aligned}
 1 &= \mathbb{P}(|X - \mu| \leq 1) + \mathbb{P}(|X - \mu| > 1) \\
 \mathbb{P}(|X - \mu| > 1) &= 1 - \mathbb{P}(|X - \mu| \leq 1)
 \end{aligned}$$

Así,

$$\begin{aligned}
 \mathbb{P}(|X - \mu| \leq 1) &= \mathbb{P}(|X - \mu| \leq 1) \\
 &= \mathbb{P}(-1 \leq X - \mu \leq 1) \\
 &= \mathbb{P}\left(-1 + \frac{1}{2} \leq X \leq 1 + \frac{1}{2}\right) \\
 &= \mathbb{P}\left(-\frac{1}{2} \leq X \leq \frac{3}{2}\right) \\
 &= \int_0^{3/2} 2e^{-2x} dx \\
 &= 1 - e^{-3}
 \end{aligned}$$

En suma:

$$\begin{aligned}\mathbb{P}(|X - \mu| > 1) &= 1 - \mathbb{P}(|X - \mu| \leq 1) \\ &= 1 - (1 - e^{-3}) \\ &= e^{-3} \\ &= 0.04979\end{aligned}$$

5.2. Tipos de convergencia

El aspecto más importante de la teoría de la probabilidad se refiere al comportamiento de las sucesiones de variables aleatorias. Esta parte de la probabilidad se denomina teoría de muestras grandes, teoría del límite o teoría asintótica. La pregunta básica es la siguiente: ¿qué podemos decir sobre el comportamiento límite de una secuencia de variables aleatorias X_1, X_2, X_3, \dots ? Dado que las estadísticas y la minería de datos se basan en la recopilación de datos, naturalmente nos interesará lo que sucede a medida que recopilamos más y más datos.

En cálculo decimos que una secuencia de números reales x_n converge a un límite x si, para cada $\epsilon > 0$, $|x_n - x| < \epsilon$ para todo n grande. En probabilidad, la convergencia es más sutil.

Volviendo al cálculo por un momento, supongamos que $x_n = x$ para todo n . Entonces, trivialmente,

$$\lim_{n \rightarrow \infty} x_n = x$$

Consideremos una versión probabilística de este ejemplo. Supongamos que X_1, X_2, \dots es una secuencia de variables aleatorias que son independientes y supongamos que cada una tiene una distribución $N(0, 1)$. Como todas tienen la misma distribución, nos sentimos tentados a decir que X_n “converge” a $X \sim N(0, 1)$. Pero esto no puede ser del todo correcto ya que $\mathbb{P}(X_n = X) = 0$ para todo n (Dos variables aleatorias continuas son iguales con probabilidad cero).

A continuación, se presenta otro ejemplo. Consideremos X_1, X_2, \dots donde $X_i \sim N(0, 1/n)$. Intuitivamente, X_n está muy concentrado alrededor de 0 para un valor n grande, por lo que nos gustaría decir que X_n converge a 0. Pero $\mathbb{P}(X_n = 0) = 0$ para todo n . Claramente, necesitamos desarrollar algunas herramientas para discutir la convergencia de manera rigurosa.

Planteamos dos ideas principales que enunciamos de manera informal:

1. La **ley de los grandes números** dice que la media muestral:

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

converge en probabilidad al valor esperado $\mu = \mathbb{E}[X_i]$.

Esto significa que X_n está cerca de μ con alta probabilidad.

2. El **teorema del límite central** dice que:

$$\sqrt{n}(\bar{X}_n - \mu)$$

converge en distribución a una distribución Normal.

Esto significa que la media muestral tiene aproximadamente una distribución Normal para n grande.

5.2.1. Tipos de convergencia

Definición 5.1 Sea X_1, X_2, \dots una sucesión de variables aleatorias y sea X otra variable aleatoria. Sea F_n la función distribución de probabilidad (CDF) de X_n y sea F la CDF de X . Entonces,

1. X_n converge a X en **probabilidad**, $X_n \xrightarrow{P} X$, si para cada $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \rightarrow 0$$

cuando $n \rightarrow \infty$.

2. X_n converge a X en **distribución**, $X_n \xrightarrow{D} X$, si

$$\lim_{n \rightarrow \infty} F_n(t) = F(t)$$

para todo t en el que F sea continua.

5.2.2. La Ley de los Grandes Números

Ahora llegamos a un logro supremo en probabilidad, la ley de los grandes números. Este teorema dice que la media de una muestra grande está cerca de la media de la distribución. Por ejemplo, se espera que la proporción de caras de una gran cantidad de lanzamientos sea cercana a $1/2$. Ahora lo haremos más preciso.

Teorema 5.3 *La Ley Débil de los Grandes Números.* Si X_1, X_2, \dots son IID, entonces

$$\bar{X}_n \xrightarrow{P} \mu$$

Interpretación de la Ley Débil de los Grandes Números: La distribución de X_n se vuelve más concentrada alrededor de μ a medida que n , se hace grande.

Demostración. Supongamos que $\sigma < \infty$. Esto no es necesario pero simplifica la prueba. Recondando la definición de convergencia en probabilidad, con lo que X_n , o \bar{X}_n en este caso, converge a μ en probabilidad, $\bar{X}_n \xrightarrow{P} \mu$, si para cada $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \rightarrow 0$$

cuando $n \rightarrow \infty$.

Usando la 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 $\epsilon > 0$:

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

O equivalentemente:

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

Utilicemos la segunda expresión. De esta forma, si consideramos que \bar{X}_n es una variable aleatoria:

$$\begin{aligned} \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) &\leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \\ \lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) &\leq \lim_{n \rightarrow \infty} \frac{\sigma^2}{n\epsilon^2} \\ &= 0 \end{aligned}$$

Teorema 5.4 *La Ley Fuerte de los Grandes Números.* Sea X_1, X_2, \dots IID. Si $\mu = \mathbb{E}[|X_1|] < \infty$ entonces

$$\bar{X}_n \xrightarrow{CS} \mu$$

Ejemplo. Consideremos el lanzamiento de una moneda con probabilidad de éxito p . Sea X_i el resultado del lanzamiento (0 o 1). Partamos de que $p = \mathbb{P}(X_i = 1) = \mathbb{E}[X_i]$, así, la fracción de casos de éxito después de n lanzamientos es \bar{X}_n .

De acuerdo con la Ley de los Grandes Números, \bar{X}_n converge a p en probabilidad. Esto no significa que \bar{X}_n sea numéricamente igual a p . Significa que, cuando n es grande, la distribución de \bar{X}_n esta estrechamente concentrada al rededor de p .

Supongamos que $p = 1/2$, ¿que tan grande tiene que ser n para que $\mathbb{P}(0.4 \leq \bar{X}_n \leq 0.6) \geq 0.7$?

Partamos de que:

$$\begin{aligned}\mathbb{E}[\bar{X}_n] &= p = \frac{1}{2} \\ \text{Var}(\bar{X}_n) &= \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = \frac{1}{4n}\end{aligned}$$

Recordemos la desigualdad de Chebyshev. Así, para cualquier $\epsilon > 0$:

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

O equivalentemente:

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

Tomemos el primer caso y asumamos un ϵ particular:

$$\begin{aligned}\mathbb{P}(0.4 \leq \bar{X}_n \leq 0.6) &= \mathbb{P}(-0.1 \leq \bar{X}_n - 1/2 \leq 0.1) \\ &= \mathbb{P}(-0.1 \leq \bar{X}_n - \mu \leq 0.1) \\ &= \mathbb{P}(|\bar{X}_n - \mu| \leq 0.1) \\ &= \mathbb{P}(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2} \\ &= \mathbb{P}(|\bar{X}_n - \mu| \leq 0.1) = 1 - \mathbb{P}(|\bar{X}_n - \mu| > 0.1) \\ &\geq 1 - \frac{1}{4n(0.1)^2} = 1 - \frac{25}{n}\end{aligned}$$

$$\begin{aligned}
1 - \frac{25}{n} &\geq 0.7 \\
0.3 &\geq \frac{25}{n} \\
n &\geq \frac{25}{0.3} \\
n &\geq 83.3333
\end{aligned}$$

Donde $n = 84$.

5.2.3. El Teorema del Límite Central

La ley de los grandes números dice que la distribución de \bar{X}_n se acumula cerca de μ . Esto no es suficiente para ayudarnos a aproximarnos a los enunciados de probabilidad sobre \bar{X}_n . Para ello necesitamos el Teorema del Límite Central.

Supongamos que X_1, \dots, X_n son IID con media μ y varianza σ^2 . El Teorema del Límite Central (TLC) dice que:

$$\bar{X}_n = \frac{\sum_i X_i}{n}$$

tiene una distribución que es aproximadamente normal con media μ y varianza σ^2/n . Esto es notable ya que no se supone nada sobre la distribución de X_i , excepto la existencia de la media y la varianza.

Teorema 5.5 *El Teorema del Límite Central (TLC)*. Sea X_1, \dots, X_n IID con media μ y varianza σ^2 . Sea:

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

Entonces:

$$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{\text{Var}(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} Z$$

Donde, $Z \sim N(0, 1)$. En otras palabras,

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Interpretación: Las proposiciones de probabilidad sobre \bar{X}_n se pueden aproximar utilizando una distribución normal. Son las proposiciones de probabilidad las que estamos aproximando, no la variable aleatoria en sí.

Además de $Z_n \xrightarrow{D} N(0, 1)$, existen varias formas de notación para indicar que la distribución de Z_n está convergiendo hacia una distribución normal. Todas significan lo mismo. Aquí están:

$$\begin{aligned} Z_n &\approx N(0, 1) \\ \bar{X}_n &\approx N\left(\mu, \frac{\sigma^2}{n}\right) \\ \bar{X}_n - \mu &\approx N\left(0, \frac{\sigma^2}{n}\right) \\ \sqrt{n}(\bar{X}_n - \mu) &\approx N(0, \sigma^2) \\ \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} &\approx N(0, 1) \end{aligned}$$

Ejemplo. Supongamos que el número de errores por programa de ordenador tiene una distribución de Poisson con media 5. Obtenemos 125 programas. Sea X_1, \dots, X_{125} el número de errores en los programas. Queremos aproximar:

$$\mathbb{P}(\bar{X}_n < 5.5)$$

Sea $\mu = \mathbb{E}[X_i] = \lambda = 5$ y sea $\text{Var}(X_i) = \lambda = 5$. Entonces,

$$\begin{aligned} \mathbb{P}(\bar{X}_n < 5.5) &= \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5.5 - \mu)}{\sigma}\right) \\ &= \mathbb{P}(Z < 2.5) = 0.9938 \end{aligned}$$

6

Modelos, Inferencia y Aprendizaje Estadísticos

6.1. 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 $X_1, X_2, \dots, X_n \sim F$, ¿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.

6.2. Modelos paramétricos y No paramétricos

Definición 6.1 *Modelo estadístico.* Un modelo estadístico \mathfrak{F} es un conjunto de distribuciones, densidades o funciones.

Definición 6.2 *Modelo paramétrico.* Un modelo paramétrico es un conjunto \mathfrak{F} que puede ser parametrizado por un número finito de parámetros. En general, un modelo paramétrico toma la forma:

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

Donde θ es un parámetro desconocido que puede tomar valores en el espacio de parámetros Θ .

Definición 6.3 *Modelo no paramétrico.* Un modelo no paramétrico es un conjunto \mathfrak{F} que no puede ser parametrizado por un número finito de parámetros.

A continuación, veamos unos ejemplos:

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

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

6.3. 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 parámetros θ , o $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ – para facilitar la notación utilizaremos indistintamente θ o $\boldsymbol{\theta}$. Así, el problema de inferencia se reduce al problema de estimar el parámetro θ .

A partir de muestras de datos el proceso de estimación de parámetros 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 parámetro.

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

Otra forma de verlo 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 encuentran 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.

7

Estimación Puntual

7.1. Definiciones y ejemplos

La estimación puntual se refiere a la mejor adivinanza del valor del parámetro 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 7.1 *Estadística.* *Una estadística (coloquialmente mencionada como ‘estadístico’) es una función de las variables aleatorias observables en una muestra y de constantes conocidas.*

Definición 7.2 *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 7.3 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 para estimar a θ .*

Más formalmente, decimos que un estimador de θ es:

$$\hat{\theta}_n = g(X_1, X_2, \dots, X_n) \quad (7.1)$$

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 propiedades son: insesgamiento, varianza mínima, eficiencia, consistencia, suficiencia y robustez. 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 aguilas 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 aguilas y S denota sol. Basados en esos resultados, ¿cuál es el valor que consideramos apropiado 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 aguilas 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:

$$p_X(k) = p^k(1-p)^{1-k} = \begin{cases} p & \text{si } k = 1 \\ (1-p) & \text{si } k = 0 \end{cases} \quad (7.2)$$

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

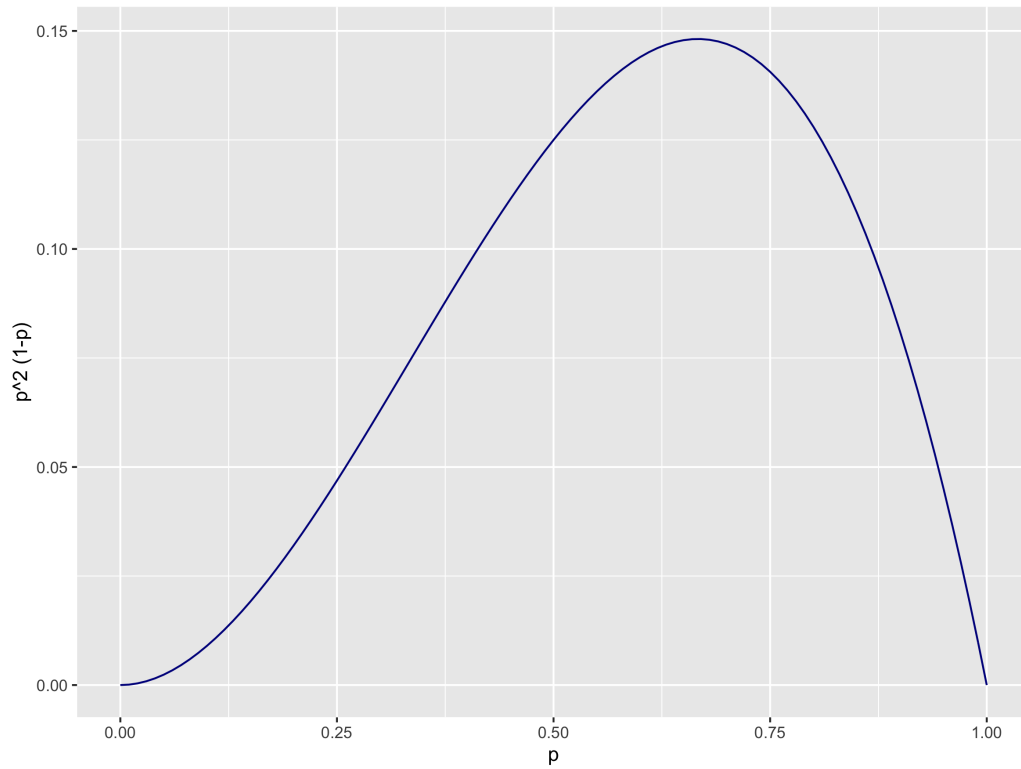


Figura 7.1: Gráfica de $p^2(1-p)$.

miento de una moneda tres veces como:

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

Así, de forma analoga el lanzamiento de una moneda n veces y considerando el resultado en la ecuación (7.2) 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 maximiza la función de probabilidad conjunta $P(X_1 = 1 \cap X_2 = 1 \cap X_3 = 0)$?. La Figura 7.1 muestra la evolución de la probabilidad y el punto en que maximiza la probabilidad.

