Notas Elementales de Regresión Logística y Análisis de Supervivencia

Carlos E. Martínez-Rodríguez

Julio 2024

Índice general

Parte I

PRIMERA PARTE: REGRESIÓN LOGÍSTICA

Capítulo 1

Introducción

1.1. Introducción

La Estadística es una ciencia formal que estudia la recolección, análisis e interpretación de datos de una muestra representativa, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional. Sin embargo, la estadística es más que eso, es decir, es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad. Se usa para la toma de decisiones en áreas de negocios o instituciones gubernamentales. Ahora bien, las técnicas estadísticas se aplican de manera amplia en mercadotecnia, contabilidad, control de calidad y en otras actividades; estudios de consumidores; análisis de resultados en deportes; administradores de instituciones; en la educación; organismos políticos; médicos; y por otras personas que intervienen en la toma de decisiones.

Definición 1.1 *La Estadística es la ciencia cuyo objetivo es reunir una información cuantitativa concerniente a individuos, grupos, series de hechos, etc. y deducir de ello gracias al análisis de estos datos unos significados precisos o unas previsiones para el futuro.*

La estadística, en general, es la ciencia que trata de la recopilación, organización presentación, análisis e interpretación de datos numéricos con el fin de realizar una toma de decisión más efectiva. Los métodos estadísticos tradicionalmente se utilizan para propósitos descriptivos, para organizar y resumir datos numéricos. La estadística descriptiva, por ejemplo trata de la tabulación de datos, su presentación en forma gráfica o ilustrativa y el cálculo de medidas descriptivas.

1.1.1. Historia de la Estadística

Es difícil conocer los orígenes de la Estadística. Desde los comienzos de la civilización han existido formas sencillas de estadística, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o ciertas cosas. Su origen empieza posiblemente en la isla de Cerdeña, donde existen monumentos prehistóricos pertenecientes a los Nuragas, las primeros habitantes de la isla; estos monumentos constan de bloques de basalto superpuestos sin mortero y en cuyas paredes de encontraban grabados toscos signos que han sido interpretados con mucha verosimilidad como muescas que servían para llevar la cuenta del ganado y la caza. Los babilonios usaban ya pequeñas tablillas de arcilla para recopilar datos en tablas sobre la producción agrícola y los géneros vendidos o cambiados mediante trueque. Otros vestigios pueden ser hallados en el antiguo Egipto, cuyos faraones lograron recopilar, hacia el año 3050 antes de Cristo, prolijos datos relativos a la población y la riqueza del país. De acuerdo al historiador griego Heródoto, dicho registro de riqueza y población se hizo con el objetivo de preparar la construcción de las pirámides. En el mismo Egipto, Ramsés II hizo un censo de las tierras con el objeto de verificar un nuevo reparto. En el antiguo Israel la Biblia da referencias, en el libro de los Números, de los datos estadísticos obtenidos en dos recuentos de la población hebrea. El rey David por otra parte, ordenó a Joab, general del ejército hacer un censo de Israel con la finalidad de conocer el número de la población. También los chinos efectuaron

censos hace más de cuarenta siglos. Los griegos efectuaron censos periódicamente con fines tributarios, sociales (división de tierras) y militares (cálculo de recursos y hombres disponibles). La investigación histórica revela que se realizaron 69 censos para calcular los impuestos, determinar los derechos de voto y ponderar la potencia guerrera.

Fueron los romanos, maestros de la organización política, quienes mejor supieron emplear los recursos de la estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos tenían la obligación de anotar nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas. Para el nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del imperio. Durante los mil años siguientes a la caída del imperio Romano se realizaron muy pocas operaciones Estadísticas, con la notable excepción de las relaciones de tierras pertenecientes a la Iglesia, compiladas por Pipino el Breve en el 758 y por Carlomagno en el 762 DC. Durante el siglo IX se realizaron en Francia algunos censos parciales de siervos. En Inglaterra, Guillermo el Conquistador recopiló el Domesday Book o libro del Gran Catastro para el año 1086, un documento de la propiedad, extensión y valor de las tierras de Inglaterra. Esa obra fue el primer compendio estadístico de Inglaterra. Aunque Carlomagno, en Francia; y Guillermo el Conquistador, en Inglaterra, trataron de revivir la técnica romana, los métodos estadísticos permanecieron casi olvidados durante la Edad Media.

Durante los siglos XV, XVI, y XVII, hombres como Leonardo de Vinci, Nicolás Copérnico, Galileo, Neper, William Harvey, Sir Francis Bacon y René Descartes, hicieron grandes operaciones al método científico, de tal forma que cuando se crearon los Estados Nacionales y surgió como fuerza el comercio internacional existía ya un método capaz de aplicarse a los datos económicos. Para el año 1532 empezaron a registrarse en Inglaterra las defunciones debido al temor que Enrique VII tenía por la peste. Más o menos por la misma época, en Francia la ley exigió a los clérigos registrar los bautismos, fallecimientos y matrimonios. Durante un brote de peste que apareció a fines de la década de 1500, el gobierno inglés comenzó a publicar estadística semanales de los decesos. Esa costumbre continuó muchos años, y en 1632 estos Bills of Mortality (Cuentas de Mortalidad) contenían los nacimientos y fallecimientos por sexo. En 1662, el capitán John Graunt usó documentos que abarcaban treinta años y efectuó predicciones sobre el número de personas que morirían de varias enfermedades y sobre las proporciones de nacimientos de varones y mujeres que cabría esperar. El trabajo de Graunt, condensado en su obra *Natural and Political Observations...Made upon the Bills of Mortality*, fue un esfuerzo innovador en el análisis estadístico. Por el año 1540 el alemán Sebastián Muster realizó una compilación estadística de los recursos nacionales, comprensiva de datos sobre organización política, instrucciones sociales, comercio y poderío militar.

Los eruditos del siglo XVII demostraron especial interés por la Estadística Demográfica como resultado de la especulación sobre si la población aumentaba, decrecía o permanecía estática. En los tiempos modernos tales métodos fueron resucitados por algunos reyes que necesitaban conocer las riquezas monetarias y el potencial humano de sus respectivos países. El primer empleo de los datos estadísticos para fines ajenos a la política tuvo lugar en 1691 y estuvo a cargo de Gaspar Neumann, un profesor alemán que vivía en Breslau. Este investigador se propuso destruir la antigua creencia popular de que en los años terminados en siete moría más gente que en los restantes, y para lograrlo hurgó pacientemente en los archivos parroquiales de la ciudad. Después de revisar miles de partidas de defunción pudo demostrar que en tales años no fallecían más personas que en los demás. Los procedimientos de Neumann fueron conocidos por el astrónomo inglés Halley, descubridor del cometa que lleva su nombre, quien los aplicó al estudio de la vida humana.

Durante el siglo XVII y principios del XVIII, matemáticos como Bernoulli, Francis Maseres, Lagrange y Laplace desarrollaron la teoría de probabilidades. No obstante durante cierto tiempo, la teoría de las probabilidades limitó su aplicación a los juegos de azar y hasta el siglo XVIII no comenzó a aplicarse a los grandes problemas científicos. Godofredo Achenwall, profesor de la Universidad de Gotinga, acuñó en 1760 la palabra estadística, que extrajo del término italiano statista (estadista). Creía, y con sobrada razón, que los datos de la nueva ciencia serían el aliado más eficaz del gobernante consciente. La raíz remota de la palabra se halla, por otra parte, en el término latino status, que significa estado o situación; Esta etimología aumenta el valor intrínseco de la palabra, por cuanto la estadística revela el sentido cuantitativo de las más variadas situaciones. Jacques Quételect es quien aplica las Estadísticas a las ciencias sociales. Este interpretó la teoría de la probabilidad para su uso en las ciencias sociales y resolver la aplicación del principio de promedios y de la variabilidad a los fenómenos sociales. Quételect fue el primero en realizar la aplicación práctica de todo el método Estadístico, entonces conocido, a las diversas ramas de la ciencia. Entretanto, en el período del 1800 al 1820 se desarrollaron dos conceptos matemáticos

fundamentales para la teoría Estadística; la teoría de los errores de observación, aportada por Laplace y Gauss; y la teoría de los mínimos cuadrados desarrollada por Laplace, Gauss y Legendre. A finales del siglo XIX, Sir Francis Gaston ideó el método conocido por Correlación, que tenía por objeto medir la influencia relativa de los factores sobre las variables. De aquí partió el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica como J. Pease Norton, R. H. Hooker y G. Udny Yule, que efectuaron amplios estudios sobre la medida de las relaciones.

La historia de la estadística está resumida en tres grandes etapas o fases.

- **Fase 1: Los Censos:** Desde el momento en que se constituye una autoridad política, la idea de inventariar de una forma más o menos regular la población y las riquezas existentes en el territorio está ligada a la conciencia de soberanía y a los primeros esfuerzos administrativos.
- **Fase 2: De la Descripción de los Conjuntos a la Aritmética Política:** Las ideas mercantilistas extrañan una intensificación de este tipo de investigación. Colbert multiplica las encuestas sobre artículos manufacturados, el comercio y la población: los intendentes del Reino envían a París sus memorias. Vauban, más conocido por sus fortificaciones o su Dime Royale, que es la primera propuesta de un impuesto sobre los ingresos, se señala como el verdadero precursor de los sondeos. Más tarde, Bufón se preocupa de esos problemas antes de dedicarse a la historia natural. La escuela inglesa proporciona un nuevo progreso al superar la fase puramente descriptiva.

  Sus tres principales representantes son Graunt, Petty y Halley. El penúltimo es autor de la famosa Aritmética Política. Chaptal, ministro del interior francés, publica en 1801 el primer censo general de población, desarrolla los estudios industriales, de las producciones y los cambios, haciéndose sistemáticos durantes las dos terceras partes del siglo XIX.
- **Fase 3: Estadística y Cálculo de Probabilidades:** El cálculo de probabilidades se incorpora rápidamente como un instrumento de análisis extremadamente poderoso para el estudio de los fenómenos económicos y sociales y en general para el estudio de fenómenos cuyas causas son demasiados complejas para conocerlos totalmente y hacer posible su análisis.

La Estadística para su mejor estudio se ha dividido en dos grandes ramas: **la Estadística Descriptiva y la Estadística Inferencial**.

- **Descriptiva:** consiste sobre todo en la presentación de datos en forma de tablas y gráficas. Esta comprende cualquier actividad relacionada con los datos y está diseñada para resumir o describir los mismos sin factores pertinentes adicionales; esto es, sin intentar inferir nada que vaya más allá de los datos, como tales.
- **Inferencial:** se deriva de muestras, de observaciones hechas sólo acerca de una parte de un conjunto numeroso de elementos y esto implica que su análisis requiere de generalizaciones que van más allá de los datos. Como consecuencia, la característica más importante del reciente crecimiento de la estadística ha sido un cambio en el énfasis de los métodos que describen a métodos que sirven para hacer generalizaciones. La Estadística Inferencial investiga o analiza una población partiendo de una muestra tomada.

Estadística Inferencial

Los métodos básicos de la estadística inferencial son la estimación y el contraste de hipótesis, que juegan un papel fundamental en la investigación. Por tanto, algunos de los objetivos que se persiguen son:

- Calcular los parámetros de la distribución de medias o proporciones muestrales de tamaño n, extraídas de una población de media y varianza conocidas.
- Estimar la media o la proporción de una población a partir de la media o proporción muestral.
- Utilizar distintos tamaños muestrales para controlar la confianza y el error admitido.
- Contrastar los resultados obtenidos a partir de muestras.

- Visualizar gráficamente, mediante las respectivas curvas normales, las estimaciones realizadas.

En definitiva, la idea es, a partir de una población se extrae una muestra por algunos de los métodos existentes, con la que se generan datos numéricos que se van a utilizar para generar estadísticos con los que realizar estimaciones o contrastes poblacionales. Existen dos formas de estimar parámetros: la *estimación puntual* y la *estimación por intervalo de confianza*. En la primera se busca, con base en los datos muestrales, un único valor estimado para el parámetro. Para la segunda, se determina un intervalo dentro del cual se encuentra el valor del parámetro, con una probabilidad determinada.

Si el objetivo del tratamiento estadístico inferencial, es efectuar generalizaciones acerca de la estructura, composición o comportamiento de las poblaciones no observadas, a partir de una parte de la población, será necesario que la proporción de población examinada sea representativa del total. Por ello, la selección de la muestra requiere unos requisitos que lo garanticen, debe ser representativa y aleatoria.

Además, la cantidad de elementos que integran la muestra (el tamaño de la muestra) depende de múltiples factores, como el dinero y el tiempo disponibles para el estudio, la importancia del tema analizado, la confiabilidad que se espera de los resultados, las características propias del fenómeno analizado, etcétera.

Así, a partir de la muestra seleccionada se realizan algunos cálculos y se estima el valor de los parámetros de la población tales como la media, la varianza, la desviación estándar, o la forma de la distribución, etc.

El conjunto de los métodos que se utilizan para medir las características de la información, para resumir los valores individuales, y para analizar los datos a fin de extraerles el máximo de información, es lo que se llama *métodos estadísticos*. Los métodos de análisis para la información cuantitativa se pueden dividir en los siguientes seis pasos:

- Definición del problema.
- Recopilación de la información existente.
- Obtención de información original.
- Clasificación.
- Presentación.
- Análisis.

El centro de gravedad de la metodología estadística se empieza a desplazar técnicas de computación intensiva aplicadas a grandes masas de datos, y se empieza a considerar el método estadístico como un proceso iterativo de búsqueda del modelo ideal. Las aplicaciones en este periodo de la Estadística a la Economía conducen a una disciplina con contenido propio: la Econometría. La investigación estadística en problemas militares durante la segunda guerra mundial y los nuevos métodos de programación matemática, dan lugar a la Investigación Operativa. El tratamiento de los datos de la investigación científica tiene varias etapas:

- En la etapa de recolección de datos del método científico, se define a la población de interés y se selecciona una muestra o conjunto de personas representativas de la misma, se realizan experimentos o se emplean instrumentos ya existentes o de nueva creación, para medir los atributos de interés necesarios para responder a las preguntas de investigación. Durante lo que es llamado trabajo de campo se obtienen los datos en crudo, es decir las respuestas directas de los sujetos uno por uno, se codifican (se les asignan valores a las respuestas), se capturan y se verifican para ser utilizados en las siguientes etapas.
- En la etapa de recuento, se organizan y ordenan los datos obtenidos de la muestra. Esta será descrita en la siguiente etapa utilizando la estadística descriptiva, todas las investigaciones utilizan estadística descriptiva, para conocer de manera organizada y resumida las características de la muestra.
- En la etapa de análisis se utilizan las pruebas estadísticas (estadística inferencial) y en la interpretación se acepta o rechaza la hipótesis nula.

Niveles de medición y tipos de variables

Para poder emplear el método estadístico en un estudio es necesario medir las variables.

- Medir: es asignar valores a las propiedades de los objetos bajo ciertas reglas, esas reglas son los niveles de medición.
- Cuantificar: es asignar valores a algo tomando un patrón de referencia. Por ejemplo, cuantificar es ver cuántos hombres y cuántas mujeres hay.

Variable: es una característica o propiedad que asume diferentes valores dentro de una población de interés y cuya variación es susceptible de medirse.

Las variables pueden clasificarse de acuerdo al tipo de valores que puede tomar como:

- **Discretas o categóricas** en las que los valores se relacionan a nombres, etiquetas o categorías, no existe un significado numérico directo.
- **Continuas** los valores tienen un correlato numérico directo, son continuos y susceptibles de fraccionarse y de poder utilizarse en operaciones aritméticas.
- **Dicotómica** sólo tienen dos valores posibles, la característica está ausente o presente.

En cuanto a una clasificación estadística, las varibles pueden ser:

- **Aleatoria** Aquella en la cual desconocemos el valor porque fluctúa de acuerdo a un evento debido al azar.
- **Determinística** Aquella variable de la que se conoce el valor.
- **Independiente** aquellas variables que son manipuladas por el investigador. Define los grupos.
- **Dependiente** son mediciones que ocurren durante el experimento o tratamiento (resultado de la independiente), es la que se mide y compara entre los grupos.

En lo que tiene que ver con los **Niveles de Medición** tenemoss distintos tipos de variable

- **Nominal:** Las propiedades de la medición nominal son:
  - Exhaustiva: implica a todas las opciones.
  - A los sujetos se les asignan categorías, por lo que son mutuamente excluyentes. Es decir, la variable está presente o no; tiene o no una característica.
- **Ordinal:** Las propiedades de la medición ordinal son:
  - El nivel ordinal posee transitividad, por lo que se tiene la capacidad de identificar que es mejor o mayor que otra, en ese sentido se pueden establecer jerarquías.
  - Las distancias entre un valor y otro no son iguales.
- **Intervalo:**
  - El nivel de medición intervalar requiere distancias iguales entre cada valor. Por lo general utiliza datos cuantitativos. Por ejemplo: temperatura, atributos psicológicos (CI, nivel de autoestima, pruebas de conocimientos, etc.)
  - Las unidades de calificación son equivalentes en todos los puntos de la escala. Una escala de intervalos implica: clasificación, magnitud y unidades de tamaños iguales (Brown, 2000).
  - Se pueden hacer operaciones aritméticas.
  - Cuando se le pide al sujeto que califique una situación del 0 al 10 puede tomarse como un nivel de medición de intervalo, siempre y cuando se incluya el 0.
- **Razón:**
  - La escala empieza a partir del 0 absoluto, por lo tanto incluye sólo los números por su valor en sí, por lo que no pueden existir los números con signo negativo. Por ejemplo: Peso corporal en kg., edad en años, estatura en cm.