De forma general y considerando el resultado en la ecuación (7.2), el lanzamiento de una moneda n veces estará dado por la sucesión de v.a.'s:

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

Donde k_i , para $i = 1, 2, \dots, n$ es como definimos en la ecuación (7.2). Entonces, el problema general estará dado por:

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

En este caso, para determinar el valor que maximiza 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 sucesión de v.a.'s dadas por $X_1 = k_1, X_2 = k_2, \dots, X_n = k_n$.

7.2. 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 parámetro y (2) tiene un error estándar lo más pequeño posible.

7.2.1. 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 varianza. La primera de ellas es la propiedad de insesgamiento o, visto desde la otra óptica, la propiedad de sesgo cero. Partamos de las siguientes definiciones.

Definición 7.4 Esperanza de una variable aleatoria. Sea X una v.a.. Si X es una variable 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 7.5 *Estimador Inssegado (1)*. Una estadística $\hat{\theta}$ es un estimador inssegado 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 7.6 *Estimador Inssegado (2)*. Sean $\{X_1, X_2, \dots, X_n\}$ una sucesión 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 $\hat{\theta}(= h(X_1, X_2, \dots, X_n))$ es un estimador inssegado 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 inssegado para muestras grandes.

Definición 7.7 *Estimador Inssegado (3)*. Un estimador es inssegado si el centro de la distribución muestral está al rededor del parámetro.

La Figura 7.2 ilustra el planteamiento de esta última definición en el caso la distribución de la media y su ubicación en la distribución poblacional. A manera de contraste, la Figura permite concluir que un estimados sesgado tenderá a subestimar un parámetro, en el promedio, o tenderá a sobrestimarlo.

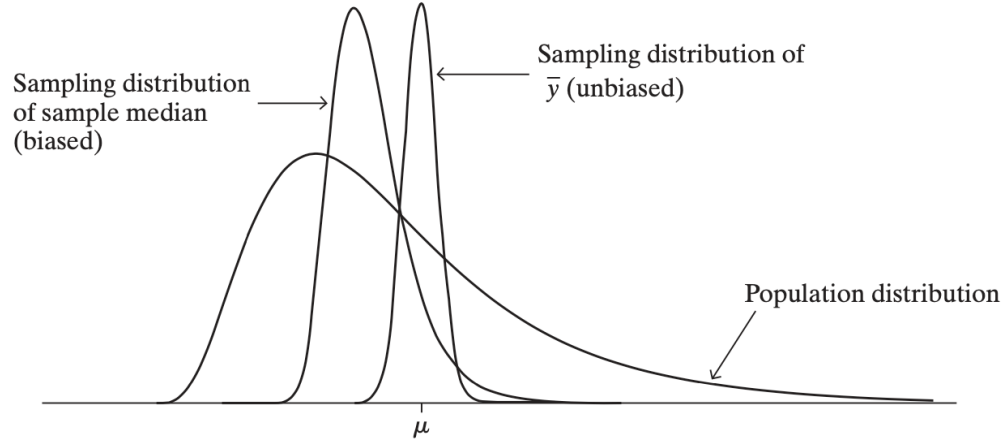


Figura 7.2: Distribución muestral de dos estimadores de la media poblacional para una distribución poblacional sesgada, retomada de Agresti, et. al. (2017; p. 104) Agresti, Franklin y Klingenberg 2017

Definición 7.8 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:

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

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

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

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 $X \sim \text{Bin}(p)$. 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 (7.4) 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 media 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 7.9 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 ejemplo, 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:

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

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 varianza 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 planteado, busquemos 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:

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

Veamos otro ejemplo.

Ejemplo. Sea $X \sim \text{Bin}(p)$ 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 estimador es insesgado. Ahora veamos algunos casos de funciones continuas.

Ejemplo. Sean X_1, X_2, \dots, 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 $\hat{\delta} = \bar{X}$. Determinemos si el estimador es sesgado.

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

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

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:

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

Ejemplo. Sea una muestra aleatoria X_1, X_2, \dots, X_n de una pdf dada por la expresión que describimos más abajo. Supongamos un par de estimadores $\hat{\theta} = 3/2 \cdot \bar{X}$ 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:

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

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} :

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

Pero si tomamos el siguiente límite:

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

Ejemplo. Sean X_1, X_2, \dots, 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, \dots, 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, \dots, 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:

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

Verifiquemos el valor esperado:

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

Donde utilizamos los siguientes resultados:

$$\begin{aligned}
\mathbb{E}\left[X_i \sum_{j=1}^n X_j\right] &= \mathbb{E}[X_i X_1 + \dots + X_i X_i + \dots + X_i X_n] \\
&= \mathbb{E}[X_i X_1] + \dots + \mathbb{E}[X_i^2] + \dots + \mathbb{E}[X_i X_n] \\
&= \mu^2 + \dots + \sigma^2 + \mu^2 + \dots + \mu^2 \\
&= n \cdot \mu^2 + \sigma^2
\end{aligned}$$

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

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:

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

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:

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

De esta forma:

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

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

7.2.2. 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 7.10 *Estimador eficiente.* Decimos 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 parámetro que tratemos de estimar. En suma, un ‘buen’ estimador de un parámetro es insesgado, o es cercano a serlo, y eficiente.

Definición 7.11 *Estimador insesgado de mínima varianza.* El estimador $\hat{\theta}$ 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 7.12 Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados del parámetro θ si:

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2) \quad (7.3)$$

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

Algunos teoremas relevantes.

Teorema 7.1 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:

$$\text{Var}(\hat{\theta}) = \left[n \cdot \mathbb{E} \left[\left(\frac{\partial \ln(f(x))}{\partial \theta} \right)^2 \right] \right]^{-1} \quad (7.4)$$

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

Definición 7.13 Condiciones de regularidad. Sean X_1, X_2, \dots, X_n una muestra aleatoria de una función de densidad $f(\cdot; \theta)$ donde $\theta \in \Theta$. Supongamos que Θ es un subconjunto de la recta real y sea $\hat{\theta}$ un estimador insesgado de θ . Dicho esto, diremos que las siguientes son las condiciones de regularidad para $f(\cdot; \theta)$:

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

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

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

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

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

Teorema 7.2 Desigualdad de Cramér-Rao. Sea $f(x)$ una pdf, supongamos que tenemos un conjunto de observaciones de x para las que $f(x) \neq 0$, que no depende de θ .

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

$$\text{Var}(\hat{\theta}) \geq \left[n \cdot \mathbb{E} \left[\left(\frac{\partial \ln(f(x))}{\partial \theta} \right)^2 \right] \right]^{-1} \quad (7.5)$$

$$= \left[-n \cdot \mathbb{E} \left[\left(\frac{\partial^2 \ln(f(x))}{\partial \theta^2} \right) \right] \right]^{-1} \quad (7.6)$$

A la expresión contenida en el término de la esperanza en ambos teoremas 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:

$$\mathbb{E}[\hat{\mu}_1] = \mathbb{E}[\hat{\mu}_2] = \mu \quad (7.7)$$

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

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

$$\begin{aligned}Var(\hat{\mu}_2) &= Var\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) \\ &= Var\left(\frac{1}{3}X_1\right) + Var\left(\frac{1}{3}X_2\right) + Var\left(\frac{1}{3}X_3\right) \\ &= \frac{1}{9}Var(X_1) + \frac{1}{9}Var(X_2) + \frac{1}{9}Var(X_3) \\ &= \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 \\ &= \frac{3}{9}\sigma^2\end{aligned}$$

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, \dots, X_n una muestra aleatoria de una población dada y $\hat{\mu} = \bar{X}$ 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: $\mathbb{E}[\bar{X}] = \mu$. Respecto del segundo implica resolver:

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

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

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

Entonces, tomando los dos teoremas anteriores:

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

Por su parte:

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

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

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

Ejemplo. Sea X_1, X_2, \dots, 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 $X = X_1 + X_2 + \dots + X_n$ 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:

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

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

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

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:

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

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:

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

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

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

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

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

7.2.3. 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:

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

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 $n \rightarrow \infty$ 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 $n \rightarrow \infty$. De esta forma construimos el concepto de cercanía en la siguiente definición:

Definición 7.14 *Estimador consistente.* *El estimador o la estadística $\hat{\theta}_n$ ($= h(X_1, X_2, \dots, X_n)$) es un estimador consistente del parámetro θ de una pdf dada, si y sólo si, para cada $\varepsilon > 0$, entonces:*

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

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, \dots, 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 verifiquemos 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:

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

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 7.3 *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. Supongamos X_1, X_2, \dots, 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, \dots, X_n una muestra aleatoria, y sea $X = \sum_{i=1}^n X_i$, entonces:

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

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:

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

De forma relacionada podemos enunciar el siguiente teorema:

Teorema 7.4 Si $\hat{\theta}_n$ es un estimador insesgado de θ y $Var(\hat{\theta}_n) \rightarrow 0$, cuando $n \rightarrow \infty$, 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

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

7.2.4. Suficiencia

Un estimador $\hat{\theta}$ se dice que es suficiente si éste 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:

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

Definición 7.15 *Estimador Suficiente.* Sea X_1, X_2, \dots, X_n una muestra aleatoria de una pdf, la estadística o estimador $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ 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, \dots, X_n una muestra aleatoria de una función de densidad de probabilidad Poisson:

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

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:

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

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

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

¿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

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

Donde asumimos que, como demostraremos abajo, $X = X_1 + X_2 + \dots + X_n$ se distribuye como una Binomial:

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

Donde $k = 0, 1, \dots, n$.

La v.a. X es equivalente a resolver:

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

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:

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

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

Teorema 7.5 *La Estadística $\hat{\theta}$ es un estimador suficiente del parámetro θ si y sólo si la función de probabilidad conjunta o densidad conjunta de la muestra aleatoria dada puede ser factorizada como:*

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

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

Ejemplo. Sea \bar{X} un estimador de la media μ de una pdf normal 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:

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

Desarrollemos:

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

Utilizando el resultado anterior, podemos construir

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

7.3. 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, \dots, 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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n)$ 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 .

7.3.1. Método de Momentos

El método de momentos es el método más sencillo de resolver cuando enfrentamos el problema de plantear múltiples estimadores de los parámetros. Este método genera estimadores que no son óptimos pero son fácilmente 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 parámetros a estimar tengamos. Tomemos como definición de momentos muestrales a:

Definición 7.16 Momentos Muestrales. *El k -ésimo momento muestral de un conjunto de observaciones $\{x_1, x_2, \dots, x_n\}$ es la media de su k -ésimas potencias:*

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

Para $k = 1, 2, \dots, K$.

Tomemos como definición de momentos poblacionales a:

Definición 7.17 Momentos Poblacionales. *El k -ésimo momento poblacional de una v.a. con pdf dada:*

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

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

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 $\{\theta_1, \theta_2, \dots, \theta_K\} = \boldsymbol{\theta}$. Así, si la población tiene K parámetros, el método de momentos consiste en resolver para cada una de las incógnitas $\{\theta_1, \theta_2, \dots, \theta_K\}$ el sistema de ecuaciones:

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

Ejemplo. Sea $\{x_1, x_2, \dots, x_n\}$ 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 $\{x_1, x_2, \dots, x_n\}$ 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 obtenemos 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 :

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

7.3.2. Método de Máxima Verosimilitud

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

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

Dicho lo anterior, podemos establecer la siguiente definición:

Definición 7.18 Sea $\{x_1, x_2, \dots, x_n\}$ una muestra aleatoria de una pdf dada y sea $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_n\}$ un conjunto de parámetros poblacionales desconocidos. La función de verosimilitud para una muestra aleatoria estará dada para el caso continuo por la expresión:

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

y para el caso discreto por la expresión:

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

De esta forma el problema consiste en resolver:

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

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 θ tales que:

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

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

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

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

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

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:

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

Tomando la primera derivada e igualando a cero obtenemos:

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

$$\frac{k}{\hat{\theta}} - \frac{n - k}{1 - \hat{\theta}} = 0$$

$$\frac{k}{\hat{\theta}} = \frac{n - k}{1 - \hat{\theta}}$$

$$\hat{\theta} = \frac{k}{n}$$

Las segundas derivadas serán:

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

Ejemplo. Sea $\{x_1, x_2, \dots, x_n\}$ 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:

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

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

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

Tomando las condiciones de primer orden:

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

Las condiciones de segundo orden estarán dadas por:

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

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

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

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

$$\frac{d\ln(L(\theta))}{d\theta} = n \neq 0 \quad (7.8)$$

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 $\{x_1, x_2, \dots, x_n\}$ 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 verosimilitud para a y b .

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

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

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:

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

8

Estimación por Intervalo

8.1. 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 8.1 *Intervalo de confianza (1)*. Sean X_1, X_2, \dots, 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 8.2 Intervalo de confianza (2). Un intervalo de confianza para un parámetro θ es un intervalo $C_n = (a, b)$ en el cual ambos extremos $a = a(X_1, X_2, \dots, X_n)$ y $b = b(X_1, X_2, \dots, X_n)$ 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 8.3 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 específica $1 - \alpha$ 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 8.1 Teorema del Límite Central. Sea una muestra aleatoria $\{W_1, W_2, \dots, W_n\}$. Supongamos que la pdf asociada a cada W_i tiene una media μ y una varianza σ^2 . Para cualesquiera valores a y b :

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

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

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

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

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

Ahora, iniciemos la discusión con un ejemplo o caso ilustrativo. Supongamos que $\{x_1, x_2, \dots, x_n\}$ 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:

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

En este punto es necesario recordar que si tenemos una muestra aleatoria $\{x_1, x_2, \dots, x_n\}$ de una distribución normal con media y varianza dada por: μ y σ^2 , entonces \bar{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?:

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

En este punto solo nos falta determinar el valor de su media y varianza. Para responder determinemos los momentos, empezando por la media:

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

Por su parte, la varianza será:

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

Por lo tanto, si $\bar{X} \sim N(\mu, \sigma^2/n)$, 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 valor a partir del cual se acumula el 2.5 % de la probabilidad en cada una de las colas. Dicho valor, como discutiremos más adelante en la sección de prueba de hipótesis es 1.96.

Por lo tanto lo siguiente es cierto:

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

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

8.2. 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 esta sección.

Teorema 8.2 *Sea \bar{X} la media de una muestra aleatoria de tamaño n de una pdf normal con varianza conocida, entonces si \bar{X} es un estimador μ , la probabilidad de $1 - \alpha$ será descrita por:*

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

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 $\bar{X} = 64.3$. 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 8.3 *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 8.4 *Distribución t-Student para una muestra aleatoria.* Sean $\{x_1, x_2, \dots, x_n\}$ una muestra aleatoria de una pdf normal, de forma que cada x_i se distribuye como una $N(\mu, \sigma^2)$, entonces:

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

Donde:

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

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

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \quad (8.1)$$

Consideremos que dada la ecuación (8.1), entonces para un muestreo de tamaño $n = 1$ ¿cómo se distribuye una transformación dada por?:

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

Verifiquemos los momentos:

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

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

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

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

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:

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

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

son chi-cuadrado con n grados de libertad:

$$\begin{aligned}
Y &= \sum_{i=1}^n \frac{n}{n-1} \left(\frac{x_i - \bar{X}}{\sigma} \right)^2 \sim \chi_{[n]}^2 \\
&= \sum_{i=1}^n \frac{n}{n-1} \left(\frac{x_i - \bar{X}}{\sigma} \right)^2 \\
&= \sum_{i=1}^n \frac{n}{\sigma^2} \left(\frac{(x_i - \bar{X})^2}{n-1} \right) \\
&= \frac{\sum_{i=1}^n \frac{(x_i - \bar{X})^2}{n-1}}{\frac{\sigma^2}{n}} \sim \chi_{[n]}^2
\end{aligned} \tag{8.2}$$

Por lo tanto, el teorema anterior afirma que una distribución t-Student resultaría de la división de la ecuación (8.1) y la raíz cuadrada de la ecuación (8.2) dividida por sus grados de libertad:

$$\begin{aligned}
T &= \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\sum_{i=1}^n \frac{(x_i - \bar{X})^2}{n-1}}{\frac{\sigma^2}{n}}}} \\
&= \frac{\bar{X} - \mu}{\sqrt{\frac{\sum_{i=1}^n \frac{(x_i - \bar{X})^2}{n-1}}{n}}} \\
&= \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{[n-1]}
\end{aligned} \tag{8.3}$$

En adelante utilizaremos la ecuación (8.3) 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 :

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

Por lo tanto podemos enunciar el siguiente teorema:

Teorema 8.5 Sean \bar{X} y $\hat{\sigma}^2$ valores de la media y varianza muestrales de una muestra de tamaño n de una población normal, entonces, si \bar{X} es un estimador de μ y $\hat{\sigma}^2$ es un estimador de σ^2 , ña probabilidad $1 - \alpha$ será descrita por:

$$P\left(\bar{X} - t_{\alpha/2, n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2, n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha \quad (8.4)$$

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 $\bar{X} = 66.3$, $\hat{\sigma} = 8.4$ 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 (8.4), 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 8.1 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 convención diremos que es el caso cuando $n > 30$.

Cuadro 8.1: 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

8.3. 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:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (8.5)$$

Donde \bar{X}_1 y \bar{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 (8.5). Así, partamos de que esa expresión es la diferencia dada por:

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

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

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

Por lo que hace a la varianza tenemos que:

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

Por lo tanto, la transformación en la ecuación (8.5) se distribulle como normal con media cero y varianza 1:

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

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

Teorema 8.6 Sean \bar{X}_1 y \bar{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:

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

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

Al respecto, hagamos algunas observaciones. Por virtud 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 8.2

Cuadro 8.2: Estadísticas de dos muestras aleatorias

Estadística	Muestra 1	Muestra 2
\bar{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:

$$T = \frac{Z}{\sqrt{\frac{Y}{n_1 + n_2 - 2}}}$$

$$= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{[n_1 + n_2 - 2]} \quad (8.6)$$

Donde en la ecuación (8.6):

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

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 8.7 Sean \bar{X}_1 , \bar{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:

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

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 8.3:

Cuadro 8.3: Tabla de muestras aleatorias

	Muestra 1	Muestra 2
\bar{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 8.1.
- Si en el intervalo de confianza se encuentra el 0 podemos concluir que no hay diferencia estadísticamente significativamente.

8.4. Intervalos de confianza para proporciones y diferencia de proporciones

En muchos casos deseamos estimar intervalos de confianza de proporciones, probabilidades, porcentajes 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 condiciones podemos establecer que:

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

Por lo tanto, podemos establecer el siguiente teorema:

Teorema 8.8 Sea $X = k$ una v.a. binomial con parámetro θ y con una muestra n grande. Asumamos un estimador de θ dado por:

$$\hat{\theta} = \frac{X}{n} \quad (8.7)$$

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 $X = 136$ individuos 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 marihuana y algunos de sus productos derivados?

Supongamos que la encuesta más reciente indica que de 1,200 adultos seleccionados aleatoriamente, 396 se pronunciaron a favor de la despenalización y otros 804 manifestaron estar en contra. Digamos que buscamos aprobar 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:

$$0.33 - 0.0136 \times 1.96 < p < 0.33 + 0.0136 \times 1.96$$

$$0.30 < p < 0.36$$

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 hombres y mujeres que votan a favor 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 independientes 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:

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

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 8.9 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,

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

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 8.4.

Cuadro 8.4: 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:

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

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

8.5. Intervalos de confianza para varianzas y razón de varianzas

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

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

Entonces, podemos afirmar que:

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

Utilizando lo anterior, podemos establecer que:

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

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

Utilizando lo anterior, podemos enunciar el siguiente teorema:

Teorema 8.10 *Sea $\hat{\sigma}^2$ el valor de la varianza muestral de una muestra de tamaño n de una población normal, entonces:*

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

Donde:

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

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 loalizamos en tablas los valores:

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

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:

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

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 8.11 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 8.5. Determine un intervalo de confianza para el 98 %.

Cuadro 8.5: 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:

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

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

9

Pruebas de Hipótesis

9.1. 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.

Esta técnica estadística 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 parámetros Θ en dos conjuntos disjuntos Θ_0 y Θ_1 , y que pretendemos probar:

$$H_0 : \theta \in \Theta_0 \text{ Vs } H_a : \theta \notin \Theta_0 \quad (9.1)$$

Sea X una variable aleatoria y sea \mathcal{X} el rango de X . Así, probaremos una hipótesis definiendo un subconjunto 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:

$$R = \{x : T(x) > c\} \quad (9.2)$$

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 9.1 *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:

$$H_0 : \theta = \theta_0 \text{ Vs } H_1 : \theta \neq \theta_0$$

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:

$$H_0 : \theta \leq \theta_0 \text{ Vs } H_1 : \theta > \theta_0$$

o

$$H_0 : \theta \geq \theta_0 \text{ Vs } H_1 : \theta < \theta_0$$

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

9.2. 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 verdadero 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 9.1: 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 $\mu_1 > \mu_0$,

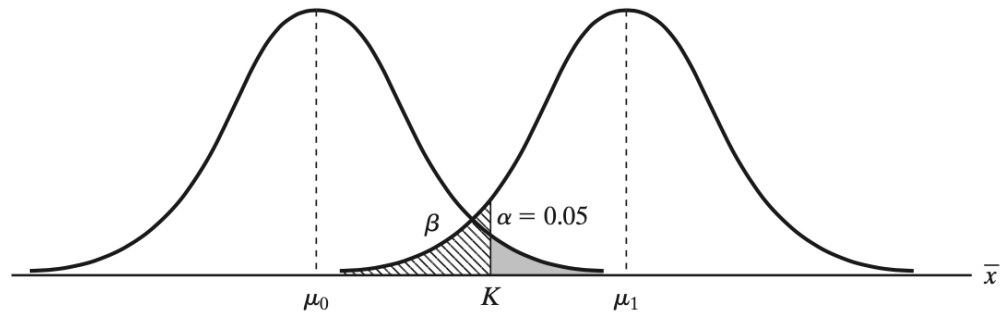


Figura 9.1: Ilustración de las probabilidades de Error Tipo I y Error Tipo II. Retomado de Miller y Miller (2014; p. 340) I. Miller y M. Miller 2014

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

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

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

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

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

Así, la constante será $K = \mu_0 + \frac{1.645}{\sqrt{n}}$. En la Figura 9.1 ilustramos el valor de esta constante. Asimismo, en la Figura 9.1 se ilustra los tipos de error y las probabilidades asociadas a cada uno.

En ese mismo sentido, definiremos el concepto de **Potencia de una Prueba**, la cual será el valor dado 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 decisión.

9.3. 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-

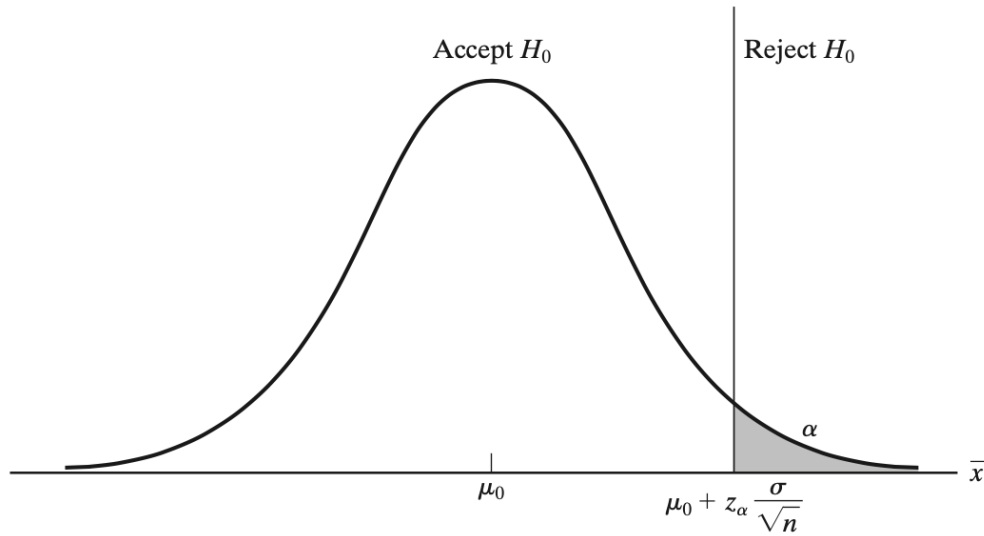


Figura 9.2: Hipótesis de una cola (derecha) y su zona de rechazo. Retomado de Miller y Miller (2014; p. 360) I. Miller y M. Miller 2014

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 $\{X_1, X_2, \dots, X_n\}$ 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:

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

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 original de \bar{X} , la zona de rechazo será como se describe en la Figura 9.2.
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 original de \bar{X} , la zona de rechazo será como se describe en la Figura 9.3.
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

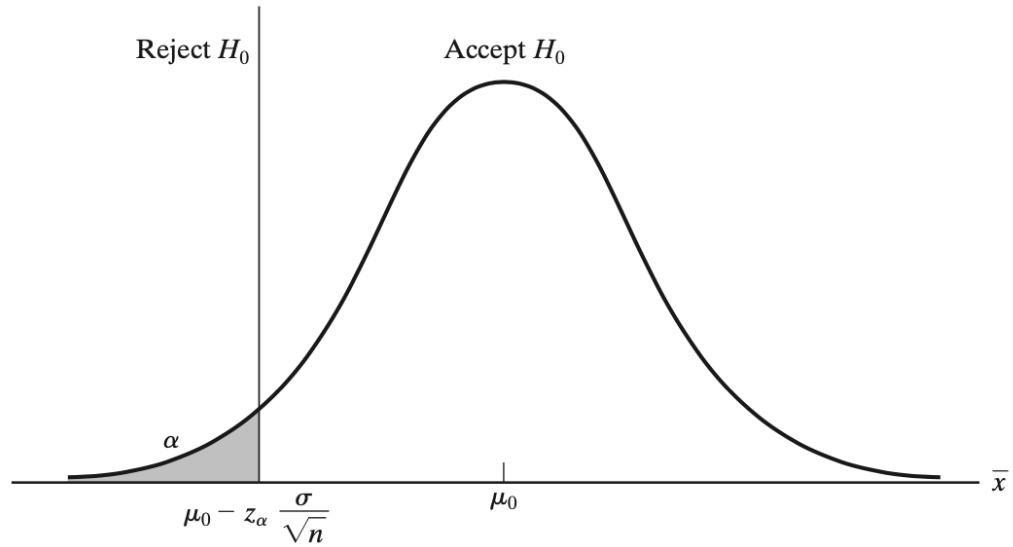


Figura 9.3: Hipótesis de una cola (izquierda) y su zona de rechazo. Retomado de Miller y Miller (2014; p. 361) I. Miller y M. Miller 2014

original de \bar{X} , las dos zonas de rechazo serán como se describe en la Figura 9.4.

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 .

9.4. Pruebas para medias, varianzas y proporciones

9.4.1. 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.

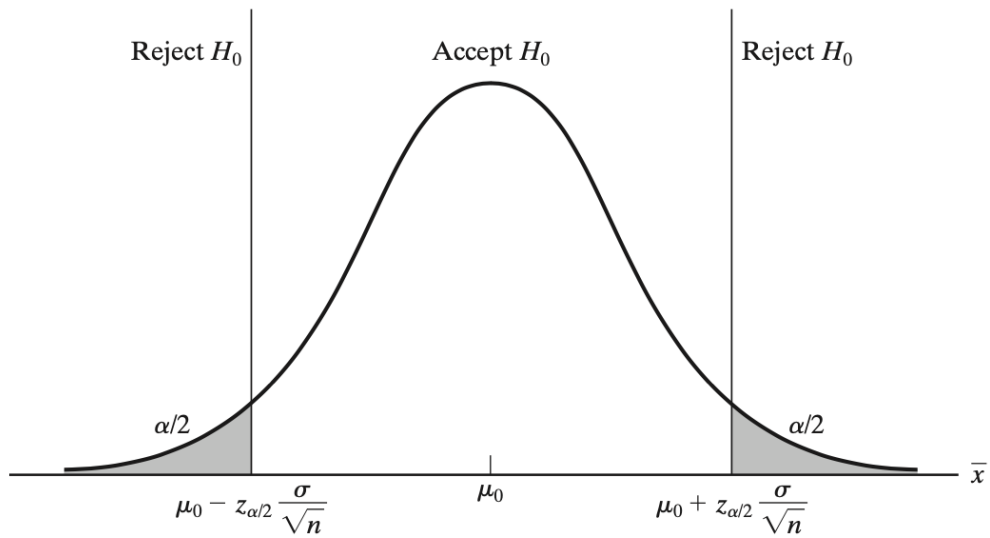


Figura 9.4: Hipótesis de dos colas y sus zonas de rechazo. Retomado de Miller y Miller (2014; p. 360) I. Miller y M. Miller 2014

Partamos suponiendo que queremos probar:

$$\begin{aligned}
 H_0 &: \mu = \mu_0 \\
 H_a &: \begin{cases} \mu \neq \mu_0 \\ \text{o} \\ \mu > \mu_0 \\ \text{o} \\ \mu < \mu_0 \end{cases}
 \end{aligned}
 \tag{9.3}$$

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, con las hipótesis alternativas en la ecuación (9.3):

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

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:

$$H_0 : \mu = 8$$

$$H_a : \mu \neq 8$$

Asumamos que queremos una región crítica o de rechazo del 1 %, es decir, es una prueba con una confianza del 99 %. Dicho esto, el valor crítico 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 suficientemente 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:

$$H_0 : \mu = 22$$

$$H_a : \mu < 22$$

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:

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

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:

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

En estos casos, de forma similar queremos probar las hipótesis de la ecuación (9.3), para la cual proponemos las siguientes regiones críticas o de su rechazo, respectivamente, con las hipótesis alternativas en la ecuación (9.3):

$$\left\{ \begin{array}{l} |T| \geq t_{[\alpha/2, n-1]} \\ \text{o} \\ T \geq t_{[\alpha, n-1]} \\ \text{o} \\ T \leq -t_{[\alpha, n-1]} \end{array} \right.$$

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

$$H_0 : \mu = 185$$

$$H_a : \mu < 185$$

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:

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

Por lo tanto, no podemos rechazar la hipótesis nula.

9.4.2. 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 plantear la siguiente estructura de hipótesis nula e hipótesis alternativa:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = \delta \\ H_a &: \begin{cases} \mu_1 - \mu_2 \neq \delta \\ \text{o} \\ \mu_1 - \mu_2 > \delta \\ \text{o} \\ \mu_1 - \mu_2 < \delta \end{cases} \end{aligned} \quad (9.4)$$

De forma similar al caso de hipótesis para medias podemos establecer la siguiente transformación para evaluar la estructura de hipótesis planteada en la ecuación (9.4):

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (9.5)$$

Dada la hipótesis nula en la ecuación (9.4) proponemos las siguientes regiones críticas o de su rechazo, respectivamente, con las hipótesis alternativas en la ecuación (9.5):

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

Ejemplo. Supongamos dos muestras aleatorias tomadas de poblaciones normales, para las cuales observamos los datos del cuadro 9.2. 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 plantear como:

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

Cuadro 9.2: Estadísticas de dos muestras aleatorias

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

Con los datos anteriores tenemos que:

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

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 cambia las hipótesis cuando desconocemos las varianzas de las poblaciones de donde provienen las muestras? En estos casos utilizaremos una estructura de hipótesis similar a la ecuación (9.4) y una transformación dada por:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{[n_1+n_2-2]} \quad (9.6)$$

Donde:

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

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

De forma similar establecemos las reglas de rechazo para la ecuación (9.6) 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 9.3. Supongamos que queremos determinar si con una confianza del 95 % la media de la población 1 es mayor que la media de la población 2.