Definiciones adicionales

- **Variable:** Consideraciones que una variable son una característica o fenómeno que puede tomar distintos valores.
- **Dato:** Mediciones o cualidades que han sido recopiladas como resultado de observaciones.
- **Población:** Se considera el área de la cual son extraídos los datos. Es decir, es el conjunto de elementos o individuos que poseen una característica común y medible acerca de lo cual se desea información. Es también llamado Universo.
- **Muestra:** Es un subconjunto de la población, seleccionado de acuerdo a una regla o algún plan de muestreo.
- **Censo:** Recopilación de todos los datos (de interés para la investigación) de la población.
- **Estadística:** Es una función o fórmula que depende de los datos de la muestra (es variable).
- **Parámetro:** Característica medible de la población. Es un resumen numérico de alguna variable observada de la población. Los parámetros normales que se estudian son: *La media poblacional, Proporción.*
- **Estimador:** Un estimador de un parámetro es un estadístico que se emplea para conocer el parámetro desconocido.
- **Estadístico:** Es una función de los valores de la muestra. Es una variable aleatoria, cuyos valores dependen de la muestra seleccionada. Su distribución de probabilidad, se conoce como *Distribución muestral del estadístico.*
- **Estimación:** Este término indica que a partir de lo observado en una muestra (un resumen estadístico con las medidas que conocemos de Descriptiva) se extrapola o generaliza dicho resultado muestral a la población total, de modo que lo estimado es el valor generalizado a la población. Consiste en la búsqueda del valor de los parámetros poblacionales objeto de estudio. Puede ser puntual o por intervalo de confianza:
  - ***Puntual:*** cuando buscamos un valor concreto. Un estimador de un parámetro poblacional es una función de los datos muestrales. En pocas palabras, es una fórmula que depende de los valores obtenidos de una muestra, para realizar estimaciones. Lo que se pretende obtener es el valor exacto de un parámetro.
  - ***Intervalo de confianza:*** cuando determinamos un intervalo, dentro del cual se supone que va a estar el valor del parámetro que se busca con una cierta probabilidad. El intervalo de confianza está determinado por dos valores dentro de los cuales afirmamos que está el verdadero parámetro con cierta probabilidad. Son unos límites o margen de variabilidad que damos al valor estimado, para poder afirmar, bajo un criterio de probabilidad, que el verdadero valor no los rebasará.

    Este intervalo contiene al parámetro estimado con una determinada certeza o nivel de confianza.

  En la estimación por intervalos se usan los siguientes conceptos:
- **Variabilidad del parámetro:** Si no se conoce, puede obtenerse una aproximación en los datos o en un estudio piloto. También hay métodos para calcular el tamaño de la muestra que prescinden de este aspecto. Habitualmente se usa como medida de esta variabilidad la desviación típica poblacional.
- **Error de la estimación:** Es una medida de su precisión que se corresponde con la amplitud del intervalo de confianza. Cuanta más precisión se desee en la estimación de un parámetro, más estrecho deberá ser el intervalo de confianza y, por tanto, menor el error, y más sujetos deberán incluirse en la muestra estudiada.
- **Nivel de confianza:** Es la probabilidad de que el verdadero valor del parámetro estimado en la población se sitúe en el intervalo de confianza obtenido. El nivel de confianza se denota por $1 - \alpha$

- **p-value:** También llamado nivel de significación. Es la probabilidad (en tanto por uno) de fallar en nuestra estimación, esto es, la diferencia entre la certeza (1) y el nivel de confianza $1-\alpha$.
- **Valor crítico:** Se representa por $Z_{\alpha/2}$. Es el valor de la abscisa en una determinada distribución que deja a su derecha un área igual a $1/2$, siendo $1-\alpha$ el nivel de confianza. Normalmente los valores críticos están tabulados o pueden calcularse en función de la distribución de la población.

Para un tamaño fijo de la muestra, los conceptos de error y nivel de confianza van relacionados. Si admitimos un error mayor, esto es, aumentamos el tamaño del intervalo de confianza, tenemos también una mayor probabilidad de éxito en nuestra estimación, es decir, un mayor nivel de confianza. Por tanto, un aspecto que debe de tenerse en cuenta es el tamaño muestral, ya que para disminuir el error que se comente habrá que aumentar el tamaño muestral. Esto se resolverá, para un intervalo de confianza cualquiera, despejando el tamaño de la muestra en cualquiera de las formulas de los intervalos de confianza que veremos a continuación, a partir del error máximo permitido. Los intervalos de confianza pueden ser unilaterales o bilaterales:

- **Contraste de Hipótesis:** Consiste en determinar si es aceptable, partiendo de datos muestrales, que la característica o el parámetro poblacional estudiado tome un determinado valor o esté dentro de unos determinados valores.
- **Nivel de Confianza:** Indica la proporción de veces que acertaríamos al afirmar que el parámetro está dentro del intervalo al seleccionar muchas muestras.

1.1.2. Muestreo:

Muestreo: Una muestra es representativa en la medida que es imagen de la población. En general, podemos decir que el tamaño de una muestra dependerá principalmente de: *Nivel de precisión deseado, Recursos disponibles, Tiempo involucrado en la investigación.* Además el plan de muestreo debe considerar *La población, Parámetros a medir.* Existe una gran cantidad de tipos de muestreo, en la práctica los más utilizados son los siguientes:

- **MUESTREO ALEATORIO SIMPLE:** Es un método de selección de n unidades extraídas de N, de tal manera que cada una de las posibles muestras tiene la misma probabilidad de ser escogida. (En la práctica, se enumeran las unidades de 1 a N, y a continuación se seleccionan n números aleatorios entre 1 y N, ya sea de tablas o de alguna urna con fichas numeradas).
- **MUESTREO ESTRATIFICADO ALEATORIO:** Se usa cuando la población está agrupada en pocos estratos, cada uno de ellos son muchas entidades. Este muestreo consiste en sacar una muestra aleatoria simple de cada uno de los estratos. (Generalmente, de tamaño proporcional al estrato).
- **MUESTREO SISTEMÁTICO:** Se utiliza cuando las unidades de la población están de alguna manera totalmente ordenadas. Para seleccionar una muestra de n unidades, se divide la población en n subpoblaciones de tamaño $K = N/n$ y se toma al azar una unidad de la K primeras y de ahí en adelante cada K-ésima unidad.
- **MUESTREO POR CONGLOMERADO:** Se emplea cuando la población está dividida en grupos o conglomerados pequeños. Consiste en obtener una muestra aleatoria simple de conglomerados y luego CENSAR cada uno de éstos.
- **MUESTREO EN DOS ETAPAS (Bietápico):** En este caso la muestra se toma en dos pasos:
  - Seleccionar una muestra de unidades primarias, y
  - Seleccionar una muestra de elementos a partir de cada unidad primaria escogida.
  - *Observación:* En la realidad es posible encontrarse con situaciones en las cuales no es posible aplicar libremente un tipo de muestreo, incluso estaremos obligados a mezclarlas en ocasiones.

1.1.3. Errores Estadísticos Comunes

El propósito de esta sección es solamente indicar los malos usos comunes de datos estadísticos, sin incluir el uso de métodos estadísticos complicados. Un estudiante debería estar alerta en relación con estos malos usos y debería hacer un gran esfuerzo para evitarlos a fin de ser un verdadero estadístico.

Datos estadísticos inadecuados: Los datos estadísticos son usados como la materia prima para un estudio estadístico. Cuando los datos son inadecuados, la conclusión extraída del estudio de los datos se vuelve obviamente inválida. Por ejemplo, supongamos que deseamos encontrar el ingreso familiar típico del año pasado en la ciudad Y de 50,000 familias y tenemos una muestra consistente del ingreso de solamente tres familias: 1 millón, 2 millones y no ingreso. Si sumamos el ingreso de las tres familias y dividimos el total por 3, obtenemos un promedio de 1 millón. Entonces, extraemos una conclusión basada en la muestra de que el ingreso familiar promedio durante el año pasado en la ciudad fue de 1 millón. Es obvio que la conclusión es falsa, puesto que las cifras son extremas y el tamaño de la muestra es demasiado pequeño; por lo tanto la muestra no es representativa.

Hay muchas otras clases de datos inadecuados. Por ejemplo, algunos datos son respuestas inexactas de una encuesta, porque las preguntas usadas en la misma son vagas o engañosas, algunos datos son toscas estimaciones porque no hay disponibles datos exactos o es demasiado costosa su obtención, y algunos datos son irrelevantes en un problema dado, porque el estudio estadístico no está bien planeado. Al momento de recopilar los datos que serán procesados se es susceptible de cometer errores así como durante los cómputos de los mismos. No obstante, hay otros errores que no tienen nada que ver con la digitación y que no son tan fácilmente identificables. Algunos de éstos errores son:

- **Sesgo:** Es imposible ser completamente objetivo o no tener ideas preconcebidas antes de comenzar a estudiar un problema, y existen muchas maneras en que una perspectiva o estado mental pueda influir en la recopilación y en el análisis de la información. En estos casos se dice que hay un sesgo cuando el individuo da mayor peso a los datos que apoyan su opinión que a aquellos que la contradicen. Un caso extremo de sesgo sería la situación donde primero se toma una decisión y después se utiliza el análisis estadístico para justificar la decisión ya tomada.
- **Datos No Comparables:** el establecer comparaciones es una de las partes más importantes del análisis estadístico, pero es extremadamente importante que tales comparaciones se hagan entre datos que sean comparables.
- **Proyección descuidada de tendencias:** la proyección simplista de tendencias pasadas hacia el futuro es uno de los errores que más ha desacreditado el uso del análisis estadístico.
- **Muestreo Incorrecto:** en la mayoría de los estudios sucede que el volumen de información disponible es tan inmenso que se hace necesario estudiar muestras, para derivar conclusiones acerca de la población a que pertenece la muestra. Si la muestra se selecciona correctamente, tendrá básicamente las mismas propiedades que la población de la cual fue extraída; pero si el muestreo se realiza incorrectamente, entonces puede suceder que los resultados no signifiquen nada.

Sesgo significa que un usuario dé los datos perjudicialmente de más énfasis a los hechos, los cuales son empleados para mantener su predeterminada posición u opinión. Los estadísticos son frecuentemente degradados por lemas tales como: *Hay tres clases de mentiras: mentiras, mentiras reprobables y estadística, y Las cifras no mienten, pero los mentirosos piensan.* Hay dos clases de sesgos: conscientes e inconscientes. Ambos son comunes en el análisis estadístico. Hay numerosos ejemplos de sesgos conscientes. Un anunciante frecuentemente usa la estadística para probar que su producto es muy superior al producto de su competidor. Un político prefiere usar la estadística para sostener su punto de vista. Gerentes y líderes de trabajadores pueden simultáneamente situar sus respectivas cifras estadísticas sobre la misma tabla de trato para mostrar que sus rechazos o peticiones son justificadas. Es casi imposible que un sesgo inconsciente esté completamente ausente en un trabajo estadístico. En lo que respecta al ser humano, es difícil obtener una actitud completamente objetiva al abordar un problema, aun cuando un científico debería tener una mente abierta. Un estadístico debería estar enterado del hecho de que su interpretación de los resultados del análisis estadístico está influenciado por su propia experiencia, conocimiento y antecedentes con relación al problema dado.

1.2. Estadística Descriptiva

La estadística descriptiva resume y describe las características de un conjunto de datos. Incluye medidas de tendencia central, medidas de dispersión y medidas de forma.

1.2.1. Medidas de Tendencia Central

Las medidas de tendencia central incluyen la media, la mediana y la moda.

Media

La media aritmética es la suma de los valores dividida por el número de valores:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

donde x_i son los valores de la muestra y n es el tamaño de la muestra.

Mediana

La mediana es el valor medio cuando los datos están ordenados. Si el número de valores es impar, la mediana es el valor central. Si es par, es el promedio de los dos valores centrales.

Moda

La moda es el valor que ocurre con mayor frecuencia en un conjunto de datos.

1.2.2. Medidas de Dispersión

Las medidas de dispersión incluyen el rango, la varianza y la desviación estándar.

Rango

El rango es la diferencia entre el valor máximo y el valor mínimo de los datos:

$$Rango = x_{\max} - x_{\min}$$

Varianza

La varianza es la media de los cuadrados de las diferencias entre los valores y la media:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Desviación Estándar

La desviación estándar es la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

1.3. Probabilidad

1.3.1. Espacio Muestral y Eventos

1.3.2. Definiciones de Probabilidad

Probabilidad Clásica

Probabilidad Frecuentista

Probabilidad Bayesiana

1.4. Estadística Bayesiana

1.4.1. Prior y Posterior

Distribución Prior

Verosimilitud

La verosimilitud (likelihood) es la probabilidad de observar los datos dados los parámetros. Es una función de los parámetros θ dada una muestra de datos X:

$$L(\theta; X) = P(X|\theta)$$

donde X son los datos observados y θ son los parámetros del modelo.

Distribución Posterior

La distribución posterior (a posteriori) combina la información de la prior y la verosimilitud utilizando el teorema de Bayes. Representa nuestra creencia sobre los parámetros después de observar los datos:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

donde $P(\theta|X)$ es la distribución posterior, $P(X|\theta)$ es la verosimilitud, $P(\theta)$ es la prior y $P(X)$ es la probabilidad marginal de los datos.

La probabilidad marginal de los datos $P(X)$ se puede calcular como:

$$P(X) = \int_{\Theta} P(X|\theta)P(\theta)d\theta$$

donde Θ es el espacio de todos los posibles valores del parámetro θ.

1.5. Distribuciones de Probabilidad

1.5.1. Distribuciones Discretas

Distribución Binomial

La distribución binomial describe el número de éxitos en una serie de ensayos de Bernoulli independientes y con la misma probabilidad de éxito. La función de probabilidad es:

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

donde X es el número de éxitos, n es el número de ensayos, p es la probabilidad de éxito en cada ensayo, y $\binom{n}{k}$ es el coeficiente binomial.

La función generadora de momentos (MGF) para la distribución binomial es:

$$M_X(t) = \left(1 - p + pe^t\right)^n$$

El valor esperado y la varianza de una variable aleatoria binomial son:

$$\begin{aligned} E(X) &= np \\ \text{Var}(X) &= np(1-p) \end{aligned}$$

Distribución de Poisson

La distribución de Poisson describe el número de eventos que ocurren en un intervalo de tiempo fijo o en un área fija. La función de probabilidad es:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

donde X es el número de eventos, λ es la tasa media de eventos por intervalo, y k es el número de eventos observados.

La función generadora de momentos (MGF) para la distribución de Poisson es:

$$M_X(t) = e^{\lambda(e^t - 1)}$$

El valor esperado y la varianza de una variable aleatoria de Poisson son:

$$\begin{aligned} E(X) &= \lambda \\ \text{Var}(X) &= \lambda \end{aligned}$$

1.5.2. Distribuciones Continuas

Distribución Normal

La distribución normal, también conocida como distribución gaussiana, es una de las distribuciones más importantes en estadística. La función de densidad de probabilidad es:

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

donde x es un valor de la variable aleatoria, μ es la media, y σ es la desviación estándar.

La función generadora de momentos (MGF) para la distribución normal es:

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

El valor esperado y la varianza de una variable aleatoria normal son:

$$\begin{aligned} E(X) &= \mu \\ \text{Var}(X) &= \sigma^2 \end{aligned}$$

Distribución Exponencial

La distribución exponencial describe el tiempo entre eventos en un proceso de Poisson. La función de densidad de probabilidad es:

$$f(x) = \lambda e^{-\lambda x}$$

donde x es el tiempo entre eventos y λ es la tasa media de eventos.