Cuadro 9.3: Estadísticas de dos muestras aleatorias

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

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

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

Con los datos anteriores tenemos que:

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

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

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.

9.4.3. 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 función de densidad de probabilidad normal.

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

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

$$\begin{aligned} H_0 &: \sigma^2 = \sigma_0^2 \\ H_a &: \begin{cases} \sigma^2 \neq \sigma_0^2 \\ \text{o} \\ \sigma^2 > \sigma_0^2 \\ \text{o} \\ \sigma^2 < \sigma_0^2 \end{cases} \end{aligned} \quad (9.7)$$

Para los cual establecemos la siguiente transformación para evaluar la ecuación (9.7):

$$\chi^2 = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2} \sim \chi_{[n-1]}^2 \quad (9.8)$$

Donde

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

Dada la hipótesis nula en la ecuación (9.7) proponemos las siguientes regiones críticas o de su rechazo, respectivamente, con las hipótesis alternativas en la ecuación (9.8):

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

Ejemplo. Supongamos una 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 afirmación de que la varianza muestral es más grande que la varianza poblacional $\sigma^2 = 0.36$.

De esta forma podemos establecer 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 siguiente 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 $\chi_{[0.05;17]}^2 = 27.587$.

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

¿Qué pasa cuando 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:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_a &: \begin{cases} \sigma_1^2 \neq \sigma_2^2 \\ \text{o} \\ \sigma_1^2 > \sigma_2^2 \\ \text{o} \\ \sigma_1^2 < \sigma_2^2 \end{cases} \end{aligned} \quad (9.9)$$

Para los cual establecemos la siguiente transformación para evaluar la ecuación (9.9):

$$F = \frac{\frac{\hat{\sigma}_1^2}{\sigma_1^2}}{\frac{\hat{\sigma}_2^2}{\sigma_2^2}} = \frac{\hat{\sigma}_1^2 \sigma_2^2}{\hat{\sigma}_2^2 \sigma_1^2} \sim f_{[n_1-1, n_2-1]} \quad (9.10)$$

Dada la hipótesis nula en la ecuación (9.9) proponemos las siguientes regiones críticas o de su rechazo, respectivamente, con las hipótesis alternativas en la ecuación (9.10):

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

Ejemplo. Supongamos que tenemos 2 muestras aleatorias de 2 poblaciones con distribuciones normales, para las cuales observamos los datos del Cuadro 9.4.

Cuadro 9.4: 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:

$$F = \frac{\hat{\sigma}_2^2 \sigma_1^2}{\hat{\sigma}_1^2 \sigma_2^2} \sim f_{[n_1-1, n_2-1]} = \frac{19.2}{3.5} = 5.49 \quad (9.11)$$

De acuerdo con la tabla de F de Fisher, tenemos que $f_{[0.001:12,15]} = 3.67$. Así, $F = 5.49 > f_{[0.001:12,15]} = 3.67$, por lo que podemos rechazar la hipótesis nula de que las varianzas sean iguales.

9.4.4. 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.

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_a &: \begin{cases} \theta \neq \theta_0 \\ \text{o} \\ \theta > \theta_0 \\ \text{o} \\ \theta < \theta_0 \end{cases} \end{aligned} \quad (9.12)$$

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 críticas:

$$\left\{ \begin{array}{l} |Z| \geq z_{\alpha/2} \\ \text{o} \\ Z \geq z_{\alpha} \\ \text{o} \\ Z \leq -z_{\alpha} \end{array} \right.$$

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 porcentaje 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 porcentaje ha disminuido.

9.5. 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:

$$\lambda = \frac{L(\theta_0)}{L(\theta_u)} \quad (9.13)$$

Donde $L(\theta_0)$ es la función de verosimilitud del modelo restringido 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 9.2 Prueba de razón de verosimilitud. Sea X_1, X_2, \dots, 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 ser 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 es no concluyente si $\lambda = k$.

Donde k es una constante no negativa y:

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

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:

$$-2\ln\lambda \sim \chi_Q^2$$

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

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

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 $\bar{X} = 20$ y $\bar{Y} = 22$.

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

$$H_0 : \theta_1 = \theta_2 \quad (9.14)$$

$$H_1 : \theta_1 \neq \theta_2 \quad (9.15)$$

Recordemos que una función Poisson es de la siguiente forma:

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

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

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

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

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

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

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

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

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

Donde:

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

Así, las condicionnes de primer orden serían:

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

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:

$$\begin{aligned} \lambda &= \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_u)} \\ &= \frac{\frac{1}{x_1!, x_2!, \dots, x_n! \cdot y_1!, y_2!, \dots, y_n!} \cdot e^{-n2\hat{\theta} \sum_{i=1}^n x_i + \sum_{i=1}^n y_i}}{\frac{1}{x_1!, x_2!, \dots, x_n! \cdot y_1!, y_2!, \dots, y_n!} \cdot e^{(-n\hat{\theta}_1) + (-n\hat{\theta}_2) \hat{\theta}_1^{\sum_{i=1}^n x_i} \hat{\theta}_2^{\sum_{i=1}^n y_i}}} \\ &= \frac{e^{-n2\hat{\theta} \sum_{i=1}^n x_i + \sum_{i=1}^n y_i}}{e^{(-n\hat{\theta}_1) + (-n\hat{\theta}_2) \hat{\theta}_1^{\sum_{i=1}^n x_i} \hat{\theta}_2^{\sum_{i=1}^n y_i}}} \\ &= \frac{e^{-n2\hat{\theta} \hat{\theta}^{n\bar{X} + n\bar{Y}}}}{e^{(-n\hat{\theta}_1) + (-n\hat{\theta}_2) \hat{\theta}_1^{n\bar{X}} \hat{\theta}_2^{n\bar{Y}}}} \\ &= \frac{2^{1^{100 \times 20 + 100 \times 22}}}{20^{100 \times 20} 22^{100 \times 22}} \end{aligned}$$

Al finalizar,

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

De acuerdo con las tablas:

$$\chi_{0.01}^2 = 6.635$$

Por lo tanto, rechazamos la hipótesis nula.

9.6. Bondad de ajuste

10

Análisis de varianza y diseño de experimentos

10.1. Análisis de varianza: Introducción y motivación

El análisis de varianza está basado en una 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:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad (10.1)$$

Para probar si la hipótesis en la ecuación (10.1) es válida 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 construimos 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 $j = 1, 2, \dots, K$ y una varianza dada por σ^2 para todos los grupos j . Esto es:

$$f_{Y_{ij}}(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{y-\mu_j}{\sigma}\right)^2} \quad (10.2)$$

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: $y_{1j}, y_{2j}, \dots, y_{n_j j}$.

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

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

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:

$$SSTR = \sum_{j=1}^K \sum_{i=1}^{n_j} (\bar{Y}_j - \bar{\mathbf{Y}})^2 \quad (10.3)$$

Podemos desarrollar la ecuación (10.3) de forma que tendríamos:

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

A la expresión de la ecuación (10.3) también se le conoce como la variabilidad total de las medias muestrales respecto de la media total. A esta

también se le puede descomponer en una expansión alrededor de \bar{Y}_j , \bar{Y} y μ :

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

Una vez determinada una expresión alterna para la ecuación (10.3) podemos verificar su valor esperado:

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

Para continuar, requerimos recordar que:

$$\mathbb{E}[\bar{Y}] = \mu$$

Así:

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

Por otro lado, analicemos la siguiente expresión:

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

Por lo tanto:

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

Dados los resultados anteriores, podemos determinar que:

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

10.2. Planteamiento de la prueba de hipótesis

Prueba de $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$ 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 $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$ es cierta, entonces:

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

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

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

Pruebas de hipótesis cuando σ^2 es desconocida

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

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

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):

$$SSE = \sum_{j=1}^K (n_j - 1) \hat{\sigma}_j^2 = \sum_{j=1}^K \sum_{i=1}^{n_j} (y_{ij} - \bar{Y}_j)^2 \quad (10.4)$$

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

$$\frac{SSE}{\sigma^2} = \sum_{j=1}^K \sum_{i=1}^{n_j} \frac{(y_{ij} - \bar{Y}_j)^2}{\sigma^2} \sim \chi_{[n-K]}^2 \quad (10.5)$$

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

$$SSTOT = SSTR + SSE$$

Es decir,

$$SSTOT = \sum_{j=1}^K \sum_{i=1}^{n_j} (y_{ij} - \bar{Y})^2 = \sum_{j=1}^K \sum_{i=1}^{n_j} ((\bar{Y}_j - \bar{Y}) + (y_{ij} - \bar{Y}_j))^2 \quad (10.6)$$

Podemos demostrar que:

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

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

$$F = \frac{SSTR/(K-1)}{SSE/(n-K)} \sim F_{[K-1, n-K]} \quad (10.7)$$

Por lo tanto, rechazaremos la hipótesis nula siempre que:

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

10.3. 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: $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$. Así, la tabla se puede construir como se muestra en el Cuadro 10.1.

Cuadro 10.1: Tabla ANOVA					
Fuente	df	SS	MS	F	Prob.
Tratamiento	$K-1$	$SSTR$	$MSTR = \frac{SSTR}{K-1}$	$\frac{MSTR}{MSE}$	$\mathbb{P}(F_{[K-1, n-K]} \geq F)$
Error	$n-K$	SSE	$MSE = \frac{SSE}{n-K}$		
Total	$n-1$	$SSTOT$			

Veamos 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 10.2 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 10.2: 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
\bar{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 diferencia 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:

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

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

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

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, 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 10.3: 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			

10.4. 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 mecanismo menos formal, el cual consiste en refinar el experimento minimizando el error asociado a cada individuo. Por otro lado, existe un mecanismo más formal, el cual permite reducir el error realizando experimentos aleatorios por bloques. La Figura 10.1 muestra ejemplos de cómo realizar los experimentos por bloques.

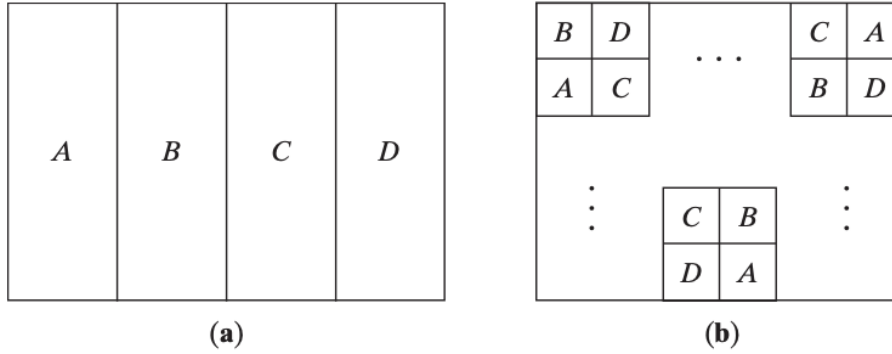


Figura 10.1: Dos diseños experimentales diferentes

10.5. 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 una 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:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_K$$

La misma que hemos planteado previamente. No obstante, el modelo matemático planteado es diferente, en el cual asumimos un término adicional.

Así, incorporamos al análisis el efecto del bloque i -ésimo β_i como:

$$Y_{ij} = \mu_j + \beta_i + \varepsilon_{ij} \quad (10.8)$$

Donde ε_{ij} se distribuye de forma normal con media cero (0) y varianza σ^2 para $i = 1, 2, \dots, b$ y $j = 1, 2, \dots, K$. 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:

$$\mu = \frac{1}{K} \sum_{j=1}^K \mu_j \quad (10.9)$$

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: $\beta_1, \beta_2, \dots, \beta_b$.

Sea $\bar{Y}_i = \frac{1}{K} \sum_{j=1}^K Y_{ij}$ 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:

$$\begin{aligned} \bar{Y}_i &= \frac{1}{K} \sum_{j=1}^K (\mu_j + \beta_i) \\ &= \frac{1}{K} \sum_{j=1}^K \mu_j + \frac{1}{K} \sum_{j=1}^K \beta_i \\ &= \mu + \beta_i \end{aligned}$$

Si \bar{Y} es sustituido por el parámetro del cual es su estimador, entonces podemos expresar el efecto de bloque como:

$$\beta_i = \bar{Y}_i - \bar{Y}$$

Sabiendo lo anterior, podemos alterar la forma en que hemos calculado SSE para incorporar o separar el efecto del bloque:

$$\begin{aligned}
\sum_{i=1}^b \sum_{j=1}^K (Y_{ij} - \bar{Y}_j)^2 &= \sum_{i=1}^b \sum_{j=1}^K [(Y_{ij} - \bar{Y}_j) + (\bar{Y}_i - \bar{Y}) - (\bar{Y}_i - \bar{Y})]^2 \\
&= \sum_{i=1}^b \sum_{j=1}^K [(\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_j - \bar{Y}_i + \bar{Y})]^2 \\
&= \sum_{i=1}^b \sum_{j=1}^K (\bar{Y}_i - \bar{Y})^2 \\
&\quad + \sum_{i=1}^b \sum_{j=1}^K (Y_{ij} - \bar{Y}_j - \bar{Y}_i + \bar{Y})^2 \\
&\quad + 2 \sum_{i=1}^b \sum_{j=1}^K (\bar{Y}_i - \bar{Y})(Y_{ij} - \bar{Y}_j - \bar{Y}_i + \bar{Y}) \\
&= \sum_{i=1}^b \sum_{j=1}^K (\bar{Y}_i - \bar{Y})^2 \\
&\quad + \sum_{i=1}^b \sum_{j=1}^K (Y_{ij} - \bar{Y}_j - \bar{Y}_i + \bar{Y})^2
\end{aligned}$$

En este caso hemos empleado un resultado para el término:

$$\begin{aligned}
&\sum_{i=1}^b \sum_{j=1}^K (\bar{Y}_i - \bar{Y})(Y_{ij} - \bar{Y}_j - \bar{Y}_i + \bar{Y}) \\
&= \sum_{i=1}^b (\bar{Y}_i - \bar{Y}) \sum_{j=1}^K (Y_{ij} - \bar{Y}_j - \bar{Y}_i + \bar{Y}) \\
&= \sum_{i=1}^b (\bar{Y}_i - \bar{Y}) \left(\sum_{j=1}^K Y_{ij} - \sum_{j=1}^K \bar{Y}_j - \sum_{j=1}^K \bar{Y}_i + \sum_{j=1}^K \bar{Y} \right) \\
&= \sum_{i=1}^b (\bar{Y}_i - \bar{Y}) (K\bar{Y}_i - K\bar{Y} - K\bar{Y}_i + K\bar{Y}) \\
&= 0
\end{aligned}$$

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:

$$SSB = \sum_{i=1}^b \sum_{j=1}^K (\bar{Y}_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^b \sum_{j=1}^K (Y_{ij} - \bar{Y}_j - \bar{Y}_i + \bar{Y})^2$$

Así, estableceremos que:

$$SSTOT = \sum_{i=1}^b \sum_{j=1}^K (Y_{ij} - \bar{Y})^2$$

$$SSTR = \sum_{i=1}^b \sum_{j=1}^K (\bar{Y}_j - \bar{Y})^2$$

Una propiedad similar al caso de ANOVA es que:

$$SSTOT = SSTR + SSB + SSE$$

En el Cuadro 10.4 reportamos una tabla que ilustra las distintas relaciones que existen en la medias por bloque y tratamiento.

Cuadro 10.4: Tabla de bloques y diseño experimental

	Nivel de tratammiento					Total de bloques	Media de bloques	Efecto verdadero del bloque
	1	2	...	K				
Bloques	1	Y_{11}	Y_{12}	\cdots	Y_{1K}	T_1	\bar{Y}_1	β_1
	2	Y_{21}	Y_{22}	\cdots	Y_{2K}	T_2	\bar{Y}_2	β_2
	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots	\vdots
	b	Y_{b1}	Y_{b2}	\cdots	Y_{bK}	T_b	\bar{Y}_b	β_b
Total de muestra	T_1	T_2	\cdots	T_K	T			
Media de muestra	\bar{Y}_1	\bar{Y}_2	\cdots	\bar{Y}_K			\bar{Y}	
Verdadera media	μ_1	μ_2	\cdots	μ_K				

Teorema 10.1 *Supongamos K niveles de tratamiento, con $\mu_1, \mu_2, \dots, \mu_K$ que son medias de un conjunto de b bloques, donde los efectos del bloque son $\beta_1, \beta_2, \dots, \beta_b$. Entonces:*

- Cuando $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$ es cierta, entonces:

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

- Cuando $H_0 : \beta_1 = \beta_2 = \dots = \beta_b$ 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 10.2 *Supongamos que K tratamientos con medias $\mu_1, \mu_2, \dots, \mu_K$ son medias de los bloques, entonces:*

- Si $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$ 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 , $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$ será rechazada si:

$$F \geq F_{[1-\alpha, K-1, (b-1)(K-1)]}$$

Teorema 10.3 *Supongamos que k tratamientos con medias sobre un conjunto de b bloques, dado el efecto del bloque son $\beta_1, \beta_2, \dots, \beta_b$ entonces:*

- Si $H_0 : \beta_1 = \beta_2 = \dots = \beta_b$ 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 , $H_0 : \beta_1 = \beta_2 = \dots = \beta_b$ será rechazada si:

$$F \geq F_{[1-\alpha, b-1, (b-1)(K-1)]}$$

Cuadro 10.5: Tabla ANOVA para Bloques

Fuente	df	SS	MS	F	Prob.
Tratamiento	$K - 1$	$SSTR$	$\frac{SSTR}{K-1}$	$\frac{SSTR/(K-1)}{SSE/(b-1)(K-1)}$	$P(F_{[K-1,(b-1)(K-1)]} \geq F)$
Bloques	$b - 1$	SSB	$\frac{SSB}{b-1}$	$\frac{SSB/(b-1)}{SSE/(b-1)(K-1)}$	$P(F_{[b-1,(b-1)(K-1)]} \geq F)$
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: $H_0 : \mu_1 = \mu_2 = \dots = \mu_K$ y $H_0 : \beta_1 = \beta_2 = \dots = \beta_b$. Así, la tabla se puede construir como se muestra en el Cuadro 10.5.

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 de 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 10.6 muestra los resultados.

Al respecto, pruebe si las medias de 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 10.6: 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 responder consideremos un nivel de significancia del 95 %.
El Cuadro 10.7 muestra los totales por bloque de tratamiento.

Cuadro 10.7: 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:

$$\begin{aligned}
 SSTR &= \sum_{j=1}^K \frac{T_j^2}{b} - \frac{T^2}{bK} \\
 &= \frac{511.4^2}{4} + \frac{516.8^2}{4} - \frac{1028.2^2}{8} \\
 &= 3.645
 \end{aligned}$$

$$\begin{aligned}
 SSB &= \sum_{i=1}^b \frac{T_i^2}{K} - \frac{T^2}{bK} \\
 &= \frac{228.0^2}{2} + \frac{248.7^2}{2} + \frac{267.6^2}{2} + \frac{283.9^2}{2} - \frac{1028.2^2}{8} \\
 &= 872.925
 \end{aligned}$$

$$\begin{aligned}
SSTOT &= \sum_{i=1}^b \sum_{j=1}^K Y_{ij}^2 - \frac{1028.2^2}{8} \\
&= (112.9)^2 + (123.1)^2 + \cdots + (143.7)^2 - \frac{1028.2^2}{8} \\
&= 888.515
\end{aligned}$$

$$\begin{aligned}
SSE &= SSTOT - SSTR - SSB \\
&= 888.515 - 3.645 - 872.925 \\
&= 11.945
\end{aligned}$$

El Cuadro 10.8 muestra los resultados en forma de tabla ANOVA.

Cuadro 10.8: 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 hipótesis nulas son: i) Igualdad de medias: $F_{[1,3]} = 10.13 > F = 0.92$ y ii) Igualdad de efecto de bloque $F_{[3,3]} = 9.28 < F = 73.07$

10.6. Análisis de datos categóricos

10.6.1. 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 resultado de 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 ; $i = 1, 2, \dots, K$. Donde: $p_1 + p_2 + \dots + p_K = 1$.
4. Los ensayos son independientes.
5. Estamos interesados en n_1, n_2, \dots, n_K , donde n_i , $i = 1, 2, \dots, K$, es igual al número de ensayos para los cuales el resultado cae en la categoría i y $n = n_1 + n_2 + \dots + n_K$.

10.6.2. Prueba de Ji Cuadrado

En un modelo multinomial cada una de las n_i tiene distribución binomial con parametros n y p_i . Entonces:

$$\mathbb{E}[n_i] = np_i; \text{ para } i = 1, 2, \dots, K$$

En nuestro caso tenemos probabilidades por grupo p_1, p_2, \dots, 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 , $i = 1, 2, \dots, K$. Entonces podemos calcular K desviaciones:

$$n_i - \mathbb{E}[n_i] = n_i - np_i; \text{ para } i = 1, 2, \dots, K$$

En 1900, PEARSON propuso la siguiente estadística de prueba:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^K \frac{(n_i - \mathbb{E}[n_i])^2}{\mathbb{E}[n_i]} \\ &= \sum_{i=1}^K \frac{(n_i - np_i)^2}{np_i} \sim \chi_{[K-1]}^2 \end{aligned}$$

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 10.9. 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 10.9: Número de accidentes por semana

$Y = y$	Frequency
0	32
1	12
2	6
3 o más	0

Para dar solución a este problema partamos la función de distribución Poisson:

$$p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}; \text{ para } y = 0, 1, 2, \dots$$

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 encontrar cada uno de los varloes esperados $\mathbb{E}[n_i] = np_i$. Los calculos se muestran en el Cuadro 10.10.

Cuadro 10.10: Número de accidentes por semana y calculos relacionados

$Y = y$	Frequency (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 $\chi_{[3]}^2 = 7.81 > \chi^2 = 2.923$, por lo que no podemos rechazar la hipótesis nula de que los datos sigan una distribución Poisson.

10.7. Consideraciones para el diseño de experimentos

10.7.1. 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).

10.7.2. 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 conocíamos la varianzas σ_1^2 y σ_2^2 :

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Así, cuanto menor sea $\sigma_{\bar{Y}_1 - \bar{Y}_2}$, menor será el correspondiente error del cálculo y mayor la cantidad de información de la muestra.

10.7.3. El experimento de observaciones pareadas

En muchos experimentos las muestras 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 individuo 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 individuo recibe el otro medicamento.

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{Y})^2 &= \mathbf{Y}'\mathbf{M}^0\mathbf{Y} \\
\sum_{i=1}^n (y_i - \bar{Y}) &= \sum_{i=1}^n (y_i^2 - 2y_i\bar{Y} + \bar{Y}^2) \\
&= \sum_{i=1}^n y_i^2 - 2\bar{Y} \sum_{i=1}^n y_i + \sum_{i=1}^n \bar{Y}^2 \\
&= \sum_{i=1}^n y_i^2 - 2 \frac{\sum_{i=1}^n y_i}{n} \sum_{i=1}^n y_i + n\bar{Y}^2 \\
&= \sum_{i=1}^n y_i^2 - \frac{2}{n} \left(\sum_{i=1}^n y_i \right)^2 + n \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2 \\
&= \sum_{i=1}^n y_i^2 - \frac{2}{n} \left(\sum_{i=1}^n y_i \right)^2 + \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \\
&= \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2
\end{aligned}$$

$$\begin{aligned}
\mathbf{e}'\mathbf{e} &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
&= 18310.63 - 2 \begin{bmatrix} 559.60 & 7375.44 & 324972.70 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \\
&\quad + \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \end{bmatrix} \begin{bmatrix} 25.00 & 219.00 & 9932.00 \\ 219.00 & 3055.00 & 129099.00 \\ 9932.00 & 129099.00 & 6402888.00 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
Cov(x, y) &= \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \\
Cov(p_i, u_i) &= \mathbb{E}[(p_i - \mathbb{E}[p_i])(u_i - \mathbb{E}[u_i])] \\
&= \mathbb{E} \left[\left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1} - \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} \right) (u_i) \right] \\
&= \mathbb{E} \left[\left(\frac{v_i - u_i}{\alpha_1 - \beta_1} \right) (u_i) \right] \\
&= \dots
\end{aligned}$$

11

Regresión Lineal

11.1. 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 11.1, 11.2, 11.3, 11.4 muestran la relación geométrica entre la latitud y diversas variables del clima para una muestra de 500 ciudades en la fecha que se señala.

11.2. El concepto de regresión entre dos variables

Definición 11.1 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 $Y = y$.

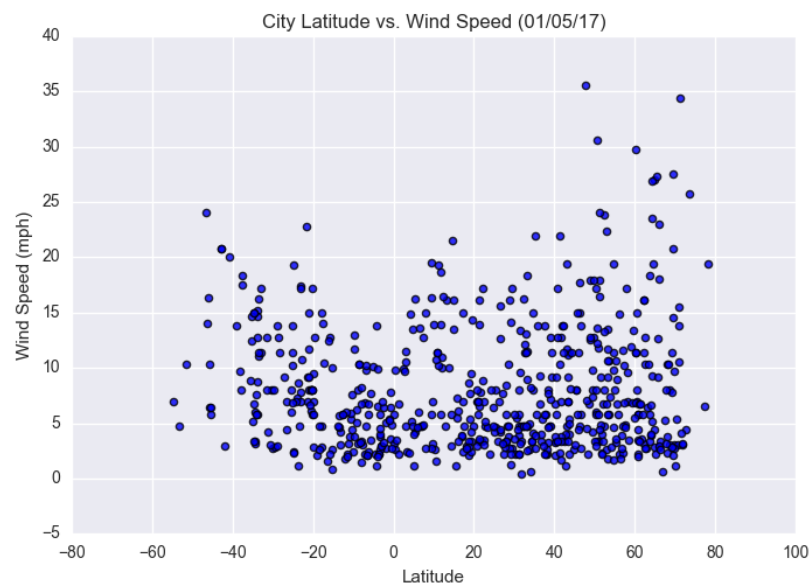


Figura 11.1: Relación entre Latitud y Velocidad del Viento

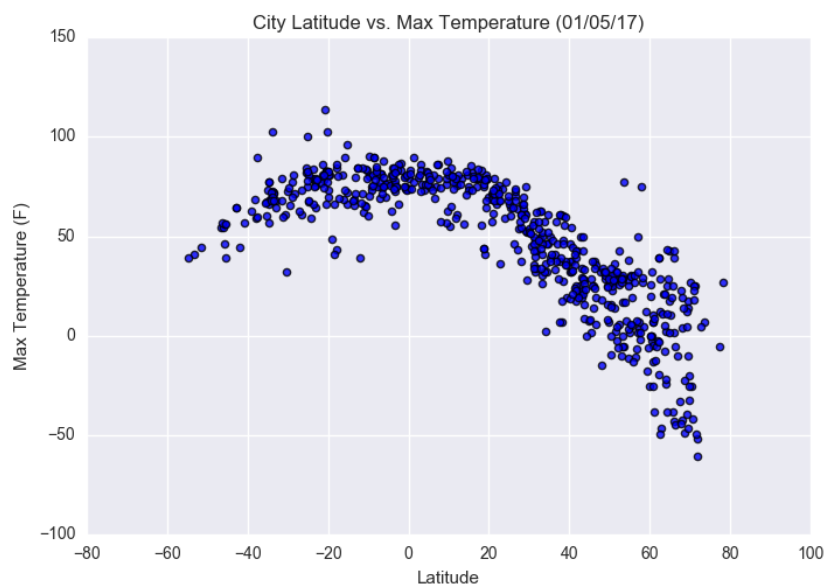


Figura 11.2: Relación entre Latitud y Temperatura

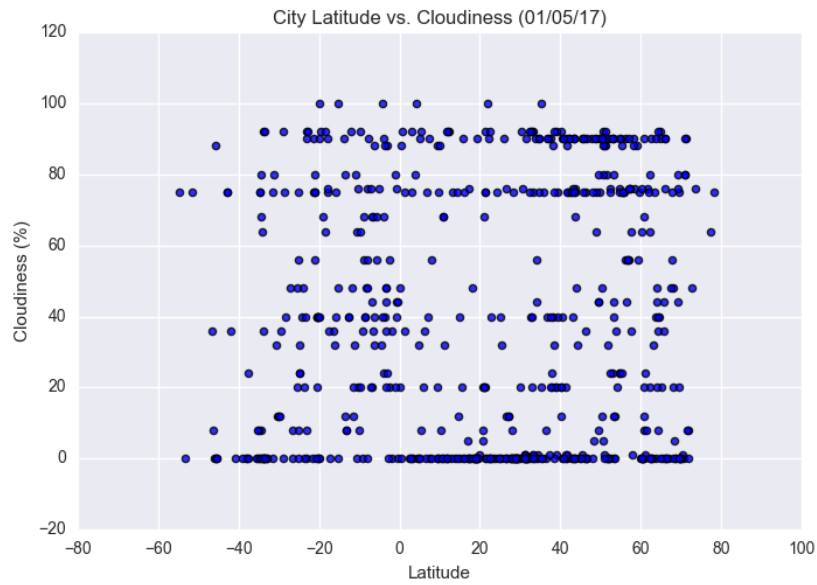


Figura 11.3: Relación entre Latitud y Nubocidad

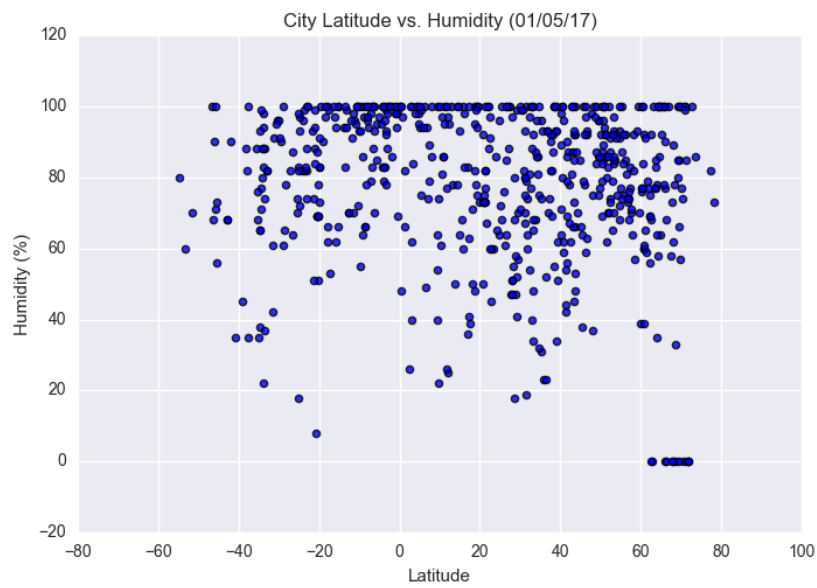


Figura 11.4: Relación entre Latitud y Humedad

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:

$$\begin{aligned} \int_D f(Y|x) dy &= \int_0^1 \frac{x+y}{x+\frac{1}{2}} dy = \int_0^1 \frac{x}{x+\frac{1}{2}} dy + \int_0^1 \frac{y}{x+\frac{1}{2}} dy \\ &= \frac{xy}{x+\frac{1}{2}} \Big|_0^1 + \frac{y^2/2}{x+\frac{1}{2}} \Big|_0^1 = \frac{x}{x+\frac{1}{2}} + \frac{\frac{1}{2}}{x+\frac{1}{2}} \\ &= \frac{x+\frac{1}{2}}{x+\frac{1}{2}} = 1 \end{aligned}$$