La función generadora de momentos (MGF) para la distribución exponencial es:

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad \text{para } t < \lambda$$

El valor esperado y la varianza de una variable aleatoria exponencial son:

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ \text{Var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

1.5.3. Pruebas de Hipótesis

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

Una prueba de hipótesis está formada por cinco partes:

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

Definición 1.2 *Las dos hipótesis en competencia son la* ***hipótesis alternativa*** H_1*, usualmente la que se desea apoyar, y la* ***hipótesis nula*** H_0*, opuesta a* H_1*.*

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

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

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

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

Ejemplo 1.2 *Un determinado proceso produce un promedio de* 5 % *de piezas defectuosas. Se está interesado en demostrar que un simple ajuste en una máquina reducirá* p*, la proporción de piezas defectuosas producidas en este proceso. Entonces se tiene* $H_0 : p < 0{,}3$ *y* $H_1 : p = 0{,}03$*. Si se puede rechazar* H_0*, se concluye que el proceso ajustado produce menos del* 5 % *de piezas defectuosas.*

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

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

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

Definición 1.3 *El p-value es la probabilidad de observar un estadístico de prueba tanto o más alejado del valor observado, si en realidad* H_0 *es verdadera. Valores grandes del estadístico de prueba y valores pequeños de p significan que se ha observado un evento muy poco probable, si* H_0 *en realidad es verdadera.*

Todo el conjunto de valores que puede tomar el estadístico de prueba se divide en dos regiones. Un conjunto, formado de valores que apoyan la hipótesis alternativa y llevan a rechazar H_0, se denomina **región de rechazo**. El otro, conformado por los valores que sustentan la hipótesis nula, se le denomina **región de aceptación**. Cuando la región de rechazo está en la cola izquierda de la distribución, la prueba se denomina **prueba lateral izquierda**. Una prueba con región de rechazo en la cola derecha se le llama **prueba lateral derecha**. Si el estadístico de prueba cae en la región de rechazo, entonces se rechaza H_0. Si el estadístico de prueba cae en la región de aceptación, entonces la hipótesis nula se acepta o la prueba se juzga como no concluyente. Dependiendo del nivel de confianza que se desea agregar a las conclusiones de la prueba, y el **nivel de significancia** α, el riesgo que está dispuesto a correr si se toma una decisión incorrecta.

Definición 1.4 *Un **error de tipo I** para una prueba estadística es el error que se tiene al rechazar la hipótesis nula cuando es verdadera. El **nivel de significancia** para una prueba estadística de hipótesis es*

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

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

1.5.4. Muestras grandes: una media poblacional

Definición 1.5 *El **valor de** p (p**-value**) o nivel de significancia observado de un estadístico de prueba es el valor más pequeño de α para el cual H_0 se puede rechazar. El riesgo de cometer un error tipo I, si H_0 es rechazada con base en la información que proporciona la muestra.*

Nota 1.1 *Valores pequeños de p indican que el valor observado del estadístico de prueba se encuentra alejado del valor hipotético de μ, es decir se tiene evidencia de que H_0 es falsa y por tanto debe de rechazarse.*

Nota 1.2 *Valores grandes de p indican que el estadístico de prueba observado no está alejado de la media hipotética y no apoya el rechazo de H_0.*

Definición 1.6 *Si el valor de p es menor o igual que el nivel de significancia α, determinado previamente, entonces H_0 es rechazada y se puede concluir que los resultados son estadísticamente significativos con un nivel de confianza del $100(1-\alpha)\,\%$.*

Es usual utilizar la siguiente clasificación de resultados:

p	H_0	Significativa
$p \leq 0{,}01$	rechazada	Result. altamente significativos y en contra de H_0
$p \leq 0{,}05$	rechazada	Result. Estadísticamente significativos y en contra de H_0
$p \leq 0{,}10$	rechazada	Result. posiblemente significativos con Tendencia estadística y en contra de H_0
$p > 0{,}10$	no rechazada	Result. estadísticamente no significativos y no rechazar H_0

Nota 1.3 *Para determinar el valor de p, encontrar el área en la cola después del estadístico de prueba. Si la prueba es de una cola, este es el valor de p. Si es de dos colas, éste valor encontrado es la mitad del valor de p. Rechazar H_0 cuando el valor de $p < \alpha$.*

Hay dos tipos de errores al realizar una prueba de hipótesis

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

Definición 1.7 *La probabilidad de cometer el error tipo II se define por β donde*

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

Nota 1.4 *Cuando H_0 es falsa y H_1 es verdadera, no siempre es posible especificar un valor exacto de μ, sino más bien un rango de posibles valores.En lugar de arriesgarse a tomar una decisión incorrecta, es mejor conlcuir que no hay evidencia suficiente para rechazar H_0, es decir en lugar de aceptar H_0, no rechazar H_0.*

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

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

Definición 1.8 *La **potencia de la prueba**, $1-\beta$, mide la capacidad de que la prueba funciona como se necesita.*

Ejemplo 1.3 *La producción diariade una planta química local ha promediado 880 toneladas en los últimos años. A la gerente de control de calidad le gustaría saber si este promedio ha cambiado en meses recientes. Ella selecciona al azar 50 días de la base de datos computarizada y calcula el promedio y la desviación estándar de las $n = 50$ producciones como $\overline{x} = 871$ toneladas y $s = 21$ toneladas, respectivamente. Pruebe la hipótesis apropiada usando $\alpha = 0{,}05$.*

La hipótesis nula apropiada es:

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

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

$$\begin{aligned}z &= \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \\ &= \frac{871-880}{21/\sqrt{50}} = -3{,}03\end{aligned}$$

*Para esta prueba de dos colas, hay que determinar los dos valores de $z_{\alpha/2}$, es decir, $z_{\alpha/2} = \pm 1{,}96$, como $z > z_{\alpha/2}$, z cae en la zona de rechazo, por lo tanto la gerente puede rechazar la hipótesis nula y concluir que el promedio efectivamente ha cambiado. La probabilidad de rechazar H_0 cuando esta es verdadera es de 0,05. Recordemos que el valor observado del estadístico de prueba es $z = -3{,}03$, la región de rechazo más pequeña que puede usarse y todavía seguir rechazando H_0 es $|z| > 3{,}03$, entonces $p = 2(0{,}012) = 0{,}0024$, que a su vez es menor que el nivel de significancia α asignado inicialmente, y además los resultados son **altamente significativos**. Finalmente determinemos la potencia de la prueba cuando μ en realidad es igual a 870 toneladas.*

Recordar que la región de aceptación está entre $-1{,}96$ y $1{,}96$, para $\mu = 880$, equivalentemente

$$874{,}18 < \overline{x} < 885{,}82$$

β es la probabilidad de aceptar H_0 cuando $\mu = 870$, calculemos los valores de z correspondientes a 874,18 y 885,82 Entonces

$$\begin{aligned}z_1 &= \frac{\overline{x}-\mu}{s/\sqrt{n}} = \frac{874{,}18-870}{21/\sqrt{50}} = 1{,}41 \\ z_1 &= \frac{\overline{x}-\mu}{s/\sqrt{n}} = \frac{885{,}82-870}{21/\sqrt{50}} = 5{,}33\end{aligned}$$

por lo tanto

$$\begin{aligned}\beta &= P\{\textit{aceptar } H_0 \textit{ cuando } H_0 \textit{ es falsa}\} = P\{874{,}18 < \mu < 885{,}82 \textit{ cuando } \mu = 870\} \\ &= P\{1{,}41 < z < 5{,}33\} = P\{1{,}41 < z\} = 1 - 0{,}9207 = 0{,}0793\end{aligned}$$

entonces, la potencia de la prueba es

$$1 - \beta = 1 - 0{,}0793 = 0{,}9207$$

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

Prueba de hipótesis para la diferencia entre dos medias poblacionales

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

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{1.1}$$

cuyo estimador está dado por

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{1.2}$$

El procedimiento para muestras grandes es:

1) **Hipótesis Nula** $H_0 : (\mu_1 - \mu_2) = D_0$, donde D_0 es el valor, la diferencia, específico que se desea probar. En algunos casos se querrá demostrar que no hay diferencia alguna, es decir $D_0 = 0$.

2) **Hipótesis Alternativa**

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

3) Estadístico de prueba:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1.3}$$

4) Región de rechazo: rechazar H_0 cuando

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

Ejemplo 1.4 *Para determinar si ser propietario de un automóvil afecta el rendimiento académico de un estudiante, se tomaron dos muestras aleatorias de 100 estudiantes varones. El promedio de calificaciones para los $n_1 = 100$ no propietarios de un auto tuvieron un promedio y varianza de $\overline{x}_1 = 2{,}7$ y $s_1^2 = 0{,}36$, respectivamente, mientras que para para la segunda muestra con $n_2 = 100$ propietarios de un auto, se tiene $\overline{x}_2 = 2{,}54$ y $s_2^2 = 0{,}4$. Los datos presentan suficiente evidencia para indicar una diferencia en la media en el rendimiento académico entre propietarios y no propietarios de un automóvil? Hacer pruebas para $\alpha = 0{,}01, 0{,}05$ y $\alpha = 0{,}1$.*

- *Solución utilizando la técnica de regiones de rechazo:realizando las operaciones* $z = 1{,}84$*, determinar si excede los valores de* $z_{\alpha/2}$.
- *Solución utilizando el p-value:Calcular el valor de p, la probabilidad de que z sea mayor que* $z = 1{,}84$ *o menor que* $z = -1{,}84$*, se tiene que* $p = 0{,}0658$.
- *Si el intervalo de confianza que se construye contiene el valor del parámetro especificado por* H_0*, entonces ese valor es uno de los posibles valores del parámetro y* H_0 *no debe ser rechazada.*
- *Si el valor hipotético se encuentra fuera de los límites de confianza, la hipótesis nula es rechazada al nivel de significancia* α.

Prueba de Hipótesis para una Proporción Binomial

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

$$SE = \sqrt{\frac{pq}{n}}. \tag{1.4}$$

La prueba de hipótesis de la forma

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

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

El procedimiento es

1) Hipótesis nula: $H_0 : p = p_0$
2) Hipótesis alternativa

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

3) Estadístico de prueba:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{pq}{n}}}, \hat{p} = \frac{x}{n} \tag{1.5}$$

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

4) Región de rechazo: rechazar H_0 cuando

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

Prueba de Hipótesis diferencia entre dos Proporciones Binomiales

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

$$SE = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}, \tag{1.6}$$

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

La hipótesis nula a probarse es de la forma

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

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

$$p = \frac{\text{número total de éxitos}}{\text{Número total de pruebas}} = \frac{x_1 + x_2}{n_1 + n_2}. \tag{1.7}$$

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

2) **Hipótesis Alternativa:** H_1 :

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

3) Estadístico de prueba:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}}, \tag{1.8}$$

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

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{1.9}$$

4) Región de rechazo: rechazar H_0 cuando

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

1.5.5. Muestras Pequeñas

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

2) **Hipótesis Alternativa:** H_1 :

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

3) Estadístico de prueba:

$$t = \frac{\overline{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}, \tag{1.10}$$

4) Región de rechazo: rechazar H_0 cuando

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

Diferencia entre dos medias poblacionales: MAI

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

$$ES = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{1.11}$$

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

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

2) **Hipótesis Alternativa**

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

3) Estadístico de prueba:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1.12}$$

donde

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}. \tag{1.13}$$

4) Región de rechazo: rechazar H_0 cuando

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

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

Diferencia entre dos medias poblacionales: Diferencias Pareadas

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

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

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

3) Estadístico de prueba:

$$t = \frac{\overline{d}}{\sqrt{\frac{s_d^2}{n}}} \tag{1.14}$$

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

4) Región de rechazo: rechazar H_0 cuando

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

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

Inferencias con respecto a la Varianza Poblacional

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

2) **Hipótesis Alternativa:** H_1

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

3) Estadístico de prueba:

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2}, \tag{1.15}$$

4) Región de rechazo: rechazar H_0 cuando

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

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

Comparación de dos varianzas poblacionales

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

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

2) **Hipótesis Alternativa**

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

3) Estadístico de prueba:

$$F = \frac{s_1^2}{s_2^2} \tag{1.16}$$

donde s_1^2 es la varianza muestral más grande.

4) Región de rechazo: rechazar H_0 cuando

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

1.5.6. Estimación por intervalos

Recordemos que S^2 es un estimador insesgado de σ^2, entonces se tiene la siguiente definición

Definición 1.9 *Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados de θ, parámetro poblacional. Si $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$, decimos que $\hat{\theta}_1$ un estimador más eficaz de θ que $\hat{\theta}_2$.*

Algunas observaciones que es preciso realizar

Nota 1.5 *a) Para poblaciones normales, $\overline{X}$ y $\tilde{X}$ son estimadores insesgados de μ, pero con $\sigma^2_{\overline{X}} < \sigma^2_{\tilde{X}_2}$.*

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

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

Nota 1.6 *Para $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$, si $n \to \infty$,*

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

$$P\left\{\hat{\theta}_L < \hat{\theta} < \hat{\theta}_U\right\} = 1 - \alpha, \tag{1.17}$$

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

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

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

$$P\left\{-Z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right\} = 1 - \alpha, \tag{1.18}$$

es equivalente a

$$P\left\{\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha. \tag{1.19}$$

f) Si $\overline{X}$ es la media muestral de una muestra de tamaño n de una población con varianza conocida σ^2, el intervalo de confianza de $100\,(1-\alpha)$ % para μ es

$$\mu \in \left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \tag{1.20}$$

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

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

Teorema 1.1 *Si $\overline{X}$ es un estimador de μ, podemos tener $100\,(1-\alpha)$ % de confianza en que el error no excederá a $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, error entre $\overline{X}$ y μ.*

Teorema 1.2 *Si $\overline{X}$ es un estimador de μ, podemos tener $100\,(1-\alpha)$ % de confianza en que el error no excederá una cantidad e cuando el tamaño de la muestra es*

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2. \tag{1.21}$$

Nota 1.7 *Para intervalos unilaterales*

$$P\left\{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_\alpha\right\} = 1 - \alpha \tag{1.22}$$

equivalentemente

$$P\left\{\mu < \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha. \tag{1.23}$$

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

a) Límite unilateral superior: $\overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$

b) Límite unilateral inferior: $\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}$

c) Para σ desconocida recordar que $T = \frac{\overline{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, donde s es la desviación estándar de la muestra. Entonces

$$P\left\{-t_{\alpha/2} < T < t_{\alpha/2}\right\} = 1 - \alpha, \text{equivalentemente} \tag{1.24}$$

$$P\left\{\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right\} = 1 - \alpha. \tag{1.25}$$

d) Un intervalo de confianza del $100\,(1-\alpha)$ % de confianza para μ, σ^2 desconocida y población normal es

$$\mu \in \left(\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) \tag{1.26}$$

,

donde $t_{\alpha/2}$ es una t-student con $\nu = n - 1$ grados de libertad.

e) Los límites unilaterales para μ con σ desconocida son $\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}$ y $\overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$.

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

$$\overline{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}. \tag{1.27}$$

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

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

$$\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}. \tag{1.28}$$

Intervalos de confianza sobre la varianza

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

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

$$\begin{array}{l} P\left\{\chi^2_{1-\frac{\alpha}{2},n-1} \leq \chi^2 \leq \chi^2_{\frac{\alpha}{2},n-1}\right\} = 1-\alpha \\ P\left\{\chi^2_{1-\frac{\alpha}{2},n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n-1}\right\} = 1-\alpha \\ P\left\{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right\} = 1-\alpha, \end{array} \tag{1.29}$$

es decir

$$\sigma^2 \in \left[\frac{(n-1)\, s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)\, s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right], \tag{1.30}$$

los intervalos unilaterales son

$$\sigma^2 \in \left[\frac{(n-1)\, s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \infty \right], \tag{1.31}$$

y

$$\sigma^2 \in \left[-\infty, \frac{(n-1)\, s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right]. \tag{1.32}$$

Intervalos de confianza para proporciones

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

$$\hat{p} = \frac{\overline{X}}{n}, \tag{1.33}$$

es el estimador puntual de la proporción de la población que pertenece a dicha clase. n y p son los parámetros de la distribución binomial, entonces

$$\hat{p} \sim N\left(p, \frac{p\,(1-p)}{n} \right) \tag{1.34}$$

aproximadamente si p es distinto de 0 y 1; o si n es suficientemente grande. Entonces

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N\,(0,1)\,, \text{ aproximadamente.} \tag{1.35}$$

Entonces

$$\begin{aligned} 1-\alpha &= P\left\{ -z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2} \right\} \\ &= P\left\{ \hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right\} \end{aligned} \tag{1.36}$$

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

$$1-\alpha = P\left\{ \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \right\} \tag{1.37}$$

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

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

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

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

para minimizar el error estándar, se propone que el tamaño de la muestra sea

$$n = \left(\frac{z_{\alpha/2}}{E} \right)^2 p\,(1-p)\,, \tag{1.38}$$

donde

$$E = \mid p - \hat{p} \mid .$$

Intervalos de confianza para dos muestras: Varianzas conocidas

Sean X_1 y X_2 variables aleatorias independientes. X_1 con media desconocida μ_1 y varianza conocida σ_1^2; y X_2 con media desconocida μ_2 y varianza conocida σ_2^2. Se busca encontrar un intervalo de confianza de $100\,(1-\alpha)$ % de la diferencia entre medias μ_1 y μ_2. Sean $X_{11}, X_{12}, \ldots, X_{1n_1}$ muestra aleatoria de n_1 observaciones de X_1, y sean $X_{21}, X_{22}, \ldots, X_{2n_2}$ muestra aleatoria de n_2 observaciones de X_2.