Por lo tanto, observamos que $f(Y|x)$ es una función de densidad de probabilidad. Ahora determinemos $\mu_{Y|x}$:

$$\begin{aligned} \mu_{Y|x} &= \mathbb{E}[Y|x] = \int_0^1 y \frac{x+y}{x+\frac{1}{2}} dy = \int_0^1 \frac{xy + y^2}{x+\frac{1}{2}} dy = \frac{\frac{xy^2}{2} + \frac{y^3}{3}}{x+\frac{1}{2}} \Big|_0^1 \\ &= \frac{\frac{x}{2} + \frac{1}{3}}{x+\frac{1}{2}} = \frac{\frac{3x}{2} + 1}{3x + \frac{3}{2}} = \frac{3x+2}{6x+3} \end{aligned}$$

Podemos plantear una definición más amplia:

Definición 11.2 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:

$$\begin{aligned}g(x_1, x_3) &= \int_{D_{x_2}} f(x_1, x_2, x_3) dx_2 = \int_0^1 (x_1 + x_2)e^{-x_3} dx_2 \\ &= \left(x_1 x_2 + \frac{x_2^2}{2} \right) e^{-x_3} \Big|_0^1 \\ &= \begin{cases} \left(x_1 + \frac{1}{2} \right) e^{-x_3} & \text{para } 1 > x_1 > 0, x_3 > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}\end{aligned}$$

Determinemos entonces la función de densidad condicional, la cual estará dada por:

$$\begin{aligned}f(X_2|x_1, X_3) &= \frac{f(x_1, x_2, x_3)}{g(x_1, x_3)} = \frac{(x_1 + x_2)e^{-x_3}}{\left(x_1 + \frac{1}{2} \right) e^{-x_3}} \\ &= \begin{cases} \frac{x_1 + x_2}{x_1 + \frac{1}{2}} & \text{para } 1 > x_1 > 0, 1 > x_2 > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}\end{aligned}$$

Entonces, la ecuación de regresión estará dada por:

$$\begin{aligned}\mu_{X_2|x_1, x_3} &= \int_{D_{x_2}} x_2 \cdot f(X_2|x_1, X_3) dx_2 = \int_0^1 x_2 \frac{x_1 + x_2}{x_1 + \frac{1}{2}} dx_2 = \frac{\frac{x_1 x_2^2}{2} + \frac{x_2^3}{3}}{x_1 + \frac{1}{2}} \Big|_0^1 \\ &= \frac{\frac{x_1}{2} + \frac{1}{3}}{x_1 + \frac{1}{2}} = \begin{cases} \frac{x_1 + \frac{2}{3}}{2x_1 + 1} & \text{para } 1 > x_1 > 0 \\ 0 & \text{en cualquier otro caso} \end{cases}\end{aligned}$$

11.3. Modelo de regresión lineal

11.3.1. 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. **Homocedasticidad.** La desviación estándar, 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 11.5.

Ahora enunciaremos un teorema relevante para el caso del modelo lineal.

Teorema 11.1 *Sean Y y X dos variables aleatorias que cumplen con los supuestos del modelo lineal. Si la regresión de Y en X es lineal, 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:

$$\mu_{Y|x} = \int_{D_y} y \cdot f(Y|x) dy = \beta_0 + \beta_1 x \quad (11.1)$$

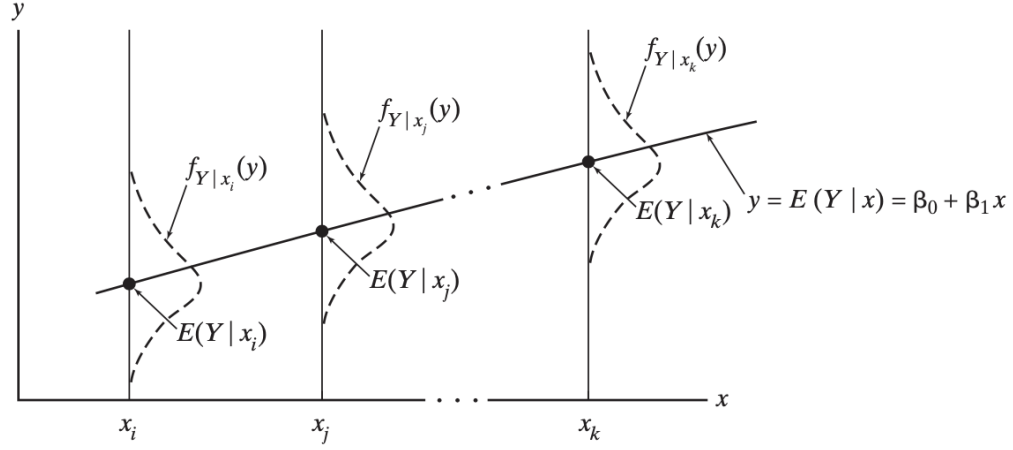


Figura 11.5: Supuestos del modelo lineal (Retomado de Larsen (2012, p 545) Larsen y Marx 2012), donde se muestra una notación con las siguientes equivalencias $f_{Y|x}(y) = f(Y|x)$ y $y = \mu_{Y|x} = \mathbb{E}[Y|x]$

Multiplicando cada uno de los lados de la última parte de la ecuación (11.1) por $g(x)$, la función de densidad de probabilidad marginal, e integrando ambos lados tenemos lo siguiente:

$$\begin{aligned}
 \int_{D_x} \int_{D_y} y \cdot f(Y|x) g(x) dy dx &= \beta_0 \int_{D_x} g(x) dx + \beta_1 \int_{D_x} x \cdot g(x) dx \\
 \iff \int_{D_x} \int_{D_y} y \cdot f(x, y) dy dx &= \beta_0 + \beta_1 \mu_X \\
 \iff \int_{D_y} \int_{D_x} y \cdot f(x, y) dx dy &= \beta_0 + \beta_1 \mu_X \\
 \iff \int_{D_y} y \cdot g(y) dy &= \beta_0 + \beta_1 \mu_X \\
 \iff \mu_Y &= \beta_0 + \beta_1 \mu_X
 \end{aligned} \tag{11.2}$$

Retomando la ecuación (11.1) y multiplicando por un factor dado por

$x \cdot g(x)$ e integrando, tenemos:

$$\begin{aligned}
\int_{D_x} \int_{D_y} y \cdot f(Y|x) \cdot x \cdot g(x) dy dx &= \beta_0 \int_{D_x} x \cdot g(x) dx + \beta_1 \int_{D_x} x^2 \cdot g(x) dx \\
\iff \int_{D_x} \int_{D_y} x \cdot y \cdot f(x, y) dy dx &= \beta_0 \mu_X + \beta_1 \mathbb{E}[X^2] \\
\iff \mathbb{E}[X \cdot Y] &= \beta_0 \mu_X + \beta_1 \mathbb{E}[X^2]
\end{aligned} \tag{11.3}$$

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 (11.2) y (11.3) 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 demuestra de forma similar.

11.3.2. Estimación de los parámetros del modelo lineal bivariado

Antes de plantear el proceso de estimación respondamos ¿cuáles son los parámetros $\boldsymbol{\theta}' = (\theta_1, \theta_2, \dots, \theta_K)'$ 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 $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ un conjunto de datos que satisfacen los supuestos del modelo lineal, entonces podemos establecer la siguiente transformación Z y la función de verosimilitud:

$$Z_i = \frac{y_i - \mathbb{E}[Y|x_i]}{\sigma}$$

$$\begin{aligned} L &= \prod_{i=1}^n f(Y_i|x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}Z_i^2} = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \mathbb{E}[Y|x_i]}{\sigma}\right)^2} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2} \end{aligned} \tag{11.4}$$

Tomando el logaritmo de la ecuación (11.4) tenemos:

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^n \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma} \right)^2 \quad (11.5)$$

$$-2\ln(L) = n \cdot \ln(2\pi) + n \cdot \ln(\sigma^2) + \sum_{i=1}^n \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma} \right)^2 \quad (11.6)$$

Ambas ecuaciones (11.5) y (11.6) son equivalentes. La ecuación (11.6) es conocida por ser la expresión de la forma log-normal o una expresión de Chi cuadrado. De la ecuación (11.5) 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:

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) &= 0 \\ \sum_{i=1}^n y_i - \sum_{i=1}^n \beta_0 - \sum_{i=1}^n \beta_1 x_i &= 0 \\ \sum_{i=1}^n y_i - n\beta_0 - \sum_{i=1}^n \beta_1 x_i &= 0 \\ \beta_0 &= \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n} \end{aligned} \quad (11.7)$$

Para la segunda ecuación del gradiente y utilizando 11.7 tenemos:

$$\begin{aligned}
\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i &= 0 \\
\sum_{i=1}^n y_i x_i - \sum_{i=1}^n \beta_0 x_i - \sum_{i=1}^n \beta_1 x_i^2 &= 0 \\
\sum_{i=1}^n y_i x_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 &= 0 \\
\sum_{i=1}^n y_i x_i - \left(\frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n} \right) \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 &= 0 \\
n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i \sum_{i=1}^n x_i - \beta_1 n \sum_{i=1}^n x_i^2 &= 0 \\
-\sum_{i=1}^n y_i \sum_{i=1}^n x_i + n \sum_{i=1}^n y_i x_i + \beta_1 \left(\left(\sum_{i=1}^n x_i \right)^2 - n \sum_{i=1}^n x_i^2 \right) &= 0 \quad (11.8)
\end{aligned}$$

Utilizando las ecuaciones (11.7) y (11.8) llegamos a las siguientes soluciones para los estimadores de los parámetros:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad (11.9)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (11.10)$$

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:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n} \quad (11.11)$$

Estos estimadores de las ecuaciones (11.10), (11.9) y (11.11) requieren de verificar sus propiedades de sesgo, consistencia y eficiencia. No obstante, lo dejaremos para más adelante en el caso de regresión múltiple.

11.4. 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 mínimos 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: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ 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 $i = 1, 2, \dots, n$ 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 planteamiento es válido para cualquier grado de polinomio, diremos que:

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

En la Figura 11.6 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. En la función introducimos los residuales al cuadrado con el objeto de aproximar el problema como uno de minimización de distancias

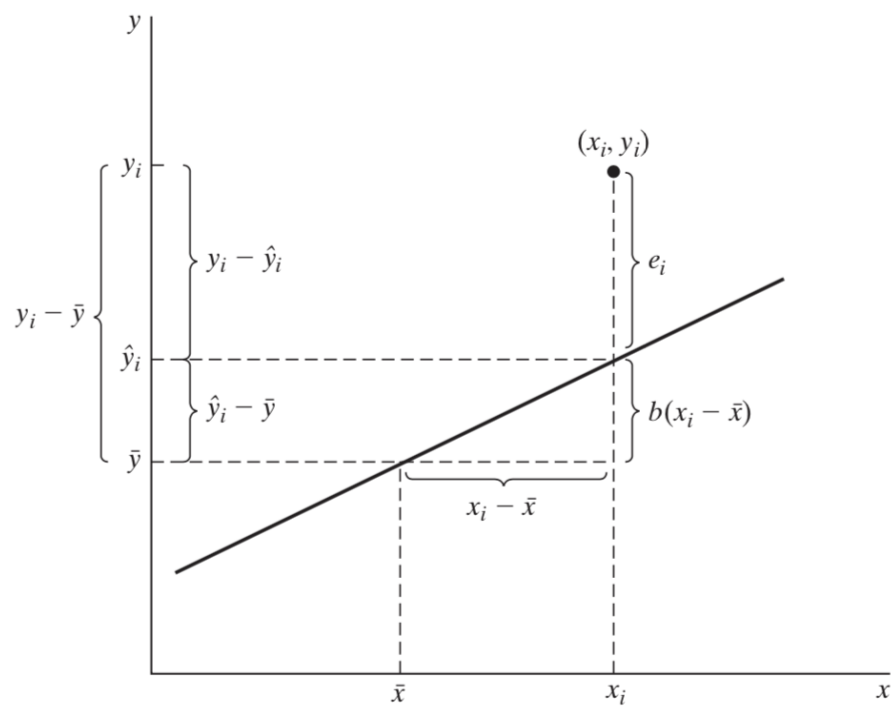


Figura 11.6: Ilustración del error de estimación, retomado de Greene (2012, pp. 40) Greene 2012

entre la coordenada y_i y la coordenada x_i de forma conjunta para todos los elementos en la muestra, $i = 1, 2, \dots, n$.

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 minimizar la función anterior requerimos de las condiciones 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 resolvemos 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:

$$\begin{aligned}\hat{\beta}_0 &= \frac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n} \\ &= \frac{\sum_{i=1}^n y_i}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^n x_i}{n} \\ &= \bar{Y} - \hat{\beta}_1 \bar{X}\end{aligned}$$

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

Por lo tanto, las soluciones serían:

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

Estas soluciones son las mismas que obtuvimos por el procedimiento de Máxima Verosimilitud.

11.5. 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 (11.12). Note que hemos eliminado los subíndices para indicar que la ecuacion se

cumple para cada uno de los elementos en la muestra, $i = 1, 2, \dots, n$. 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.

$$y = a + bx \quad (11.12)$$

Esta ecuación (11.12) tiene la característica de que tiene una tasa de variación constante para y en relación a x , de esta forma:

$$\frac{dy}{dx} = b \iff dy = bdx \iff \int dy = \int bdx \iff y = a + bx$$

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 .

11.5.1. Regresión exponencial

Supongamos que y depende de x y que los cambios en y derivados de x son proporcionales a y , es decir:

$$\frac{dy}{dx} = by$$

Donde b es una constante. De esta forma podemos encontrar que:

$$\frac{dy}{dx} = by \iff \frac{dy}{y} = bdx \iff \int \frac{dy}{y} = \int bdx \iff \ln(y) = a + bx$$

De esta forma, tenemos que:

$$y = e^{bx} e^a \iff y = ke^{bx}$$

Por lo tanto, cuando estimamos una relación: $\ln(y) = a + bx$ estamos asumiendo que existe la siguiente relación: $\frac{dy}{dx} = by$. A este tipo de ecuaciones se les conoce como log-lineales. En la Figura 11.7 se ilustra la relación log-lineal - exponencial.

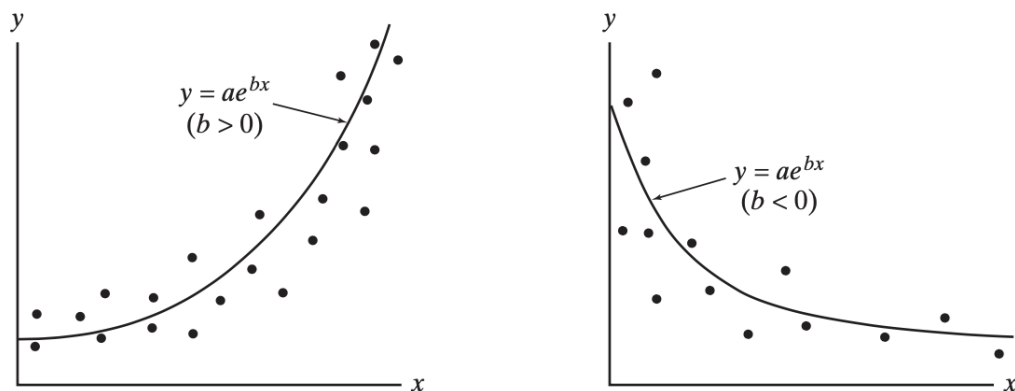


Figura 11.7: Funciones exponenciales, Retomado de Larsen y Marx (2012, p. 532) Larsen y Marx 2012

11.5.2. 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 11.8.

$$y = kx^b$$

11.5.3. 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

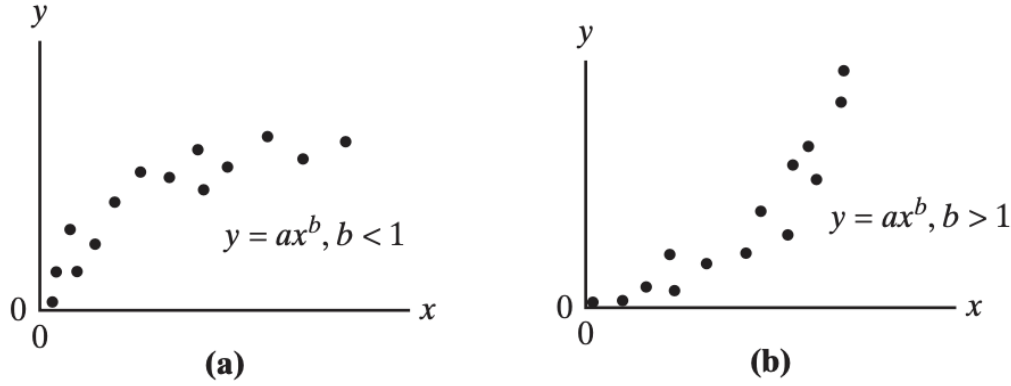


Figura 11.8: Funciones logarítmicas, Retomado de Larsen y Marx (2012, p. 536) Larsen y Marx 2012

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 la solución será como sigue, la Figura 11.9 ilustra ecuaciones como la siguiente:

$$y = \frac{1}{1 + e^{a+bx}} \quad (11.13)$$

Ahora mostremos como llegar a la solución de la ecuación logística:

$$\frac{dy}{dx} = ky(L - y) \iff \frac{dy}{y(L - y)} = kdx \iff \left(\frac{\frac{1}{L}}{y} + \frac{\frac{1}{L}}{(L - y)} \right) dy = kdx$$

$$\iff \frac{1}{L} \left(\frac{dy}{y} + \frac{dy}{(L - y)} \right) = kdx \iff \int \frac{1}{L} \left(\frac{dy}{y} + \frac{dy}{(L - y)} \right) = \int kdx$$

$$\iff \left(\int \frac{dy}{y} + \int \frac{dy}{(L - y)} \right) = \int Lkdx \iff (\ln(y) - \ln(L - y)) = Lkx + C$$

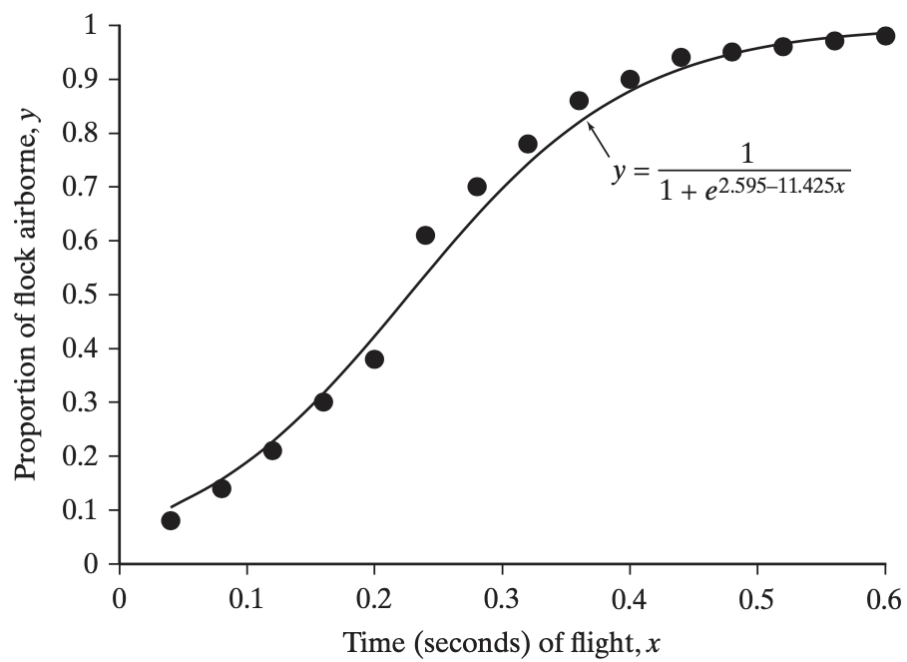


Figura 11.9: Funciones logísticas, Retomado de Larsen y Marx (2012, p. 540)
Larsen y Marx 2012

$$\Longleftrightarrow (\ln(L - y) - \ln(y)) = -LC - Lkx \Longleftrightarrow \ln\left(\frac{L - y}{y}\right) = -LC - Lkx$$

$$\Longleftrightarrow \left(\frac{L - y}{y}\right) = e^{-LC - Lkx} \Longleftrightarrow y = \frac{L}{1 + e^{-LC - Lkx}}$$

Por lo tanto, la solución general de la ecuación será:

$$y = \frac{L}{1 + e^{a+bx}}$$

Donde asumimos que la forma lineal de la ecuación logística es:

$$\ln\left(\frac{L - y}{y}\right) = a + bx$$

11.6. Modelo de regresión lineal múltiple

En esta sección ampliaremos la discusión de regresión bivariada para introducir más variables de forma que ahora supondremos una media condicional como:

$$\mu_{Y|x_1, x_2, \dots, x_K} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K = \hat{y}$$

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:

$$\{(x_{i1}, x_{i2}, \dots, x_{iK}; y_i) : i = 1, 2, \dots, n\}$$

En este caso podemos ilustrar el término de error con la Figura 11.10. Donde asumiremos que se cumplen los supuestos del modelo lineal, pero para múltiples variables independientes. De esta forma podemos establecer la función:

$$S(\beta) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

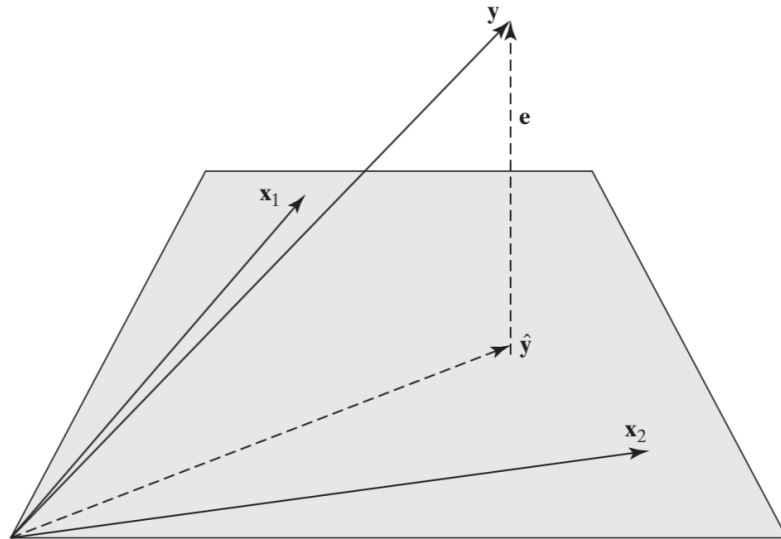


Figura 11.10: Ilustración de la proyección de y en x_1 y x_2 y el significado del error e , retomado de Greene (2012, pp. 20) Greene 2012

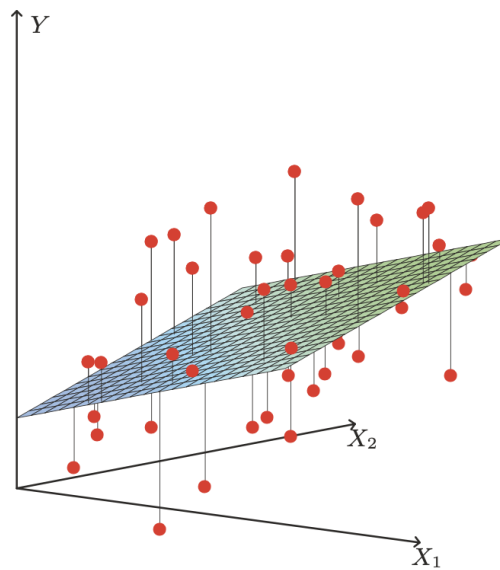


Figura 11.11: Ilustración del hiperplano generado de regesar X_1 y X_2 en Y , retomado de Hastie, Tibshirani y Friedman (2009, pp. 45) Hastie, Tibshirani y Friedman 2017

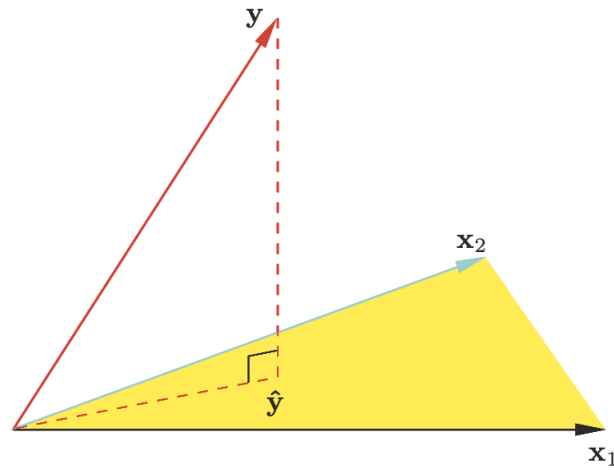


Figura 11.12: Ilustración de la proyección de y en x_1 y x_2 y el significado del error e , retomado de Hastie, Tibshirani y Friedman (2009, pp. 46) Hastie, Tibshirani y Friedman 2017

Donde:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

Entonces nuestro problema consiste en:

$$\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_K x_{iK})^2$$

Las $K + 1$ condiciones de primer orden estarán dadas por:

$$\begin{aligned}\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_K x_{iK}) \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_1} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_K x_{iK}) x_{i1} \\ &\vdots \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_K} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_K x_{iK}) x_{iK}\end{aligned}$$

De donde tenemos $K + 1$ ecuaciones normales. Resolviendo el siguiente sistema podríamos obtener los estimadores de los parámetros:

$$\begin{aligned}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_K x_{iK}) &= 0 \\ \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_K x_{iK}) x_{i1} &= 0 \\ &\vdots \\ \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_K x_{iK}) x_{iK} &= 0\end{aligned}$$

¿Cómo resolvemos este sistema? Utilicemos una alternativa. Para ello introduciremos notación matricial. Sean:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & \dots & x_{2K} \\ 1 & x_{31} & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

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:

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \\ \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} &= \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & \dots & x_{2K} \\ 1 & x_{31} & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \end{aligned}$$

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 utilizaré 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 es 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:

$$\begin{aligned}\frac{\partial S(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta}(\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta) \\ &= -2\mathbf{X}'\mathbf{Y} + 2(\mathbf{X}'\mathbf{X})'\beta \\ &= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta\end{aligned}$$

Donde hemos utilizado que la matriz $\mathbf{X}'\mathbf{X}$ es simétrica (más adelante discutimos esta condición). Finalmente, las condiciones de primer orden serán:

$$-\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

Si la inversa de la matriz $\mathbf{X}'\mathbf{X}$, entonces la solución de mínimos cuadrados será:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (11.14)$$

Para verificar si es un mínimo:

$$\frac{\partial}{\partial \beta}(-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta) = 2\mathbf{X}'\mathbf{X}$$

Donde esta ecuación es definida positiva. Ahora analicemos cada uno de los componentes de la ecuación (11.14)

$$\begin{aligned}\mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1K} & x_{2K} & x_{3K} & \dots & x_{nK} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & \dots & x_{2K} \\ 1 & x_{31} & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix} \\ &= \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \dots & \sum_{i=1}^n x_{iK} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \dots & \sum_{i=1}^n x_{i1}x_{iK} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \dots & \sum_{i=1}^n x_{i2}x_{iK} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{iK} & \sum_{i=1}^n x_{iK}x_{i1} & \sum_{i=1}^n x_{iK}x_{i2} & \dots & \sum_{i=1}^n x_{iK}^2 \end{bmatrix}\end{aligned}$$

$$\begin{aligned}
\mathbf{X}'\mathbf{Y} &= \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1K} & x_{2K} & x_{3K} & \dots & x_{nK} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{iK}y_i \end{bmatrix}
\end{aligned}$$

Ahora verifiquemos las propiedades del estimador que teníamos pendientes de la regresión bivariada. Partiendo de la ecuación (11.14), tenemos :

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|\mathbf{x}] &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{x}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{Y}|\mathbf{x}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\
&= \beta
\end{aligned}$$

Por lo tanto el estimador de mínimos cuadrados ordinarios es insesgado. Analicemos ahora la varianza:

$$\begin{aligned}
Var[\hat{\beta}|\mathbf{x}] &= \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}|\mathbf{x}])(\hat{\beta} - \mathbb{E}[\hat{\beta}|\mathbf{x}])'|\mathbf{x}] \\
&= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \beta)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \beta)'|\mathbf{x}] \\
&= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\beta' \\
&\quad - \beta\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \beta\beta'|\mathbf{x}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbb{E}[\mathbf{Y}\mathbf{Y}'|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&\quad - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbb{E}[\mathbf{Y}|\mathbf{x}] \cdot \beta' \\
&\quad - \beta \cdot \mathbb{E}[\mathbf{Y}'|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \beta\beta' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbb{E}[\mathbf{Y}\mathbf{Y}'|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta\beta' \\
&\quad - \beta\beta'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \beta\beta' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbb{E}[\mathbf{Y}\mathbf{Y}'|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - \beta\beta' - \beta\beta' + \beta\beta' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbb{E}[\mathbf{Y}\mathbf{Y}'|\mathbf{x}] \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - \beta\beta' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbb{I}_n - \mathbf{X}'\beta'\beta\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - \beta\beta' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

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 muestral 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:

$$T = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}}} \sim t_{[n-(K+1)]}$$

Donde $(\mathbf{X}'\mathbf{X})_{kk}^{-1}$ es la posición (k, k) de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. De esta forma podemos establecer una prueba de hipótesis para cada una de las β_k 's como:

$$H_0 : \hat{\beta}_k = \beta_{k,0}$$

$$H_a : \begin{cases} \hat{\beta}_k > \beta_{k,0} \\ \hat{\beta}_k < \beta_{k,0} \\ \hat{\beta}_k \neq \beta_{k,0} \end{cases}$$

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 plantear como:

$$H_0 : \mathbf{R}\hat{\boldsymbol{\beta}} = q$$

$$H_a : \begin{cases} \mathbf{R}\hat{\boldsymbol{\beta}} > q \\ \mathbf{R}\hat{\boldsymbol{\beta}} < q \\ \mathbf{R}\hat{\boldsymbol{\beta}} \neq q \end{cases}$$

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:

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \sim F_{J,n-K} \quad (11.15)$$

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 hipótesis nula de una prueba global se puede escribir como:

$$\begin{aligned} H_0 : & \quad \beta_1 = \beta_2 = \dots = \beta_K = 0 \\ H_a : & \quad \text{No } H_0 \end{aligned}$$

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 11.1. 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:

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} 1 & 3 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix} \\ \mathbf{Y} &= \begin{bmatrix} 18 \\ 16 \\ \vdots \\ 8 \end{bmatrix} \end{aligned}$$

Cuadro 11.1: 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 β seran:

$$\hat{\beta} = \begin{bmatrix} 18.3977 \\ -0.6356 \end{bmatrix}$$

Una vez estimados los parámetros podemos graficar los puntos en la muestra la recta (ver Figura 11.13).

Ejemplo. Suponga que ha estimado una ecuación de regresión del precio de las casas. Su estimación se rigue por la siguiente ecuación:

$$\ln(P) = \beta_0 + \beta_1 \ln(\text{Area}M^2) + \beta_2 \# \text{Recamaras} + \beta_3 \# \text{Banios} + \beta_4 \text{Antiguedad}$$

La estimación de dicha ecuación se reporta el en Cuadro 11.2, en el que se le pide que complete la columna 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:

$$T = \frac{\hat{\beta}_k - 0}{\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}}} = \frac{\hat{\beta}_k - 0}{\text{Error Estándar}} \sim t_{[n-(K+1)]}$$

Donde $k = 0, 1, 2, 3, 4$ y $n = 100$.