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

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N\,(0,1)\,, \tag{1.39}$$

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

$$1-\alpha = P\left\{-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right\} = P\left\{-Z_{\alpha/2} \le \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le Z_{\alpha/2}\right\} \tag{1.40}$$

$$= P\left\{\left(\overline{X}_1 - \overline{X}_2\right) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\overline{X}_1 - \overline{X}_2\right) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\}. \tag{1.41}$$

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

a)

$$\mu_1 - \mu_2 \in \left[\left(\overline{X}_1 - \overline{X}_2\right) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \left(\overline{X}_1 - \overline{X}_2\right) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right], \tag{1.42}$$

b)

$$\mu_1 - \mu_2 \in \left[-\infty, \left(\overline{X}_1 - \overline{X}_2\right) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right], \tag{1.43}$$

c)

$$\mu_1 - \mu_2 \in \left[\left(\overline{X}_1 - \overline{X}_2\right) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \infty\right]. \tag{1.44}$$

Nota 1.8 *Si σ_1 y σ_2 son conocidas, o por lo menos se conoce una aproximación, y los tamaños de las muestras n_1 y n_2 son iguales, $n_1 = n_2 = n$, se puede determinar el tamaño de la muestra para que el error al estimar $\mu_1 - \mu_2$ usando $\overline{X}_1 - \overline{X}_2$ sea menor que E (valor del error deseado) al $(1-\alpha)$ % de confianza. El tamaño n de la muestra requerido para cada muestra es*

$$n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 \left(\sigma_1^2 + \sigma_2^2\right). \tag{1.45}$$

Intervalos de confianza para dos muestras: Varianzas desconocidas e iguales

a) Si $n_1, n_2 \ge 30$ se pueden utilizar los intervalos de la distribución normal para varianza conocida

b) Si n_1, n_2 son muestras pequeñas, supongase que las poblaciones para X_1 y X_2 son normales con varianzas desconocidas y con base en el intervalo de confianza para distribuciones t-student

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

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

$$S_p^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}, \tag{1.46}$$

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

$$t_\nu = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{1.47}$$

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

Por lo tanto

$$\begin{aligned} 1 - \alpha &= P\left\{-t_{\alpha/2,\nu} \leq t \leq t_{\alpha/2,\nu}\right\} \\ &= P\left\{\left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2,\nu} S_p\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}} \leq t \leq \left(\overline{X}_1 - \overline{X}_2\right) + t_{\alpha/2,\nu} S_p\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}\right\}, \end{aligned} \tag{1.48}$$

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

a)

$$\mu_1 - \mu_2 \in \left[\left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \left(\overline{X}_1 - \overline{X}_2\right) + t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{1.49}$$

b)

$$\mu_1 - \mu_2 \in \left[-\infty, \left(\overline{X}_1 - \overline{X}_2\right) + t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{1.50}$$

c)

$$\mu_1 - \mu_2 \in \left[\left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2,\nu} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty\right]. \tag{1.51}$$

Intervalos de confianza para dos muestras: Varianzas desconocidas diferentes

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

$$t^* = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{1.52}$$

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

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^2/n_1}{n_1+1} + \frac{S_2^2/n_2}{n_2+1}} - 2. \tag{1.53}$$

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

$$\mu_1 - \mu_2 \in \left[\left(\overline{X}_1 - \overline{X}_2\right) - t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \left(\overline{X}_1 - \overline{X}_2\right) + t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right]. \tag{1.54}$$

Intervalos de confianza para razón de Varianzas

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

$$F = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2}, \tag{1.55}$$

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

Por lo tanto

$$\begin{array}{l} P\left\{F_{1-\frac{\alpha}{2},n_2-1,n_1-1} \leq F \leq F_{\frac{\alpha}{2},n_2-1,n_1-1}\right\} = 1-\alpha, \\ P\left\{F_{1-\frac{\alpha}{2},n_2-1,n_1-1} \leq \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \leq F_{\frac{\alpha}{2},n_2-1,n_1-1}\right\} = 1-\alpha, \end{array} \tag{1.56}$$

luego entonces

$$P\left\{\frac{S_1^2}{S_2^2}F_{1-\frac{\alpha}{2},n_2-1,n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2}F_{\frac{\alpha}{2},n_2-1,n_1-1}\right\} = 1-\alpha. \tag{1.57}$$

en consecuencia

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{S_1^2}{S_2^2}F_{1-\frac{\alpha}{2},n_2-1,n_1-1}, \frac{S_1^2}{S_2^2}F_{\frac{\alpha}{2},n_2-1,n_1-1}\right], \tag{1.58}$$

donde

$$F_{1-\frac{\alpha}{2},n_2-1,n_1-1} = \frac{1}{F_{\frac{\alpha}{2},n_2-1,n_1-1}}. \tag{1.59}$$

Intervalos de confianza para diferencia de proporciones

Sean dos proporciones de interés p_1 y p_2. Se busca un intervalo para p_1-p_2 al $100\,(1-\alpha)$ %. Sean dos muestras independientes de tamaño n_1 y n_2 de poblaciones infinitas de modo que X_1 y X_2 variables aleatorias binomiales independientes con parámetros (n_1,p_1) y (n_2,p_2). X_1 y X_2 son el número de observaciones que pertenecen a la clase de interés correspondientes. Entonces $\hat{p}_1 = \frac{X_1}{n_1}$ y $\hat{p}_2 = \frac{X_2}{n_2}$ son estimadores de p_1 y p_2 respectivamente. Supongamos que se cumple la aproximación normal a la binomial, entonces

$$Z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} - \frac{p_2(1-p_2)}{n_2}}} \sim N\,(0,1) \text{ aproximadamente} \tag{1.60}$$

por tanto

$$(\hat{p}_1-\hat{p}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}} \leq p_1-p_2 \leq (\hat{p}_1-\hat{p}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} - \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}} \tag{1.61}$$

1.6. Análisis de Regresion Lineal (RL)

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

1.6.1. Regresión Lineal Simple (RLS)

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

$$E\left(y|x\right) = \beta_0 + \beta_1 x,$$

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

Supongamos que cada observación y se puede describir por el modelo

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1.62}$$

donde ϵ es un error aleatorio con media cero y varianza σ^2. Para cada valor y_i se tiene ϵ_i variables aleatorias no correlacionadas, cuando se incluyen en el modelo 1.62, este se le llama *modelo de regresión lineal simple*. Suponga que se tienen n pares de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, estos datos pueden utilizarse para estimar los valores de β_0 y β_1. Esta estimación realiza por el **métodos de mínimos cuadrados**.

Entonces la ecuación (1.62) se puede reescribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ para } i = 1, 2, \ldots, n. \tag{1.63}$$

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

$$L = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2. \tag{1.64}$$

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

$$\frac{\partial L}{\partial \beta_0} = 0, \; \frac{\partial L}{\partial \beta_1} = 0.$$

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

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

simplificando

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

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

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \tag{1.65}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}, \tag{1.66}$$

entonces el modelo de regresión lineal simple ajustado es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{1.67}$$

Se intrduce la siguiente notación

$$S_{xx} = \sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2, \tag{1.68}$$

$$S_{xy} = \sum_{i=1}^{n}y_i(x_i-\overline{x}) = \sum_{i=1}^{n}x_iy_i - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right); \tag{1.69}$$

y por tanto

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}. \tag{1.70}$$

1.6.2. Propiedades de los Estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$

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

A saber

$$\begin{aligned} E\left(\hat{\beta}_1\right) &= E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}}E\left(\sum_{i=1}^{n}y_i(x_i-\overline{x})\right) = \frac{1}{S_{xx}}E\left(\sum_{i=1}^{n}(\beta_0+\beta_1x_i+\epsilon_i)(x_i-\overline{x})\right) \\ &= \frac{1}{S_{xx}}\left[\beta_0E\left(\sum_{k=1}^{n}(x_k-\overline{x})\right) + E\left(\beta_1\sum_{k=1}^{n}x_k(x_k-\overline{x})\right) + E\left(\sum_{k=1}^{n}\epsilon_k(x_k-\overline{x})\right)\right] \\ &= \frac{1}{S_{xx}}\beta_1S_{xx} = \beta_1. \end{aligned}$$

Por lo tanto

$$E\left(\hat{\beta}_1\right) = \beta_1, \tag{1.71}$$

Es decir, $\hat{\beta}_1$ es un estimador insesgado. Ahora calculemos la varianza:

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

por lo tanto

$$V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}}. \tag{1.72}$$

Entonces tenemos la siguiente proposición:

Proposición 1.1

$$E\left(\hat{\beta}_0\right) = \beta_0, \tag{1.73}$$

$$V\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right), \tag{1.74}$$

$$Cov\left(\hat{\beta}_0,\hat{\beta}_1\right) = -\frac{\sigma^2\overline{x}}{S_{xx}}. \tag{1.75}$$

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

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

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

$$SC_E = \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n} \left(y_k - \hat{y}_k\right)^2, \tag{1.76}$$

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

$$SC_E = \sum_{k=1}^{n} y_k^2 - n\overline{y}^2 - \hat{\beta}_1 S_{xy} = S_{yy} - \hat{\beta}_1 S_{xy}, \tag{1.77}$$

$$E\left(SC_E\right) = (n-2)\,\sigma^2, \text{ por lo tanto,} \tag{1.78}$$

$$\hat{\sigma}^2 = \frac{SC_E}{n-2} = \mathbf{MC_E} \text{ es un estimador insesgado de } \sigma^2. \tag{1.79}$$

1.6.3. Prueba de Hipótesis en RLS

Para evaluar la suficiencia del modelo de regresión lineal simple, es necesario lleva a cabo una prueba de hipótesis respecto de los parámetros del modelo así como de la construcción de intervalos de confianza. Para poder realizar la prueba de hipótesis sobre la pendiente y la ordenada al orígen de la recta de regresión es necesario hacer el supuesto de que el error ϵ_i se distribuye normalmente, es decir $\epsilon_i \sim N\left(0, \sigma^2\right)$. Suponga que se desea probar la hipótesis de que la pendiente es igual a una constante, $\beta_{0,1}$ las hipótesis Nula y Alternativa son:

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

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

donde dado que las $\epsilon_i \sim N\left(0, \sigma^2\right)$, se tiene que y_i son variables aleatorias normales $N\left(\beta_0 + \beta_1 x_1, \sigma^2\right)$. De las ecuaciones (1.65) se desprende que $\hat{\beta}_1$ es combinación lineal de variables aleatorias normales independientes, es decir, $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2/S_{xx}\right)$, recordar las ecuaciones (1.71) y (1.72). Entonces se tiene que el estadístico de prueba apropiado es

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{MC_E/S_{xx}}} \tag{1.80}$$

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

$$t_0| > t_{\alpha/2,n-2}. \tag{1.81}$$

Para β_0 se puede proceder de manera análoga para

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

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

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

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{MC_E\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}, \tag{1.82}$$

con el que rechazamos la hipótesis nula si

$$t_0| > t_{\alpha/2,n-2}. \tag{1.83}$$

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

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

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

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

Por lo tanto, efectivamente se tiene

$$S_{yy} = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 + \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 , \tag{1.84}$$

donde se hacen las definiciones

$$SC_E = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 \cdots \text{Suma de Cuadrados del Error} \tag{1.85}$$

$$SC_R = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \cdots \text{ Suma de Regresión de Cuadrados} \tag{1.86}$$

Por lo tanto la ecuación (1.84) se puede reescribir como:

$$S_{yy} = SC_R + SC_E, \tag{1.87}$$

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

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

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

Proposición 1.2

$$E(SC_R) = \sigma^2 + \beta_1 S_{xx}, \tag{1.88}$$

además, SC_E y SC_R son independientes.

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

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

se distribuye $F_{1,n-2}$, y se rechazaría H_0 si $F_0 > F_{\alpha,1,n-2}$. El procedimiento de prueba de hipótesis puede presentarse como la tabla de análisis de varianza siguiente

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

La prueba para la significación de la regresión puede desarrollarse basándose en la expresión (1.80), con $\hat{\beta}_{1,0} = 0$, es decir,

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MC_E/S_{xx}}}. \tag{1.89}$$

Elevando al cuadrado ambos términos:

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

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

1.6.4. Estimación de Intervalos en RLS

Además de la estimación puntual para los parámetros β_1 y β_0, es posible obtener estimaciones del intervalo de confianza de estos parámetros. El ancho de estos intervalos de confianza es una medida de la calidad total de la recta de regresión. Si los ϵ_k se distribuyen normal e independientemente, entonces

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

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

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}}. \tag{1.90}$$

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

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)} \tag{1.91}$$

1.6.5. Predicción

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

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \tag{1.92}$$

Esta nueva observación es independiente de las utilizadas para obtener el modelo de regresión, por tanto, el intervalo en torno a la recta de regresión es inapropiado, puesto que se basa únicamente en los datos empleados para ajustar el modelo de regresión. El intervalo de confianza en torno a la recta de regresión se refiere a la respuesta media verdadera $x = x_0$, no a observaciones futuras. Sea y_0 la observación futura en $x = x_0$, y sea $\hat{y}_0$ dada en la ecuación anterior, el estimador de y_0. Si se define la variable aleatoria

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

esta se distribuye normalmente con media cero y varianza

$$V(w) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{xx}} \right] \tag{1.93}$$

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

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MC_E \left[1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{xx}} \right]} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MC_E \left[1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{xx}} \right]}. \tag{1.94}$$

1.6.6. Prueba de falta de ajuste

Es común encontrar que el modelo ajustado no satisface totalmente el modelo necesario para los datos, en este caso es preciso saber qué tan bueno es el modelo propuesto. Para esto se propone la siguiente prueba de hipótesis:

H_0 : El modelo propuesto se ajusta adecuademente a los datos.

H_1 : El modelo NO se ajusta a los datos.

La prueba implica dividir la suma de cuadrados del eror o del residuo en las siguientes dos componentes:

$$SC_E = SC_{EP} + SC_{FDA} \tag{1.95}$$

donde SC_{EP} es la suma de cuadrados atribuibles al error puro, y SC_{FDA} es la suma de cuadrados atribuible a la falta de ajuste del modelo.

1.6.7. Coeficiente de Determinación

La cantidad

$$R^2 = \frac{SC_R}{S_{yy}} = 1 - \frac{SC_E}{S_{yy}}, \tag{1.96}$$

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

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

Capítulo 2

Regresión Logística

2.1. Introducción

La regresión logística es una técnica de modelado estadístico ampliamente utilizada en análisis de datos cuando el objetivo es predecir la probabilidad de un resultado binario, es decir, cuando la variable dependiente o respuesta tiene dos posibles categorías, como "éxito/fallo.º "sí/no". Esta técnica se emplea en una variedad de disciplinas, como la biomedicina, ciencias sociales, marketing y más, para resolver problemas donde la variable respuesta es discreta o categórica.

A diferencia de la regresión lineal, que asume una relación lineal entre las variables independientes y la variable dependiente y que produce valores en un rango continuo, la regresión logística está diseñada para manejar situaciones donde la respuesta es categórica. En su forma más común, la regresión logística binaria, el modelo predice la probabilidad de que un evento ocurra en función de una o más variables independientes. Este tipo de regresión toma la forma de un modelo no lineal, debido a la naturaleza discreta de la variable dependiente.

La regresión lineal busca modelar la relación entre una variable dependiente continua Y y una o más variables independientes $X_1, X_2, \ldots, X_n$ mediante una ecuación de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon, \tag{2.1}$$

donde $\beta_0, \beta_1, \ldots, \beta_n$ son los coeficientes del modelo y ϵ es el término de error. La regresión logística, en cambio, modela la probabilidad de que un evento ocurra (por ejemplo, éxito vs. fracaso) utilizando la función logística. La variable dependiente Y es binaria, tomando valores de 0 o 1. La ecuación de la regresión logística es:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n, \tag{2.2}$$

donde p es la probabilidad de que $Y = 1$. La función logística es:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}. \tag{2.3}$$

La regresión logística se utiliza en una variedad de campos para problemas de clasificación binaria, tales como:

- **Medicina**: Predicción de la presencia o ausencia de una enfermedad.
- **Marketing**: Determinación de la probabilidad de que un cliente compre un producto.
- **Finanzas**: Evaluación del riesgo de crédito, es decir, si un cliente va a incumplir o no con un préstamo.
- **Seguridad**: Detección de fraudes o intrusiones.