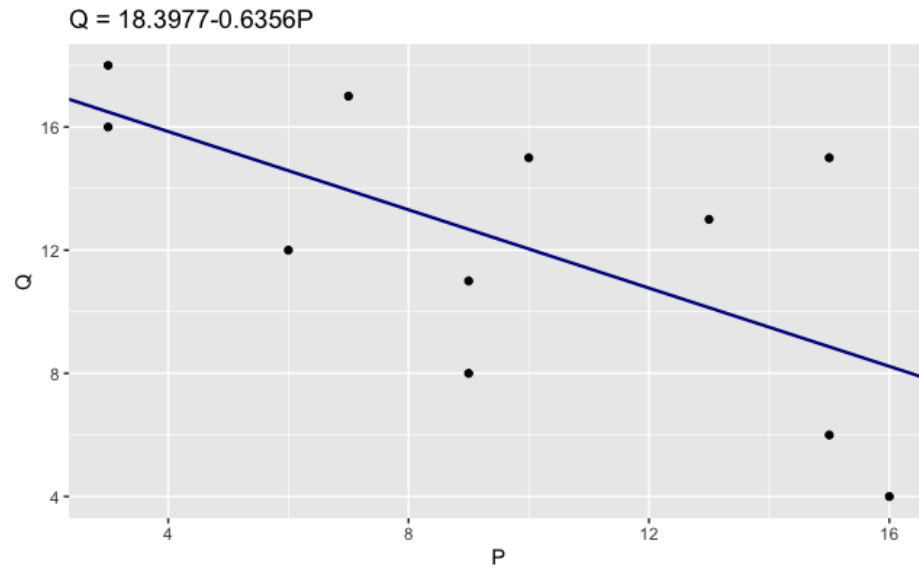


Figura 11.13: Gráfica de Precios y Cantidades, y su recta de regresión

Cuadro 11.2: 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	

12

Estadística no paramétrica

12.1. Introducción y motivación

En las secciones previas hemos hecho supuestos sobre la naturaleza de la función de densidad de probabilidad subyacente o que describe a la población de donde hemos extraído muestras de datos. Algunos de esos supuestos parecen obvios pero ¿cómo cambiarían nuestras conclusiones si esos supuestos no se cumplieran?

Ciertamente, las mismas estadísticas se podrían seguir calculando de la misma manera, así como establecer hipótesis con la misma región de rechazo. No obstante, la realidad es que el modelo estadístico ya no se sostiene.

El término, no paramétrico se puede sustituir por 'distribución libre'. En ese sentido nuestro objetivo será introducir algunas metodologías más importantes de la estadística no paramétrica para considerar situaciones en las que no hay restricciones de distribución.

12.2. Estimaciones de funciones de densidad via Kernel (Núcleo)

Este estimador fue propuesto por Rosenblatt (1956):

Definición 12.1 Sea la muestra de n observaciones X_1, X_2, \dots, X_n , donde cada $X_i \in \mathbb{R}$. Definimos la estimación de una función 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 suavizamiento es una secuencia de constantes positivas que determina el refinamiento 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 estimación, tal y como veremos en los ejemplos de clase.

La función kernel $K(x)$ 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 12.1 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 12.2 muestra la gráfica de la función kernel de Gauss.

Existen otros tipos de kernel, pero los anteriores son los más comunes.

12.3. 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.

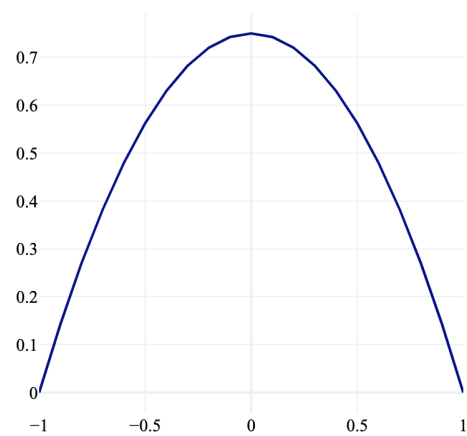


Figura 12.1: Kernel Epanechnikov

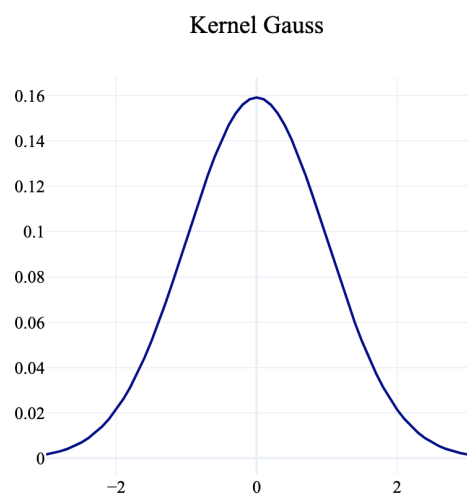


Figura 12.2: Kernel Gauss

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 observamos 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:

$$\begin{aligned}\mathbb{E}[X] &= \frac{n}{2} \\ \text{Var}(X) &= \frac{n}{4}\end{aligned}$$

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 12.1 Sean y_1, y_2, \dots, 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 simétrica, la mediana es la misma que la media, así probar si $\tilde{\mu} \neq \tilde{\mu}_0$ es equivalente a probar $\mu \neq \mu_0$. 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 12.1 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 poblacional es de \$20.30. Considerando la muestra y un nivel de confianza del 5 %, verifique:

$$H_0 : \tilde{\mu} = 20.30$$

$$H_1 : \tilde{\mu} \neq 20.30$$

Cuadro 12.1: 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 HUIITEPEC, 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 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 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 difícilmente 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:

$$p = P(X_i > Y_i), i = 1, 2, \dots, b$$

Este es el un caso particular de la prueba de signo; asumiento $p = 1/2$. Sea U una variable aleatoria binomislal 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:

$$H_0 : p = 0.5$$

$$H_1 : p \neq 0.5$$

$$z = \frac{u - \frac{b}{2}}{\sqrt{\frac{b}{4}}} = \frac{42 - \frac{64}{2}}{\sqrt{\frac{64}{4}}} = \frac{42 - 32}{\sqrt{16}} = 2.5$$

Rechazamos la hipótesis nula.

12.4. 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 mínima 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 asignaríamos 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 estadísticas, 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 típica 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 12.2. Pruebe la hipótesis de que no hay efecto entre las dos variables relacionadas.

Cuadro 12.2: 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 12.3 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 12.3: 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.

12.5. 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, \dots, y_{n_1} y z_1, z_2, \dots, 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 escalonadas calculadas para cada valor. El parámetro D viene definido según la siguiente 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 proporción 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 parámetro 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:

$$\chi^2 = 2 \cdot D^2 \left(\frac{n_1 n_2}{n_1 + n_2} \right) \sim \chi^2_{[2]}$$

13

C N

13.1. Introducción y motivación

14

Introducción al Aprendizaje Estadístico

14.1. Motivación e introducción

El aprendizaje estadístico juega un rol esencial en muchas áreas de la ciencia, finanzas y la industria. Algunos ejemplos son:

1. Predecir si un paciente—que se encuentra hospitalizado debido a un ataque al corazón—tendrá un segundo ataque. La predicción estará basada en métricas demográficas, de la dieta y de registros clínicos.
2. Predecir el precio de una acción en los siguientes 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 pide 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 susceptible 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 $i = 1, 2, \dots, n$ denota a los individuos en la muestra y $k = 1, 2, \dots, K$ denota al número de variables. Por convención diremos que $x_{i1} = 1$ para todo $i = 1, 2, \dots, n$, 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 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 serán 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 predichos.

El aprendizaje estadístico parte del establecimiento de que una variable dependiente Y es una función de un conjunto de variables explicativas $\mathbf{X} = [X_1, X_2, \dots, X_p]$. De esta forma plantearemos:

$$Y = f(\mathbf{X}) + \varepsilon \quad (14.1)$$

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:

$$\hat{Y} = \hat{f}(\mathbf{X}) \quad (14.2)$$

14.2. Modelos lineales 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 $i = 1, 2, \dots, n$ y un vector de variables independientes o explicativas X descrita de la forma:

$$\mathbf{X}_i = (x_{i1}, x_{i2}, \dots, x_{iK})$$

Así, podemos describir a la variable Y (o y_i para una observación específica) con una ecuación:

$$\hat{y}_i = \mathbf{X}_i \boldsymbol{\beta}$$

Donde $i = 1, 2, \dots, n$ 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 válidos los análisis de pruebas de hipótesis t y F descritos anteriormente.

14.3. Método de regresiones restringidas

Este tipo de métodos permite restringir los posibles valores de las estimaciones del parámetro $\boldsymbol{\beta}$ a un subconjunto seleccionado. Veámos dos casos particulares i) una regresión de cresta restringida (ridge) y ii) la regresión Lasso.

14.3.1. Regresión Ridge

La regresión de ridge restringe los coeficientes de la regresión mediante la imposición de una penalización en su magnitud. Los coeficientes estimados por este método resultan de resolver el problema:

$$\hat{\beta}^{Ridge} = \min_{\beta} \left[\sum_{i=1}^n (y_i - \mathbf{X}_i \beta)^2 + \lambda \sum_{k=2}^K \beta_k^2 \right] \quad (14.3)$$

Donde $\lambda \geq 0$ y suponemos que β_1 es el término constante de la regresión.

La idea de la penalización de los parámetros se deriva de la forma equivalente del problema en la ecuación (14.3) es:

$$\hat{\beta}^{Ridge} = \min_{\beta} \left[\sum_{i=1}^n (y_i - \mathbf{X}_i \beta)^2 \right]$$

Sujeto a:

$$\sum_{k=2}^K \beta_k^2 \leq t$$

Note que la penalización no aplica al término constante, ¿por qué?, para garantizar 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:

$$\bar{Y} = \frac{\sum_{i=1}^n y_i}{n}$$

Para los restantes coeficientes los determinaremos mediante un procedimiento dado por:

$$\min_{\beta_R, \lambda} [(\mathbf{Y} - \mathbf{X}\beta_R)'(\mathbf{Y} - \mathbf{X}\beta_R) + \lambda \beta_R' \beta_R] \quad (14.4)$$

Donde β_R contiene sólo las pendientes. Para determinar un valor estimado debemos resolver el problema descrito en la ecuación (14.4):

$$\begin{aligned} \frac{\partial}{\partial \beta_R} S(\beta_R) &= \frac{\partial}{\partial \beta_R} ((\mathbf{Y} - \mathbf{X}\beta_R)'(\mathbf{Y} - \mathbf{X}\beta_R) + \lambda \beta_R' \beta_R) \\ &= \frac{\partial}{\partial \beta_R} (\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\beta_R + \beta_R' \mathbf{X}'\mathbf{X}\beta_R + \lambda \beta_R' \beta_R) \\ &= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta_R + 2\lambda \beta_R \end{aligned}$$

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:

$$\hat{\boldsymbol{\beta}}_R^{Ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbb{I}_{K-1})^{-1}\mathbf{X}'\mathbf{Y} \quad (14.5)$$

Así, el problema de la estimación es que tiene 1 grado de libertad, λ , que es un valor de penalización y que resulta arbitrario.

14.3.2. Regresión Lasso (Least Absolute Shrinkage and Selection Operator)

La regresión Lasso se define por la solución al problema:

$$\hat{\boldsymbol{\beta}}^{Lasso} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^n (y_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right] \quad (14.6)$$

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:

$$\bar{Y} = \frac{\sum_{i=1}^n y_i}{n}$$

La idea de la penalización de los parámetros se deriva de la forma equivalente del problema en la ecuación (14.6) es:

$$\hat{\boldsymbol{\beta}}^{Lasso} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^n (y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda \sum_{k=2}^K |\beta_k| \right] \quad (14.7)$$

14.4. 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 $j = \{0, 1, 2, \dots, J\}$. Este tipo de modelos se les conoce como discriminante por probabilidades:

$$\mathbb{P}(G = j|X = x) \quad (14.8)$$

De esta forma, supongamos un problema de determinar la probabilidad de dos clases $j = \{1, 2\}$, para cada un tendríamos:

$$\begin{aligned} \mathbb{P}(G = 1|X = x) &= \frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}} \\ \mathbb{P}(G = 2|X = x) &= \frac{1}{1 + e^{\mathbf{X}\beta}} \end{aligned}$$

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:

$$\ln \left\{ \frac{\mathbb{P}(G = 1|X = x)}{\mathbb{P}(G = 2|X = x)} \right\} = \mathbf{X}\beta \quad (14.9)$$

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 columna incluye al término constante. Consideremos un caso generalizado en el que la clases o valores que puede tomar $G(X)$ son $j = 0, 1, 2, \dots, J$. 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:

$$\mathbb{P}(G = j|\mathbf{X} = \mathbf{x}) = \frac{e^{\mathbf{x}\beta_j}}{1 + \sum_{h=1}^J e^{\mathbf{x}\beta_h}} \quad (14.10)$$

Donde $j = 1, 2, \dots, J$ y β_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:

$$\mathbb{P}(G = 0|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + \sum_{h=1}^J e^{\mathbf{x}\beta_h}} \quad (14.11)$$

Estimación

Los modelos de regresión logístca soos estimados usualmente por métodos de Máxima Verosimilitud mediante el uso de $\mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i)$, para $j = 0, 1, \dots, J$ e $i = 1, 2, \dots, n$. Así, plateamos la función de verosimilitud como:

$$L(\theta) = \prod_{i=1}^n \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \quad (14.12)$$

Utilizando una transformación logarítmica:

$$\ln L(\theta) = l(\theta) = \sum_{i=1}^n \ln \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \quad (14.13)$$

Por simplicidad continuaremos con el caso de una estimación para un modelo de respuesta binomial en el cual $G = \{0, 1\}$. De esta forma establezcamos la función de verosimilitud como:

$$\begin{aligned} \ln L(\theta) = l(\theta) &= \sum_{i=1}^n G_i \ln \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i) \\ &\quad + (1 - G_i) \ln (1 - \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i)) \\ &= \sum_{i=1}^n G_i \mathbf{X}_i \beta - \ln(1 + e^{\mathbf{X}_i \beta}) \end{aligned}$$

Las condiciones de primer orden son:

$$\frac{\partial l(\theta)}{\partial \beta} = \sum_{i=1}^n \left(G_i \mathbf{X}_i - \mathbf{X}_i \frac{e^{\mathbf{X}_i \beta}}{1 + e^{\mathbf{X}_i \beta}} \right) = (G_i - \mathbb{P}(G_i = j|\mathbf{X} = \mathbf{x}_i)) \mathbf{X}_i \quad (14.14)$$

Así, la solución estará dada por aquellas β que satisfacen:

$$\sum_{i=1}^n \left(G_i - \frac{e^{\mathbf{X}_i \beta}}{1 + e^{\mathbf{X}_i \beta}} \right) \mathbf{X}_i = 0 \quad (14.15)$$



Figura 14.1: 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

Supongamos 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:

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{error cuadrático} \\ |Y - \hat{f}(X)| & \text{error absoluto} \end{cases} \quad (14.16)$$

De esta forma podemos establecer un error cuadrático como:

$$\begin{aligned} \mathbb{E}[Y - \hat{Y}]^2 &= \mathbb{E}[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= \mathbb{E}[f(X) - \hat{f}(X)]^2 + \text{Var}(\varepsilon) \end{aligned}$$

Donde $\mathbb{E}[f(X) - \hat{f}(X)]^2$ es el único componente reducible. Para hacer predicciones requerimos de un conjunto de datos de entrenamiento y otro más de prueba – en el primero estimamos $f(\cdot)$ y en 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 14.1.

14.5. Aprendizaje no supervisado

14.6. Otros

El Cuadro (14.1) muestra los resultados para este nuevo modelo. No lo mostramos en esta sección, pero ambos modelos reportados tienen raíces de sus respectivos polinomios característicos menores a 1 en valor absoluto.

Cuadro 14.1: 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.
- [Rin12] Luis Rincón. *Curso Elemental de Probabilidad y Estadística*. México: Sociedad Matemática Mexicana, 2012.
- [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