2.1.1. Implementación Básica en R

Para implementar una regresión logística en R, primero es necesario instalar y cargar los paquetes necesarios. Aquí se muestra un ejemplo básico de implementación:

- Descargue e instale R desde `https://cran.r-project.org/`.
- Descargue e instale RStudio desde `https://rstudio.com/products/rstudio/download/`.

```
# Instalaci\'on del paquete necesario
install.packages("stats")

# Carga del paquete
library(stats)

# Ejemplo de conjunto de datos
data <- data.frame(
  outcome = c(1, 0, 1, 0, 1, 1, 0, 1, 0, 0),
  predictor = c(2.3, 1.9, 3.1, 2.8, 3.6, 2.4, 2.1, 3.3, 2.2, 1.7)
)

# Ajuste del modelo de regresi\'on log\'istica
model <- glm(outcome ~ predictor, data = data, family = binomial)

# Resumen del modelo
summary(model)
```

2.2. Conceptos Básicos

La regresión logística es una técnica de modelado estadístico utilizada para predecir la probabilidad de un evento binario (es decir, un evento que tiene dos posibles resultados) en función de una o más variables independientes. Un modelo de regresión logística describe cómo una variable dependiente binaria Y (que puede tomar los valores 0 o 1) está relacionada con una o más variables independientes $X_1, X_2, \ldots, X_n$. A diferencia de la regresión lineal, que predice un valor continuo, la regresión logística predice una probabilidad que puede ser interpretada como la probabilidad de que $Y = 1$ dado un conjunto de valores para $X_1, X_2, \ldots, X_n$.

2.2.1. Regresión Lineal

La regresión lineal es utilizada para predecir el valor de una variable dependiente continua en función de una o más variables independientes. El modelo de regresión lineal tiene la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{2.4}$$

donde:

a) Y es la variable dependiente.

b) β_0 es la intersección con el eje Y o término constante.

c) $\beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes que representan la relación entre las variables independientes y la variable dependiente.

d) $X_1, X_2, \ldots, X_n$ son las variables independientes.

e) ϵ es el término de error, que representa la desviación de los datos observados de los valores predichos por el modelo.

El objetivo de la regresión lineal es encontrar los valores de los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ que minimicen la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos. Este método se conoce como mínimos cuadrados ordinarios (OLS, por sus siglas en inglés). La función de costo a minimizar es:

$$J(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.5}$$

donde:

a) y_i es el valor observado de la variable dependiente para la i-ésima observación.

b) $\hat{y}_i$ es el valor predicho por el modelo para la i-ésima observación, dado por:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} \tag{2.6}$$

Para encontrar los valores óptimos de los coeficientes, se toman las derivadas parciales de la función de costo con respecto a cada coeficiente y se igualan a cero:

$$\frac{\partial J}{\partial \beta_j} = 0 \quad \text{para } j = 0, 1, \ldots, n \tag{2.7}$$

Resolviendo este sistema de ecuaciones, se obtienen los valores de los coeficientes que minimizan la función de costo.

2.2.2. Regresión Logística

La deducción de la fórmula de la regresión logística comienza con la necesidad de modelar la probabilidad de un evento binario. Queremos encontrar una función que relacione las variables independientes con la probabilidad de que la variable dependiente tome el valor 1. La probabilidad de que el evento ocurra, $P(Y = 1)$, se denota como p. La probabilidad de que el evento no ocurra, $P(Y = 0)$, es $1 - p$. Los **odds** (chances) de que ocurra el evento se definen como:

$$\text{odds} = \frac{p}{1-p} \tag{2.8}$$

Los odds indican cuántas veces más probable es que ocurra el evento frente a que no ocurra. Para simplificar el modelado de los *odds*, se aplica el logaritmo natural, obteniendo la función **logit**:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{2.9}$$

La transformación logit es útil porque convierte el rango de la probabilidad (0, 1) al rango de números reales $(-\infty, \infty)$. La idea clave de la regresión logística es modelar la transformación logit de la probabilidad como una combinación lineal de las variables independientes:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{2.10}$$

Aquí, β_0 es el término constante y $\beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes asociados con las variables independientes $X_1, X_2, \ldots, X_n$. Para expresar p en función de una combinación lineal de las variables independientes, invertimos la transformación logit. Partimos de la ecuación 2.10, aplicando la función exponencial en ambos lados:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n} \tag{2.11}$$

Despejando p:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}} \tag{2.12}$$

La expresión final que obtenemos es conocida como la **función logística**:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}} \tag{2.13}$$

Esta función describe cómo las variables independientes se relacionan con la probabilidad de que el evento de interés ocurra. Los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ se estiman a partir de los datos utilizando el método de máxima verosimilitud.

2.3. Método de Máxima Verosimilitud

Para estimar los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ en la regresión logística, utilizamos el método de máxima verosimilitud. La idea es encontrar los valores de los coeficientes que maximicen la probabilidad de observar los datos dados. Esta probabilidad se expresa mediante la función de verosimilitud L. La función de verosimilitud $L(\beta_0, \beta_1, \ldots, \beta_n)$ para un conjunto de n observaciones se define como el producto de las probabilidades de las observaciones dadas las variables independientes:

$$L(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{2.14}$$

donde:

a) p_i es la probabilidad predicha de que $Y_i = 1$,

b) y_i es el valor observado de la variable dependiente para la i-ésima observación.

Trabajar directamente con esta función de verosimilitud puede ser complicado debido al producto de muchas probabilidades, especialmente si n es grande. Para simplificar los cálculos, se utiliza el logaritmo de la función de verosimilitud, conocido como la función de *log-verosimilitud*. El uso del logaritmo simplifica significativamente la diferenciación y maximización de la función. La función de log-verosimilitud se define como:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (1-y_i)\log(1-p_i) \right]. \tag{2.15}$$

Aquí, log representa el logaritmo natural. Esta transformación es válida porque el logaritmo es una función monótona creciente, lo que significa que maximizar la log-verosimilitud es equivalente a maximizar la verosimilitud original. En la regresión logística, de acuerdo a la ecuación 2.13 la probabilidad p_i está dada por la función logística:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}} \tag{2.16}$$

Sustituyendo esta expresión en la función de log-verosimilitud (2.15), obtenemos:

$$\begin{array}{l} \log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}} \right) + \right. \\ \left. (1-y_i)\log\left(1 - \frac{1}{1+e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}}\right) \right] \end{array} \tag{2.17}$$

Simplificando esta expresión,

$$\log\left(\frac{1}{1+e^{-z}}\right) = -\log(1+e^{-z}), \text{ por tanto} \tag{2.18}$$

$$\log\left(1 - \frac{1}{1+e^{-z}}\right) = \log\left(\frac{e^{-z}}{1+e^{-z}}\right) = -z - \log(1+e^{-z}) \tag{2.19}$$

Aplicando las ecuaciones anteriores, la ecuación 2.17 se convierte en:

$$\begin{array}{l} \log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i(-\log(1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}))+ \right. \\ \left. (1-y_i)\left(-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) - \log(1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})})\right)\right] \end{array}$$

Simplificando aún más, obtenemos:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) - \log(1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}}) \right] \tag{2.20}$$

2.3.1. Notación Matricial

Para simplificar aún más la notación, podemos utilizar notación matricial. Definimos la matriz $\mathbf{X}$ de tamaño $n \times (k+1)$ y el vector de coeficientes $\boldsymbol{\beta}$ de tamaño $(k+1) \times 1$ como sigue:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \tag{2.21}$$

Entonces, la expresión para la función de log-verosimilitud, ecuación 2.20, es:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\mathbf{X}_i\boldsymbol{\beta}) - \log(1 + e^{\mathbf{X}_i\boldsymbol{\beta}}) \right] \tag{2.22}$$

donde $\mathbf{X}_i$ es la i-ésima fila de la matriz $\mathbf{X}$. Esta notación matricial simplifica la implementación y la derivación de los estimadores de los coeficientes en la regresión logística. Utilizando el método de Newton-Raphson, es posible encontrar los coeficientes β_j que maximizan la función de log-verosimilitud. Para maximizar la función de log-verosimilitud, derivamos esta función con respecto a cada uno de los coeficientes β_j y encontramos los puntos críticos. La derivada parcial de la función de log-verosimilitud con respecto a β_j es:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i X_{ij} - \frac{X_{ij} e^{\mathbf{X}_i\boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i\boldsymbol{\beta}}} \right] \tag{2.23}$$

Simplificando, esta derivada se puede expresar como:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij}(y_i - p_i), \text{ donde } p_i = \frac{1}{1 + e^{-\mathbf{X}_i\boldsymbol{\beta}}} \tag{2.24}$$

Para encontrar los coeficientes que maximizan la log-verosimilitud, se requiere resolver el sistema de ecuaciones

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = 0 \text{ para todos los } j = 0, 1, \dots, k. \tag{2.25}$$

Este sistema de ecuaciones no tiene una solución analítica cerrada, por lo que se propone resolver utilizando el método de Newton-Raphson.

2.3.2. Método de Newton-Raphson

El método de Newton-Raphson es un algoritmo iterativo que se utiliza para encontrar las raíces de una función. En el contexto de la regresión logística, se utiliza para maximizar la función de log-verosimilitud encontrando los valores de los coeficientes $\beta_0, \beta_1, \dots, \beta_n$. Este método se basa en una aproximación de segundo orden de la función objetivo. Dado un valor inicial de los coeficientes $\boldsymbol{\beta}^{(0)}$, se actualiza iterativamente el valor de los coeficientes utilizando la fórmula:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\mathbf{H}(\boldsymbol{\beta}^{(t)}) \right]^{-1} \nabla \log L(\boldsymbol{\beta}^{(t)}) \tag{2.26}$$

donde:

a) $\boldsymbol{\beta}^{(t)}$ es el vector de coeficientes en la t-ésima iteración.

b) $\nabla \log L(\boldsymbol{\beta}^{(t)})$ es el gradiente de la función de log-verosimilitud con respecto a los coeficientes $\boldsymbol{\beta}$:

$$\nabla \log L(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \mathbf{p}) \tag{2.27}$$

donde $\mathbf{y}$ es el vector de valores observados y $\mathbf{p}$ es el vector de probabilidades, recordar la ecuación (2.24).

- $\mathbf{H}(\boldsymbol{\beta}^{(t)})$ es la matriz Hessiana (matriz de segundas derivadas) evaluada en $\boldsymbol{\beta}^{(t)}$:

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{W}\mathbf{X} \tag{2.28}$$

  donde $\mathbf{W}$ es una matriz diagonal de pesos con elementos $w_i = p_i(1 - p_i)$.

En resumen:

Algoritmo 2.1 *El algoritmo Newton-Raphson para la regresión logística se puede resumir en los siguientes pasos:*

i) Inicializar el vector de coeficientes $\boldsymbol{\beta}^{(0)}$ (por ejemplo, con ceros o valores pequeños aleatorios).

ii) Calcular el gradiente $\nabla \log L(\boldsymbol{\beta}^{(t)})$ y la matriz Hessiana $\mathbf{H}(\boldsymbol{\beta}^{(t)})$ en la iteración t.

iii) Actualizar los coeficientes utilizando la fórmula:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\mathbf{H}(\boldsymbol{\beta}^{(t)})\right]^{-1} \nabla \log L(\boldsymbol{\beta}^{(t)}) \tag{2.29}$$

iv) Repetir los pasos ii) y iii) hasta que la diferencia entre $\boldsymbol{\beta}^{(t+1)}$ y $\boldsymbol{\beta}^{(t)}$ sea menor que un umbral predefinido (criterio de convergencia).

En resumen, el método de Newton-Raphson permite encontrar los coeficientes que maximizan la función de log-verosimilitud de manera eficiente.

2.3.3. Notas finales

En el contexto de la regresión logística, los vectores $X_1, X_2, \ldots, X_n$ representan las variables independientes. Cada X_j es un vector columna que contiene los valores de la variable independiente j para cada una de las n observaciones. Es decir,

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \tag{2.30}$$

Para simplificar la notación y los cálculos, a menudo combinamos todos los vectores de variables independientes en una única matriz de diseño $\mathbf{X}$ de tamaño $n \times (k + 1)$, donde n es el número de observaciones y $k + 1$ es el número de variables independientes más el término de intercepto. La primera columna de $\mathbf{X}$ corresponde a un vector de unos para el término de intercepto, y las demás columnas corresponden a los valores de las variables independientes:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \tag{2.31}$$

revisar la ecuación 2.21. De esta forma, el modelo logit puede ser escrito de manera compacta utilizando la notación matricial:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \mathbf{X}\boldsymbol{\beta} \tag{2.32}$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \tag{2.33}$$

Así, la probabilidad p se puede expresar como:

$$p = \frac{1}{1 + e^{-\mathbf{X}\boldsymbol{\beta}}} \tag{2.34}$$

Comparar la ecuación anterior con la ecuación 2.13. Esta notación matricial simplifica la implementación y la derivación de los estimadores de los coeficientes en la regresión logística. Para estimar los coeficientes $\boldsymbol{\beta}$ en la regresión logística, se utiliza el método de máxima verosimilitud. La función de verosimilitud $L(\boldsymbol{\beta})$ se define como el producto de las probabilidades de las observaciones dadas las variables independientes, recordemos la ecuación 2.14:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \tag{2.35}$$

donde y_i es el valor observado de la variable dependiente para la i-ésima observación, y p_i es la probabilidad predicha de que $Y_i = 1$. La función de log-verosimilitud, que es más fácil de maximizar, se obtiene tomando el logaritmo natural de la función de verosimilitud (**??**):

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{2.36}$$

Sustituyendo $p_i = \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}}$, donde $\mathbf{X}_i$ es la i-ésima fila de la matriz de diseño $\mathbf{X}$, obtenemos:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i (\mathbf{X}_i \boldsymbol{\beta}) - \log(1 + e^{\mathbf{X}_i \boldsymbol{\beta}}) \right] \tag{2.37}$$

Para encontrar los valores de $\boldsymbol{\beta}$ que maximizan la función de log-verosimilitud, se utiliza un algoritmo iterativo como el método de Newton-Raphson. Este método requiere calcular el gradiente y la matriz Hessiana de la función de log-verosimilitud.

El gradiente de la función de log-verosimilitud con respecto a $\boldsymbol{\beta}$ es (2.27 y **??**):

$$\nabla \log L(\boldsymbol{\beta}) = \mathbf{X}^T (\mathbf{y} - \mathbf{p}) \tag{2.38}$$

donde $\mathbf{y}$ es el vector de valores observados y $\mathbf{p}$ es el vector de probabilidades predichas.

La matriz Hessiana de la función de log-verosimilitud es (2.28 y **??**):

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^T \mathbf{W} \mathbf{X} \tag{2.39}$$

donde $\mathbf{W}$ es una matriz diagonal de pesos con elementos $w_i = p_i(1 - p_i)$.

El método de Newton-Raphson actualiza los coeficientes $\boldsymbol{\beta}$ de la siguiente manera:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [\mathbf{H}(\boldsymbol{\beta}^{(t)})]^{-1} \nabla \log L(\boldsymbol{\beta}^{(t)}) \tag{2.40}$$

Iterando este proceso hasta que la diferencia entre $\boldsymbol{\beta}^{(t+1)}$ y $\boldsymbol{\beta}^{(t)}$ sea menor que un umbral predefinido (2.26 y 2.29), se obtienen los estimadores de máxima verosimilitud para los coeficientes de la regresión logística.

2.4. Validación del Modelo

Una vez que se han estimado los coeficientes del modelo de regresión logística, es importante validar el modelo para asegurarse de que proporciona predicciones precisas.

2.4.1. Curva ROC y AUC

La curva ROC (Receiver Operating Characteristic) es una herramienta gráfica utilizada para evaluar el rendimiento de un modelo de clasificación binaria. El área bajo la curva (AUC) mide la capacidad del modelo para distinguir entre las clases. La curva ROC (Receiver Operating Characteristic) es una representación gráfica de la sensibilidad (verdaderos positivos) frente a 1 - especificidad (falsos positivos). El área bajo la curva (AUC) mide la capacidad del modelo para distinguir entre las clases.

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.41}$$

$$\text{Especificidad} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.42}$$

2.4.2. Matriz de Confusión

La matriz de confusión es una tabla que resume el rendimiento de un modelo de clasificación al comparar las predicciones del modelo con los valores reales. Los términos en la matriz de confusión incluyen verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. La matriz de confusión es una tabla que muestra el rendimiento del modelo comparando las predicciones con los valores reales. Los términos incluyen:

- **Verdaderos Positivos (TP)**: Predicciones correctas de la clase positiva.
- **Falsos Positivos (FP)**: Predicciones incorrectas de la clase positiva.
- **Verdaderos Negativos (TN)**: Predicciones correctas de la clase negativa.
- **Falsos Negativos (FN)**: Predicciones incorrectas de la clase negativa.

	Predicción Positiva	**Predicción Negativa**
Real Positiva	TP	FN
Real Negativa	FP	TN

Cuadro 2.1: Matriz de Confusión

2.4.3. Precisión, Recall y F1-Score

$$\text{Precisión} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.43}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.44}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}} \tag{2.45}$$

2.4.4. Log-Loss

La pérdida logarítmica (Log-Loss) mide la precisión de las probabilidades predichas. La fórmula es:

$$\text{Log-Loss} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{2.46}$$

donde y_i son los valores reales y p_i son las probabilidades predichas.

2.4.5. K-Fold Cross-Validation

La validación cruzada es una técnica para evaluar la capacidad de generalización de un modelo. Existen varios tipos de validación cruzada:

En K-Fold Cross-Validation, los datos se dividen en K subconjuntos. El modelo se entrena K veces, cada vez utilizando K-1 subconjuntos para el entrenamiento y el subconjunto restante para la validación.

$$\text{Error Medio} = \frac{1}{K}\sum_{k=1}^{K}\text{Error}_k \tag{2.47}$$

2.4.6. Leave-One-Out Cross-Validation (LOOCV)

En LOOCV, cada observación se usa una vez como conjunto de validación y las restantes como conjunto de entrenamiento. Este método es computacionalmente costoso pero útil para conjuntos de datos pequeños.

2.5. Ajuste y Sobreajuste del Modelo

El ajuste adecuado del modelo es crucial para evitar el sobreajuste (*overfitting*) y el subajuste (*underfitting).* Ambos problemas impactan negativamente en la capacidad del modelo para generalizar a datos nuevos.

- **Sobreajuste**: El sobreajuste ocurre cuando un modelo se ajusta demasiado bien a los datos de entrenamiento, capturando no solo los patrones reales, sino también el ruido y las peculiaridades del conjunto de datos. Esto resulta en un modelo que funciona extremadamente bien en los datos de entrenamiento, pero que falla al ser evaluado con datos nuevos. Los síntomas típicos del sobreajuste incluyen una **alta precisión en el conjunto de entrenamiento** y **una baja precisión en el conjunto de validación o prueba**.

  Algunos factores que pueden llevar al sobreajuste incluyen:

  a) **Modelos excesivamente complejos**: Un modelo con demasiados parámetros o características puede adaptarse demasiado a los datos.

  b) **Conjuntos de datos pequeños**: Cuando los datos de entrenamiento no son suficientes para capturar la variedad de situaciones posibles.

  c) **Falta de regularización**: Si no se utiliza ninguna técnica de regularización, el modelo puede ajustarse de manera demasiado precisa a los datos.

- **Subajuste**: El subajuste ocurre cuando el modelo no es capaz de capturar los patrones subyacentes de los datos, generalmente porque es demasiado simple o porque el entrenamiento no ha sido suficiente. En este caso, tanto en el conjunto de entrenamiento como en el de validación se observan errores elevados, lo que indica que el modelo no está aprendiendo correctamente.

  Las causas comunes del subajuste incluyen:

  a) **Modelos demasiado simples**: Modelos con pocos parámetros o de baja complejidad, como la regresión lineal para problemas no lineales.

  b) **Insuficiente entrenamiento**: El modelo no ha sido entrenado adecuadamente, lo que puede requerir más iteraciones de entrenamiento o ajustes en los hiperparámetros.

  c) **Datos insuficientemente procesados**: El preprocesamiento incorrecto o insuficiente de los datos puede impedir que el modelo identifique patrones importantes.

Existen varias estrategias para prevenir el sobreajuste, incluyendo la selección de un modelo adecuado, el uso de más datos y la aplicación de técnicas de regularización. A continuación, se describen algunas de las técnicas más comunes:

- **Regularización** La regularización es una técnica que añade un término de penalización a la función de costo, con el fin de reducir la complejidad del modelo. Las dos formas más comunes de regularización son:

  a) **Regresión Lasso (L1)**: Esta técnica añade una penalización proporcional al valor absoluto de los coeficientes del modelo. La regularización Lasso tiende a reducir algunos coeficientes a cero, lo que lleva a la selección automática de características.

$$\text{Función de Costo L1} = \sum_{i=1}^{n}(y_i - \hat{y}i)^2 + \lambda \sum j = 1^p |\beta_j| \tag{2.48}$$

  b) **Regresión Ridge (L2)**: Esta técnica penaliza el cuadrado de los coeficientes del modelo. A diferencia de Lasso, Ridge no fuerza los coeficientes a ser exactamente cero, sino que los reduce.

$$\text{Función de Costo L2} = \sum_{i=1}^{n}(y_i - \hat{y}i)^2 + \lambda \sum j = 1^p \beta_j^2 \tag{2.49}$$

- **Elastic Net**: Combina las penalizaciones L1 y L2, con un término de penalización ajustable para controlar la importancia de cada uno.
- **Validación Cruzada**: Otra técnica importante para prevenir el sobreajuste es la validación cruzada, que ayuda a estimar el rendimiento del modelo en datos no observados. El método más común es la validación cruzada $K - fold$, donde los datos se dividen en K subconjuntos, y el modelo se entrena K veces, utilizando $K - 1$ subconjuntos para el entrenamiento y el restante para la validación.
- **Aumentar el Tamaño del Conjunto de Datos**: Una estrategia efectiva, aunque a menudo difícil de implementar, es aumentar el tamaño del conjunto de datos. Más datos permiten al modelo capturar mejor los patrones subyacentes y generalizar de manera más efectiva, disminuyendo el riesgo de sobreajuste.
- **Reducción de Características** (Feature Selection): Reducir el número de características (variables) irrelevantes o redundantes puede simplificar el modelo y mejorar su capacidad de generalización. Técnicas como la selección de características basada en su importancia o la eliminación de características altamente correlacionadas pueden ayudar a evitar el sobreajuste.

Capítulo 3

Preparación de Datos y Selección de Variables

3.1. Introducción

La preparación de datos y la selección de variables son pasos cruciales en el proceso de modelado estadístico. Un modelo bien preparado y con las variables adecuadas puede mejorar significativamente la precisión y la interpretabilidad del modelo. Este capítulo proporciona una revisión detallada de las técnicas de limpieza de datos, tratamiento de datos faltantes, codificación de variables categóricas y selección de variables.

3.2. Importancia de la Preparación de Datos

La calidad de los datos es fundamental para el éxito de cualquier análisis estadístico. Los datos sin limpiar pueden llevar a modelos inexactos y conclusiones erróneas. La preparación de datos incluye varias etapas:

- Limpieza de datos: es el proceso de detectar y corregir (o eliminar) los datos incorrectos, incompletos o irrelevantes. Este proceso incluye:
  - Eliminación de duplicados
  - Corrección de errores tipográficos
  - Consistencia de formato
  - Tratamiento de valores extremos (outliers)
- Tratamiento de datos faltantes: Los datos faltantes son un problema común en los conjuntos de datos y pueden afectar la calidad de los modelos. Hay varias estrategias para manejar los datos faltantes:
  - **Eliminación de Datos Faltantes**: Se eliminan las filas o columnas con datos faltantes.
  - **Imputación**: Se reemplazan los valores faltantes con estimaciones, como la media, la mediana o la moda. Una técnica común es reemplazar los valores faltantes con la media de la variable. Esto se puede hacer de la siguiente manera:

$$x_i = \begin{cases} x_i & \text{si } x_i \text{ no es faltante} \\ \bar{x} & \text{si } x_i \text{ es faltante} \end{cases}$$

    donde $\bar{x}$ es la media de la variable.
  - **Modelos Predictivos**: Se utilizan modelos predictivos para estimar los valores faltantes.

- Codificación de variables categóricas: Las variables categóricas deben ser convertidas a un formato numérico antes de ser usadas en un modelo de regresión logística. Hay varias técnicas para codificar variables categóricas:

  - La codificación one-hot crea una columna binaria para cada categoría. Por ejemplo, si tenemos una variable categórica con tres categorías (A, B, C), se crean tres columnas:

$$\begin{aligned} \text{A} &= [1,0,0] \\ \text{B} &= [0,1,0] \\ \text{C} &= [0,0,1] \end{aligned}$$

  - La codificación ordinal asigna un valor entero único a cada categoría, preservando el orden natural de las categorías. Por ejemplo:

$$\begin{aligned} \text{Bajo} &= 1 \\ \text{Medio} &= 2 \\ \text{Alto} &= 3 \end{aligned}$$

- Selección y transformación de variables: La selección de variables es el proceso de elegir las variables más relevantes para el modelo. Existen varias técnicas para la selección de variables:

  - Métodos de Filtrado: Los métodos de filtrado seleccionan variables basadas en criterios estadísticos, como la correlación o la chi-cuadrado. Algunas técnicas comunes incluyen:
    - **Análisis de Correlación**: Se seleccionan variables con alta correlación con la variable dependiente y baja correlación entre ellas.
    - **Pruebas de Chi-cuadrado**: Se utilizan para variables categóricas para determinar la asociación entre la variable independiente y la variable dependiente.
  - Métodos de Wrapper: Los métodos de wrapper evalúan múltiples combinaciones de variables y seleccionan la combinación que optimiza el rendimiento del modelo. Ejemplos incluyen:
    - **Selección hacia Adelante**: Comienza con un modelo vacío y agrega variables una por una, seleccionando la variable que mejora más el modelo en cada paso.
    - **Selección hacia Atrás**: Comienza con todas las variables y elimina una por una, removiendo la variable que tiene el menor impacto en el modelo en cada paso.
    - **Selección Paso a Paso**: Combina la selección hacia adelante y hacia atrás, agregando y eliminando variables según sea necesario.
  - Métodos Basados en Modelos: Los métodos basados en modelos utilizan técnicas de regularización como Lasso y Ridge para seleccionar variables. Estas técnicas añaden un término de penalización a la función de costo para evitar el sobreajuste.
    - Regresión Lasso: La regresión Lasso (Least Absolute Shrinkage and Selection Operator) añade una penalización L_1 a la función de costo:

$$J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

      donde λ es el parámetro de regularización que controla la cantidad de penalización.
    - Regresión Ridge: La regresión Ridge añade una penalización L_2 a la función de costo:

$$J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

      donde λ es el parámetro de regularización.

Capítulo 4

Diagnóstico del Modelo y Ajuste de Parámetros

El diagnóstico del modelo y el ajuste de parámetros son pasos esenciales para mejorar la precisión y la robustez de los modelos de regresión logística. Este capítulo se enfoca en las técnicas para diagnosticar problemas en los modelos y en métodos para ajustar los parámetros de manera óptima.

- Diagnóstico del Modelo: El diagnóstico del modelo implica evaluar el rendimiento del modelo y detectar posibles problemas, como el sobreajuste, la multicolinealidad y la influencia de puntos de datos individuales.
- Residuos: Los residuos son las diferencias entre los valores observados y los valores predichos por el modelo. El análisis de residuos puede revelar patrones que indican problemas con el modelo.

$$\text{Residuo}_i = y_i - \hat{y}_i$$

  - Residuos Estudiantizados: Los residuos estudiantizados se ajustan por la variabilidad del residuo y se utilizan para detectar outliers.

$$r_i = \frac{\text{Residuo}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

    donde h_i es el leverage del punto de datos.

- Influencia: La influencia mide el impacto de un punto de datos en los coeficientes del modelo. Los puntos con alta influencia pueden distorsionar el modelo.
  - Distancia de Cook: La distancia de Cook es una medida de la influencia de un punto de datos en los coeficientes del modelo.

$$D_i = \frac{r_i^2}{p} \cdot \frac{h_i}{1 - h_i}$$

    donde p es el número de parámetros en el modelo.

- Multicolinealidad: La multicolinealidad ocurre cuando dos o más variables independientes están altamente correlacionadas. Esto puede inflar las varianzas de los coeficientes y hacer que el modelo sea inestable.
  - Factor de Inflación de la Varianza (VIF): El VIF mide cuánto se inflan las varianzas de los coeficientes debido a la multicolinealidad.

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el coeficiente de determinación de la regresión de la variable j contra todas las demás variables.

- Ajuste de Parámetros: El ajuste de parámetros implica seleccionar los valores óptimos para los hiperparámetros del modelo. Esto puede mejorar el rendimiento y prevenir el sobreajuste.
  - Grid Search: El grid search es un método exhaustivo para ajustar los parámetros. Se define una rejilla de posibles valores de parámetros y se evalúa el rendimiento del modelo para cada combinación.
  - Random Search: El random search selecciona aleatoriamente combinaciones de valores de parámetros dentro de un rango especificado. Es menos exhaustivo que el grid search, pero puede ser más eficiente.
  - Bayesian Optimization: La optimización bayesiana utiliza modelos probabilísticos para seleccionar iterativamente los valores de parámetros más prometedores.

Capítulo 5

Interpretación de los Resultados

Interpretar correctamente los resultados de un modelo de regresión logística es esencial para tomar decisiones informadas. Este capítulo se centra en la interpretación de los coeficientes del modelo, las odds ratios, los intervalos de confianza y la significancia estadística.

5.1. Coeficientes de Regresión Logística

Los coeficientes de regresión logística representan la relación entre las variables independientes y la variable dependiente en términos de log-odds.

Cada coeficiente β_j en el modelo de regresión logística se interpreta como el cambio en el log-odds de la variable dependiente por unidad de cambio en la variable independiente X_j.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

- **Coeficiente Positivo**: Un coeficiente positivo indica que un aumento en la variable independiente está asociado con un aumento en el log-odds de la variable dependiente.
- **Coeficiente Negativo**: Un coeficiente negativo indica que un aumento en la variable independiente está asociado con una disminución en el log-odds de la variable dependiente.

5.2. Odds Ratios

Las odds ratios proporcionan una interpretaciónón más intuitiva de los coeficientes de regresión logística. La odds ratio para una variable independiente X_j se calcula como e^{β_j}.

$$\text{OR}_j = e^{\beta_j}$$

- **OR >1**: Un OR mayor que 1 indica que un aumento en la variable independiente está asociado con un aumento en las odds de la variable dependiente.
- **OR <1**: Un OR menor que 1 indica que un aumento en la variable independiente está asociado con una disminución en las odds de la variable dependiente.
- **OR = 1**: Un OR igual a 1 indica que la variable independiente no tiene efecto sobre las odds de la variable dependiente.

5.3. Intervalos de Confianza

Los intervalos de confianza proporcionan una medida de la incertidumbre asociada con los estimadores de los coeficientes. Un intervalo de confianza del 95 % para un coeficiente β_j indica que, en el 95 % de las muestras, el intervalo contendrá el valor verdadero de β_j.

Para calcular un intervalo de confianza del 95 % para un coeficiente β_j, utilizamos la fórmula:

$$\beta_j \pm 1{,}96 \cdot \mathrm{SE}(\beta_j)$$

donde $\mathrm{SE}(\beta_j)$ es el error estándar de β_j.

5.4. Significancia Estadística

La significancia estadística se utiliza para determinar si los coeficientes del modelo son significativamente diferentes de cero. Esto se evalúa mediante pruebas de hipótesis.

Para cada coeficiente β_j, la hipótesis nula H_0 es que $\beta_j = 0$. La hipótesis alternativa H_a es que $\beta_j \neq 0$.

El p-valor indica la probabilidad de obtener un coeficiente tan extremo como el observado, asumiendo que la hipótesis nula es verdadera. Un p-valor menor que el nivel de significancia α (típicamente 0.05) indica que podemos rechazar la hipótesis nula.

Capítulo 6

Regresión Logística Multinomial y Análisis de Supervivencia

La regresión logística multinomial y el análisis de supervivencia son extensiones de la regresión logística binaria. Este capítulo se enfoca en las técnicas y aplicaciones de estos métodos avanzados.

6.1. Regresión Logística Multinomial

La regresión logística multinomial se utiliza cuando la variable dependiente tiene más de dos categorías.

El modelo de regresión logística multinomial generaliza el modelo binario para manejar múltiples categorías. La probabilidad de que una observación pertenezca a la categoría k se expresa como:

$$P(Y = k) = \frac{e^{\beta_{0k} + \beta_{1k} X_1 + \ldots + \beta_{nk} X_n}}{\sum_{j=1}^{K} e^{\beta_{0j} + \beta_{1j} X_1 + \ldots + \beta_{nj} X_n}}$$

Los coeficientes del modelo multinomial se estiman utilizando máxima verosimilitud, similar a la regresión logística binaria.

6.2. Análisis de Supervivencia

El análisis de supervivencia se utiliza para modelar el tiempo hasta que ocurre un evento de interés, como la muerte o la falla de un componente.

La función de supervivencia $S(t)$ describe la probabilidad de que una observación sobreviva más allá del tiempo t:

$$S(t) = P(T > t)$$

El modelo de Cox es un modelo de regresión semiparamétrico utilizado para analizar datos de supervivencia:

$$h(t|X) = h_0(t) e^{\beta_1 X_1 + \ldots + \beta_p X_p}$$

donde $h(t|X)$ es la tasa de riesgo en el tiempo t dado el vector de covariables X y $h_0(t)$ es la tasa de riesgo basal.

Parte II

SEGUNDA PARTE: ANALISIS DE SUPERVIVENCIA

Capítulo 7

Introducción al Análisis de Supervivencia

Introducción

El análisis de supervivencia es una rama de la estadística que se ocupa del análisis del tiempo que transcurre hasta que ocurre un evento de interés, comúnmente referido como "tiempo de falla". Este campo es ampliamente utilizado en medicina, biología, ingeniería, ciencias sociales, y otros campos. En el análisis de supervivencia, un *evento* se refiere a la ocurrencia de un evento específico, como la muerte, la falla de un componente, la recaída de una enfermedad, etc. El *tiempo de supervivencia* es el tiempo que transcurre desde un punto de inicio definido hasta la ocurrencia del evento.

La censura ocurre cuando la información completa sobre el tiempo hasta el evento no está disponible para todos los individuos en el estudio. Hay tres tipos principales de censura:

- **Censura a la derecha:** Ocurre cuando el evento de interés no se ha observado para algunos sujetos antes del final del estudio.
- **Censura a la izquierda:** Ocurre cuando el evento de interés ocurrió antes del inicio del periodo de observación.
- **Censura por intervalo:** Ocurre cuando el evento de interés se sabe que ocurrió en un intervalo de tiempo, pero no se conoce el momento exacto.

Función de Supervivencia y Función de Riesgo

La función de supervivencia, $S(t)$, se define como la probabilidad de que un individuo sobreviva más allá de un tiempo t, se puede expresar como:

$$S(t) = P(T > t) \tag{7.1}$$

donde T es una variable aleatoria que representa el tiempo hasta el evento. La función de supervivencia tiene las siguientes propiedades:

- $S(0) = 1$: Esto indica que al inicio (tiempo $t = 0$), la probabilidad de haber experimentado el evento es cero, por lo tanto, la supervivencia es del $100\,\%$.
- $\lim_{t\to\infty} S(t) = 0$: A medida que el tiempo tiende al infinito, la probabilidad de que cualquier individuo aún no haya experimentado el evento tiende a cero.
- $S(t)$ es una función no creciente: Esto significa que a medida que el tiempo avanza, la probabilidad de supervivencia no aumenta.

La función de densidad de probabilidad $f(t)$ describe la probabilidad de que el evento ocurra en un instante de tiempo específico, esta se define como:

$$f(t) = \frac{dF(t)}{dt} \tag{7.2}$$

donde $F(t)$ es la función de distribución acumulada, $F(t) = P(T \leq t)$. La relación entre $S(t)$ y $f(t)$ es:

$$f(t) = -\frac{dS(t)}{dt} \tag{7.3}$$

Función de Riesgo

La función de riesgo, $\lambda(t)$, también conocida como función de tasa de fallas o *hazard rate*, se define como la tasa instantánea de ocurrencia del evento en el tiempo t, dado que el individuo ha sobrevivido hasta el tiempo t. Matemáticamente, se expresa como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{7.4}$$

Esto se puede reescribir usando $f(t)$ y $S(t)$ como:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{7.5}$$

Relación entre Función de Supervivencia y Función de Riesgo

La función de supervivencia y la función de riesgo están relacionadas a través de la siguiente ecuación:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\, du\right) \tag{7.6}$$

Esta fórmula se deriva del hecho de que la función de supervivencia es la probabilidad acumulada de no haber experimentado el evento hasta el tiempo t, y $\lambda(t)$ es la tasa instantánea de ocurrencia del evento.

La función de riesgo también puede ser expresada como:

$$\lambda(t) = -\frac{d}{dt} \log S(t) \tag{7.7}$$

La relación entre la función de supervivencia y la función de riesgo se puede deducir integrando la función de riesgo:

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(u)\, du\right) \\ \log S(t) &= -\int_0^t \lambda(u)\, du \\ \frac{d}{dt} \log S(t) &= -\lambda(t) \\ \lambda(t) &= -\frac{d}{dt} \log S(t) \end{aligned}$$

Ejemplo de Cálculo

Supongamos que tenemos una muestra de tiempos de supervivencia $T_1, T_2, \ldots, T_n$. Podemos estimar la función de supervivencia empírica como:

$$\hat{S}(t) = \frac{\text{Número de individuos que sobreviven más allá de } t}{\text{Número total de individuos en riesgo en } t}$$

y la función de riesgo empírica como:

$$\hat{\lambda}(t) = \frac{\text{Número de eventos en } t}{\text{Número de individuos en riesgo en } t}$$

El análisis de supervivencia es una herramienta poderosa para analizar datos de tiempo hasta evento. La función de supervivencia tiene varias propiedades importantes:

- $S(0) = 1$: Indica que la probabilidad de haber experimentado el evento en el tiempo 0 es cero.
- $\lim_{t\to\infty} S(t) = 0$: A medida que el tiempo tiende al infinito, la probabilidad de supervivencia tiende a cero.
- $S(t)$ es una función no creciente: A medida que el tiempo avanza, la probabilidad de supervivencia no aumenta.

Si la función de densidad de probabilidad $f(t)$ del tiempo de supervivencia T es conocida, la función de supervivencia puede derivarse como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t) \\ &= 1 - \int_0^t f(u)\, du \end{aligned}$$

donde $F(t)$ es la función de distribución acumulada. Consideremos un ejemplo donde el tiempo de supervivencia T sigue una distribución exponencial con tasa λ. La función de densidad de probabilidad $f(t)$ es:

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$

La función de distribución acumulada $F(t)$ es:

$$F(t) = \int_0^t \lambda e^{-\lambda u}\, du = 1 - e^{-\lambda t}$$

Por lo tanto, la función de supervivencia $S(t)$ es:

$$S(t) = 1 - F(t) = e^{-\lambda t}$$

La función de riesgo, $\lambda(t)$, proporciona la tasa instantánea de ocurrencia del evento en el tiempo t, dado que el individuo ha sobrevivido hasta el tiempo t. Matemáticamente, se define como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

La función de riesgo se puede relacionar con la función de densidad de probabilidad $f(t)$ y la función de supervivencia $S(t)$ de la siguiente manera:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

La derivación de $\lambda(t)$ se basa en la definición condicional de la probabilidad:

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\frac{P(t \leq T < t+\Delta t \text{ y } T \geq t)}{P(T \geq t)}}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\frac{P(t \leq T < t+\Delta t)}{P(T \geq t)}}{\Delta t} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

La función de supervivencia y la función de riesgo están estrechamente relacionadas. La relación se expresa mediante la siguiente ecuación:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,du\right)$$

Para deducir esta relación, consideramos la derivada logarítmica de la función de supervivencia:

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(u)\,du\right) \\ \log S(t) &= -\int_0^t \lambda(u)\,du \\ \frac{d}{dt}\log S(t) &= -\lambda(t) \\ \lambda(t) &= -\frac{d}{dt}\log S(t) \end{aligned}$$

La función de riesgo, $\lambda(t)$, se interpreta como la tasa instantánea de ocurrencia del evento por unidad de tiempo, dado que el individuo ha sobrevivido hasta el tiempo t. Es una medida local del riesgo de falla en un instante específico.

Ejemplo de Cálculo de $\lambda(t)$

Consideremos nuevamente el caso donde el tiempo de supervivencia T sigue una distribución exponencial con tasa λ. La función de densidad de probabilidad $f(t)$ es:

$$f(t) = \lambda e^{-\lambda t}$$

La función de supervivencia $S(t)$ es:

$$S(t) = e^{-\lambda t}$$

La función de riesgo $\lambda(t)$ se calcula como:

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} \\ &= \lambda \end{aligned}$$

En este caso, $\lambda(t)$ es constante y igual a λ, lo que es una característica de la distribución exponencial.

Funciones de Riesgo Acumulada y Media Residual

La función de riesgo acumulada $H(t)$ se define como:

$$H(t) = \int_0^t \lambda(u)\,du$$

Esta función proporciona la suma acumulada de la tasa de riesgo hasta el tiempo t.

La función de vida media residual $e(t)$ se define como la esperanza del tiempo de vida restante dado que el individuo ha sobrevivido hasta el tiempo t:

$$e(t) = \mathbb{E}[T - t \mid T > t] = \int_t^\infty S(u)\,du$$

Ejemplo de Cálculo de Función de Riesgo Acumulada y Vida Media Residual

Consideremos nuevamente la distribución exponencial con tasa λ. La función de riesgo acumulada $H(t)$ es:

$$\begin{aligned} H(t) &= \int_0^t \lambda \, du \\ &= \lambda t \end{aligned}$$

La función de vida media residual $e(t)$ es:

$$\begin{aligned} e(t) &= \int_t^\infty e^{-\lambda u} \, du \\ &= \left[\frac{-1}{\lambda} e^{-\lambda u} \right]_t^\infty \\ &= \frac{1}{\lambda} e^{-\lambda t} \\ &= \frac{1}{\lambda} \end{aligned}$$

En este caso, la vida media residual es constante e igual a $\frac{1}{\lambda}$, otra característica de la distribución exponencial.

Capítulo 8

Estimador de Kaplan-Meier

El estimador de Kaplan-Meier, también conocido como la función de supervivencia empírica, es una herramienta no paramétrica para estimar la función de supervivencia a partir de datos censurados. Este método es especialmente útil cuando los tiempos de evento están censurados a la derecha.

Definición del Estimador de Kaplan-Meier

El estimador de Kaplan-Meier se define como:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

donde:

- t_i es el tiempo del i-ésimo evento,
- d_i es el número de eventos que ocurren en t_i,
- n_i es el número de individuos en riesgo justo antes de t_i.

Propiedades del Estimador de Kaplan-Meier

El estimador de Kaplan-Meier tiene las siguientes propiedades:

- Es una función escalonada que disminuye en los tiempos de los eventos observados.
- Puede manejar datos censurados a la derecha.
- Proporciona una estimación no paramétrica de la función de supervivencia.

Función Escalonada

La función escalonada del estimador de Kaplan-Meier significa que $\hat{S}(t)$ permanece constante entre los tiempos de los eventos y disminuye en los tiempos de los eventos. Matemáticamente, si t_i es el tiempo del i-ésimo evento, entonces:

$$\hat{S}(t) = \hat{S}(t_i) \quad \text{para } t_i \leq t < t_{i+1}$$

El estimador de Kaplan-Meier maneja datos censurados a la derecha al ajustar la estimación de la función de supervivencia sólo en los tiempos en que ocurren eventos. Si un individuo es censurado antes de experimentar el evento, no contribuye a la disminución de $\hat{S}(t)$ en el tiempo de censura. Esto asegura que la censura no sesga la estimación de la supervivencia.

El estimador de Kaplan-Meier es no paramétrico porque no asume ninguna forma específica para la distribución de los tiempos de supervivencia. En cambio, utiliza la información empírica disponible para estimar la función de supervivencia.

La deducción del estimador de Kaplan-Meier se basa en el principio de probabilidad condicional. Consideremos un conjunto de tiempos de supervivencia observados $t_1, t_2, \ldots, t_k$ con eventos en cada uno de estos tiempos. El estimador de la probabilidad de supervivencia más allá del tiempo t es el producto de las probabilidades de sobrevivir más allá de cada uno de los tiempos de evento observados hasta t.

La probabilidad de sobrevivir más allá de t_i, dado que el individuo ha sobrevivido justo antes de t_i, es:

$$P(T > t_i \mid T \geq t_i) = 1 - \frac{d_i}{n_i}$$

donde d_i es el número de eventos en t_i y n_i es el número de individuos en riesgo justo antes de t_i.

La probabilidad de sobrevivir más allá de un tiempo t cualquiera, dada la secuencia de tiempos de evento, es el producto de las probabilidades condicionales de sobrevivir más allá de cada uno de los tiempos de evento observados hasta t. Así, el estimador de Kaplan-Meier se obtiene como:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

Ejemplo de Cálculo

Supongamos que tenemos los siguientes tiempos de supervivencia observados para cinco individuos: 2, 3, 5, 7, 8. Supongamos además que tenemos censura a la derecha en el tiempo 10. Los tiempos de evento y el número de individuos en riesgo justo antes de cada evento son:

Tiempo (t_i)	Eventos (d_i)	En Riesgo (n_i)
2	1	5
3	1	4
5	1	3
7	1	2
8	1	1

Cuadro 8.1: Ejemplo de cálculo del estimador de Kaplan-Meier

Usando estos datos, el estimador de Kaplan-Meier se calcula como:

$$\begin{aligned}
\hat{S}(2) &= 1 - \frac{1}{5} = 0{,}8 \\
\hat{S}(3) &= 0{,}8 \times \left(1 - \frac{1}{4}\right) = 0{,}8 \times 0{,}75 = 0{,}6 \\
\hat{S}(5) &= 0{,}6 \times \left(1 - \frac{1}{3}\right) = 0{,}6 \times 0{,}6667 = 0{,}4 \\
\hat{S}(7) &= 0{,}4 \times \left(1 - \frac{1}{2}\right) = 0{,}4 \times 0{,}5 = 0{,}2 \\
\hat{S}(8) &= 0{,}2 \times \left(1 - \frac{1}{1}\right) = 0{,}2 \times 0 = 0
\end{aligned}$$

Intervalos de Confianza para el Estimador de Kaplan-Meier

Para calcular intervalos de confianza para el estimador de Kaplan-Meier, se puede usar la transformación logarítmica y la aproximación normal. Un intervalo de confianza aproximado para $\log(-\log(\hat{S}(t)))$

se obtiene como:

$$\log(-\log(\hat{S}(t))) \pm z_{\alpha/2}\sqrt{\frac{1}{d_i(n_i - d_i)}}$$

donde $z_{\alpha/2}$ es el percentil correspondiente de la distribución normal estándar.

Transformación Logarítmica Inversa

La transformación logarítmica inversa se utiliza para obtener los límites del intervalo de confianza para $S(t)$:

$$\hat{S}(t) = \exp\left(-\exp\left(\log(-\log(\hat{S}(t))) \pm z_{\alpha/2}\sqrt{\frac{1}{d_i(n_i - d_i)}}\right)\right)$$

Cálculo Detallado de Intervalos de Confianza

Para un cálculo más detallado de los intervalos de confianza, consideremos un tiempo específico t_j. La varianza del estimador de Kaplan-Meier en t_j se puede estimar usando Greenwood's formula:

$$\text{Var}(\hat{S}(t_j)) = \hat{S}(t_j)^2 \sum_{t_i \leq t_j} \frac{d_i}{n_i(n_i - d_i)}$$

El intervalo de confianza aproximado para $\hat{S}(t_j)$ es entonces:

$$\hat{S}(t_j) \pm z_{\alpha/2}\sqrt{\text{Var}(\hat{S}(t_j))}$$

Ejemplo de Intervalo de Confianza

Supongamos que en el ejemplo anterior queremos calcular el intervalo de confianza para $\hat{S}(3)$. Primero, calculamos la varianza:

$$\begin{aligned}
\text{Var}(\hat{S}(3)) &= \hat{S}(3)^2\left(\frac{1}{5 \times 4} + \frac{1}{4 \times 3}\right) \\
&= 0{,}6^2\left(\frac{1}{20} + \frac{1}{12}\right) \\
&= 0{,}36\,(0{,}05 + 0{,}0833) \\
&= 0{,}36 \times 0{,}1333 \\
&= 0{,}048
\end{aligned}$$

El intervalo de confianza es entonces:

$$0{,}6 \pm 1{,}96\sqrt{0{,}048} = 0{,}6 \pm 1{,}96 \times 0{,}219 = 0{,}6 \pm 0{,}429$$

Por lo tanto, el intervalo de confianza para $\hat{S}(3)$ es aproximadamente $(0{,}171, 1{,}029)$. Dado que una probabilidad no puede exceder 1, ajustamos el intervalo a $(0{,}171, 1{,}0)$.

El estimador de Kaplan-Meier proporciona una estimación empírica de la función de supervivencia que es fácil de interpretar y calcular. Su capacidad para manejar datos censurados lo hace especialmente útil en estudios de supervivencia.

El estimador de Kaplan-Meier es una herramienta poderosa para estimar la función de supervivencia en presencia de datos censurados. Su cálculo es relativamente sencillo y proporciona una estimación no paramétrica robusta de la supervivencia a lo largo del tiempo. La interpretación adecuada de este estimador y su intervalo de confianza asociado es fundamental para el análisis de datos de supervivencia.

Capítulo 9

Comparación de Curvas de Supervivencia

Comparar curvas de supervivencia es crucial para determinar si existen diferencias significativas en las tasas de supervivencia entre diferentes grupos. Las pruebas de hipótesis, como el test de log-rank, son herramientas comunes para esta comparación.

Test de Log-rank

El test de log-rank se utiliza para comparar las curvas de supervivencia de dos o más grupos. La hipótesis nula es que no hay diferencia en las funciones de riesgo entre los grupos.

Fórmula del Test de Log-rank

El estadístico del test de log-rank se define como:

$$\chi^2 = \frac{\left(\sum_{i=1}^{k}(O_i - E_i)\right)^2}{\sum_{i=1}^{k} V_i}$$

donde:

- O_i es el número observado de eventos en el grupo i.
- E_i es el número esperado de eventos en el grupo i.
- V_i es la varianza del número de eventos en el grupo i.

Cálculo de E_i y V_i

El número esperado de eventos E_i y la varianza V_i se calculan como:

$$\begin{aligned} E_i &= \frac{d_i \cdot n_i}{n} \\ V_i &= \frac{d_i \cdot (n - d_i) \cdot n_i \cdot (n - n_i)}{n^2 \cdot (n - 1)} \end{aligned}$$

donde:

- d_i es el número total de eventos en el grupo i.
- n_i es el número de individuos en riesgo en el grupo i.
- n es el número total de individuos en todos los grupos.

Grupo	Tiempo (t_i)	Eventos (O_i)	En Riesgo (n_i)
1	2	1	5
1	4	1	4
2	3	1	4
2	5	1	3

Cuadro 9.1: Ejemplo de datos para el test de log-rank

Supongamos que tenemos dos grupos con los siguientes datos de eventos:
Calculemos E_i y V_i para cada grupo:

$$\begin{aligned} E_1 &= \frac{2 \cdot 5}{9} + \frac{2 \cdot 4}{8} = \frac{10}{9} + \frac{8}{8} = 1{,}11 + 1 = 2{,}11 \\ V_1 &= \frac{2 \cdot 7 \cdot 5 \cdot 4}{81 \cdot 8} = \frac{2 \cdot 7 \cdot 5 \cdot 4}{648} = \frac{280}{648} = 0{,}432 \\ E_2 &= \frac{2 \cdot 4}{9} + \frac{2 \cdot 3}{8} = \frac{8}{9} + \frac{6}{8} = 0{,}89 + 0{,}75 = 1{,}64 \\ V_2 &= \frac{2 \cdot 7 \cdot 4 \cdot 4}{81 \cdot 8} = \frac{2 \cdot 7 \cdot 4 \cdot 4}{648} = \frac{224}{648} = 0{,}346 \end{aligned}$$

El estadístico de log-rank se calcula como:

$$\begin{aligned} \chi^2 &= \frac{((1 - 2{,}11) + (1 - 1{,}64))^2}{0{,}432 + 0{,}346} \\ &= \frac{(-1{,}11 - 0{,}64)^2}{0{,}778} \\ &= \frac{3{,}04}{0{,}778} \\ &= 3{,}91 \end{aligned}$$

El valor p se puede obtener comparando χ^2 con una distribución χ^2 con un grado de libertad (dado que estamos comparando dos grupos).

Un valor p pequeño (generalmente menos de 0.05) indica que hay una diferencia significativa en las curvas de supervivencia entre los grupos. Un valor p grande sugiere que no hay suficiente evidencia para rechazar la hipótesis nula de que las curvas de supervivencia son iguales.

Además del test de log-rank, existen otras pruebas para comparar curvas de supervivencia, como el test de Wilcoxon (Breslow), que da más peso a los eventos en tiempos tempranos.

El test de log-rank es una herramienta esencial para comparar curvas de supervivencia entre diferentes grupos. Su cálculo se basa en la diferencia entre los eventos observados y esperados en cada grupo, y su interpretación puede ayudar a identificar diferencias significativas en la supervivencia.

Capítulo 10

Modelos de Riesgos Proporcionales de Cox

El modelo de riesgos proporcionales de Cox, propuesto por David Cox en 1972, es una de las herramientas más utilizadas en el análisis de supervivencia. Este modelo permite evaluar el efecto de varias covariables en el tiempo hasta el evento, sin asumir una forma específica para la distribución de los tiempos de supervivencia.

Definición del Modelo de Cox

El modelo de Cox se define como:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta^T X)$$

donde:

- $\lambda(t \mid X)$ es la función de riesgo en el tiempo t dado el vector de covariables X.
- $\lambda_0(t)$ es la función de riesgo basal en el tiempo t.
- β es el vector de coeficientes del modelo.
- X es el vector de covariables.

Supuesto de Proporcionalidad de Riesgos

El modelo de Cox asume que las razones de riesgo entre dos individuos son constantes a lo largo del tiempo. Matemáticamente, si X_i y X_j son las covariables de dos individuos, la razón de riesgos se expresa como:

$$\frac{\lambda(t \mid X_i)}{\lambda(t \mid X_j)} = \frac{\lambda_0(t) \exp(\beta^T X_i)}{\lambda_0(t) \exp(\beta^T X_j)} = \exp(\beta^T (X_i - X_j))$$

Estimación de los Parámetros

Los parámetros β se estiman utilizando el método de máxima verosimilitud parcial. La función de verosimilitud parcial se define como:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)}$$

donde $R(t_i)$ es el conjunto de individuos en riesgo en el tiempo t_i.

Función de Log-Verosimilitud Parcial

La función de log-verosimilitud parcial es:

$$\log L(\beta) = \sum_{i=1}^{k} \left(\beta^T X_i - \log \sum_{j \in R(t_i)} \exp(\beta^T X_j) \right)$$

Derivadas Parciales y Maximización

Para encontrar los estimadores de máxima verosimilitud, resolvemos el sistema de ecuaciones obtenido al igualar a cero las derivadas parciales de $\log L(\beta)$ con respecto a β:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{k} \left(X_i - \frac{\sum_{j \in R(t_i)} X_j \exp(\beta^T X_j)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)} \right) = 0$$

Cada coeficiente β_i representa el logaritmo de la razón de riesgos asociado con un incremento unitario en la covariable X_i. Un valor positivo de β_i indica que un aumento en X_i incrementa el riesgo del evento, mientras que un valor negativo indica una reducción del riesgo.

Evaluación del Modelo

El modelo de Cox se evalúa utilizando varias técnicas, como el análisis de residuos de Schoenfeld para verificar el supuesto de proporcionalidad de riesgos, y el uso de curvas de supervivencia estimadas para evaluar la bondad de ajuste.

Residuos de Schoenfeld

Los residuos de Schoenfeld se utilizan para evaluar la proporcionalidad de riesgos. Para cada evento en el tiempo t_i, el residuo de Schoenfeld para la covariable X_j se define como:

$$r_{ij} = X_{ij} - \hat{X}_{ij}$$

donde $\hat{X}_{ij}$ es la covariable ajustada.

Curvas de Supervivencia Ajustadas

Las curvas de supervivencia ajustadas se obtienen utilizando la función de riesgo basal estimada y los coeficientes del modelo. La función de supervivencia ajustada se define como:

$$\hat{S}(t \mid X) = \hat{S}_0(t)^{\exp(\beta^T X)}$$

donde $\hat{S}_0(t)$ es la función de supervivencia basal estimada.

Ejemplo de Aplicación del Modelo de Cox

Consideremos un ejemplo con tres covariables: edad, sexo y tratamiento. Supongamos que los datos se ajustan a un modelo de Cox y obtenemos los siguientes coeficientes:

$$\hat{\beta}_{edad} = 0{,}02, \quad \hat{\beta}_{sexo} = -0{,}5, \quad \hat{\beta}_{tratamiento} = 1{,}2$$

La función de riesgo ajustada se expresa como:

$$\lambda(t \mid X) = \lambda_0(t) \exp(0{,}02 \cdot \text{edad} - 0{,}5 \cdot \text{sexo} + 1{,}2 \cdot \text{tratamiento})$$

El modelo de riesgos proporcionales de Cox es una herramienta poderosa para analizar datos de supervivencia con múltiples covariables. Su flexibilidad y la falta de suposiciones fuertes sobre la distribución de los tiempos de supervivencia lo hacen ampliamente aplicable en diversas disciplinas.

Capítulo 11

Diagnóstico y Validación de Modelos de Cox

Una vez ajustado un modelo de Cox, es crucial realizar diagnósticos y validaciones para asegurar que el modelo es apropiado y que los supuestos subyacentes son válidos. Esto incluye la verificación del supuesto de proporcionalidad de riesgos y la evaluación del ajuste del modelo.

Supuesto de Proporcionalidad de Riesgos

El supuesto de proporcionalidad de riesgos implica que la razón de riesgos entre dos individuos es constante a lo largo del tiempo. Si este supuesto no se cumple, las inferencias hechas a partir del modelo pueden ser incorrectas.

Residuos de Schoenfeld

Los residuos de *Schoenfeld* se utilizan para evaluar la proporcionalidad de riesgos. Para cada evento en el tiempo t_i, el residuo de Schoenfeld para la covariable X_j se define como:

$$r_{ij} = X_{ij} - \hat{X}_{ij} \tag{11.1}$$

donde $\hat{X}_{ij}$ es la covariable ajustada. Si los residuos de Schoenfeld no muestran una tendencia sistemática cuando se trazan contra el tiempo, el supuesto de proporcionalidad de riesgos es razonable.

Bondad de Ajuste

La bondad de ajuste del modelo de Cox se evalúa comparando las curvas de supervivencia observadas y ajustadas, y utilizando estadísticas de ajuste global.

Curvas de Supervivencia Ajustadas

Las curvas de supervivencia aaustadas se obtienen utilizando la función de riesgo basal estimada y los coeficientes del modelo. La función de supervivencia ajustada se define como:

$$\hat{S}(t \mid X) = \hat{S}_0(t)^{\exp(\beta^T X)}$$

donde $\hat{S}_0(t)$ es la función de supervivencia basal estimada. Comparar estas curvas con las curvas de Kaplan-Meier para diferentes niveles de las covariables puede proporcionar una validación visual del ajuste del modelo.

Estadísticas de Ajuste Global

Las estadísticas de ajuste global, como el test de la desviación y el test de la bondad de ajuste de Grambsch y Therneau, se utilizan para evaluar el ajuste global del modelo de Cox.

El diagnóstico de influencia identifica observaciones individuales que tienen un gran impacto en los estimados del modelo. Los residuos de devianza y los residuos de martingala se utilizan comúnmente para este propósito.

Residuos de Deviance

Los residuos de deviance se definen como:

$$D_i = \text{sign}(O_i - E_i)\sqrt{-2\left(O_i \log \frac{O_i}{E_i} - (O_i - E_i)\right)}$$

donde O_i es el número observado de eventos y E_i es el número esperado de eventos. Observaciones con residuos de deviance grandes en valor absoluto pueden ser influyentes.

Residuos de Martingala

Los residuos de martingala se definen como:

$$M_i = O_i - E_i$$

donde O_i es el número observado de eventos y E_i es el número esperado de eventos. Los residuos de martingala se utilizan para detectar observaciones que no se ajustan bien al modelo.

Ejemplo de Diagnóstico

Consideremos un modelo de Cox ajustado con las covariables edad, sexo y tratamiento. Para diagnosticar la influencia de observaciones individuales, calculamos los residuos de deviance y martingala para cada observación.

Observación	Edad	Sexo	Tratamiento	Residuo de Deviance
1	50	0	1	1.2
2	60	1	0	-0.5
3	45	0	1	-1.8
4	70	1	0	0.3

Cuadro 11.1: Residuos de deviance para observaciones individuales

Observaciones con residuos de deviance grandes en valor absoluto (como la observación 3) pueden ser influyentes y requieren una revisión adicional.

El diagnóstico y la validación son pasos críticos en el anślisis de modelos de Cox. Evaluar el supuesto de proporcionalidad de riesgos, la bondad de ajuste y la influencia de observaciones individuales asegura que las inferencias y conclusiones derivadas del modelo sean vślidas y fiables.

Capítulo 12

Modelos Acelerados de Fallos

Los modelos acelerados de fallos (AFT) son una alternativa a los modelos de riesgos proporcionales de Cox. En lugar de asumir que las covariables afectan la tasa de riesgo, los modelos AFT asumen que las covariables multiplican el tiempo de supervivencia por una constante.

Definición del Modelo AFT

Un modelo AFT se expresa como:

$$T = T_0 \exp(\beta^T X)$$

donde:

- T es el tiempo de supervivencia observado.
- T_0 es el tiempo de supervivencia bajo condiciones basales.
- β es el vector de coeficientes del modelo.
- X es el vector de covariables.

Transformación Logarítmica

El modelo AFT se puede transformar logarítmicamente para obtener una forma lineal:

$$\log(T) = \log(T_0) + \beta^T X$$

Estimación de los Parámetros

Los parámetros del modelo AFT se estiman utilizando el método de máxima verosimilitud. La función de verosimilitud se define como:

$$L(\beta) = \prod_{i=1}^{n} f(t_i \mid X_i; \beta)$$

donde $f(t_i \mid X_i; \beta)$ es la función de densidad de probabilidad del tiempo de supervivencia t_i dado el vector de covariables X_i y los parámetros β.

Función de Log-Verosimilitud

La función de log-verosimilitud es:

$$\log L(\beta) = \sum_{i=1}^{n} \log f(t_i \mid X_i; \beta)$$

Maximización de la Verosimilitud

Los estimadores de máxima verosimilitud se obtienen resolviendo el sistema de ecuaciones obtenido al igualar a cero las derivadas parciales de $\log L(\beta)$ con respecto a β:

$$\frac{\partial \log L(\beta)}{\partial \beta} = 0$$

Distribuciones Comunes en Modelos AFT

En los modelos AFT, el tiempo de supervivencia T puede seguir varias distribuciones comunes, como la exponencial, Weibull, log-normal y log-logística. Cada una de estas distribuciones tiene diferentes propiedades y aplicaciones.

Modelo Exponencial AFT

En un modelo exponencial AFT, el tiempo de supervivencia T sigue una distribución exponencial con parámetro λ:

$$f(t) = \lambda \exp(-\lambda t)$$

La función de supervivencia es:

$$S(t) = \exp(-\lambda t)$$

La transformación logarítmica del tiempo de supervivencia es:

$$\log(T) = \log\left(\frac{1}{\lambda}\right) + \beta^T X$$

Modelo Weibull AFT

En un modelo Weibull AFT, el tiempo de supervivencia T sigue una distribución Weibull con parámetros λ y k:

$$f(t) = \lambda k t^{k-1} \exp(-\lambda t^k)$$

La función de supervivencia es:

$$S(t) = \exp(-\lambda t^k)$$

La transformación logarítmica del tiempo de supervivencia es:

$$\log(T) = \log\left(\left(\frac{1}{\lambda}\right)^{1/k}\right) + \frac{\beta^T X}{k}$$

Interpretación de los Coeficientes

En los modelos AFT, los coeficientes β_i se interpretan como factores multiplicativos del tiempo de supervivencia. Un valor positivo de β_i indica que un aumento en la covariable X_i incrementa el tiempo de supervivencia, mientras que un valor negativo indica una reducción del tiempo de supervivencia.

Ejemplo de Aplicación del Modelo AFT

Consideremos un ejemplo con tres covariables: edad, sexo y tratamiento. Supongamos que los datos se ajustan a un modelo Weibull AFT y obtenemos los siguientes coeficientes:

$$\hat{\beta}_{edad} = -0{,}02, \quad \hat{\beta}_{sexo} = 0{,}5, \quad \hat{\beta}_{tratamiento} = -1{,}2$$

La función de supervivencia ajustada se expresa como:

$$S(t \mid X) = \exp\left(-\left(\frac{t\exp(-0{,}02 \cdot \text{edad} + 0{,}5 \cdot \text{sexo} - 1{,}2 \cdot \text{tratamiento})}{\lambda}\right)^{k}\right)$$

Los modelos AFT proporcionan una alternativa flexible a los modelos de riesgos proporcionales de Cox. Su enfoque en la multiplicación del tiempo de supervivencia por una constante permite una interpretación intuitiva y aplicaciones en diversas áreas.

Capítulo 13

Análisis Multivariado de Supervivencia

El análisis multivariado de supervivencia extiende los modelos de supervivencia para incluir múltiples covariables, permitiendo evaluar su efecto simultáneo sobre el tiempo hasta el evento. Los modelos de Cox y AFT son comúnmente utilizados en este contexto.

Modelo de Cox Multivariado

El modelo de Cox multivariado se define como:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta^T X)$$

donde X es un vector de covariables.

Estimación de los Parámetros

Los parámetros β se estiman utilizando el método de máxima verosimilitud parcial, como se discutió anteriormente. La función de verosimilitud parcial se maximiza para obtener los estimadores de los coeficientes.

Modelo AFT Multivariado

El modelo AFT multivariado se expresa como:

$$T = T_0 \exp(\beta^T X)$$

Estimación de los Parámetros

Los parámetros β se estiman utilizando el método de máxima verosimilitud, similar al caso univariado. La función de verosimilitud se maximiza para obtener los estimadores de los coeficientes.

Interacción y Efectos No Lineales

En el análisis multivariado, es importante considerar la posibilidad de interacciones entre covariables y efectos no lineales. Estos se pueden incluir en los modelos extendiendo las funciones de riesgo o supervivencia.

Interacciones

Las interacciones entre covariables se pueden modelar añadiendo términos de interacción en el modelo:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2)$$

donde $X_1 X_2$ es el término de interacción.

Efectos No Lineales

Los efectos no lineales se pueden modelar utilizando funciones no lineales de las covariables, como polinomios o splines:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta_1 X + \beta_2 X^2)$$

Selección de Variables

La selección de variables es crucial en el análisis multivariado para evitar el sobreajuste y mejorar la interpretabilidad del modelo. Métodos como la regresión hacia atrás, la regresión hacia adelante y la selección por criterios de información (AIC, BIC) son comúnmente utilizados.

Regresión Hacia Atrás

La regresión hacia atrás comienza con todas las covariables en el modelo y elimina iterativamente la covariable menos significativa hasta que todas las covariables restantes sean significativas.

Regresión Hacia Adelante

La regresión hacia adelante comienza con un modelo vacío y añade iterativamente la covariable más significativa hasta que no se pueda añadir ninguna covariable adicional significativa.

Criterios de Información

Los criterios de información, como el AIC (Akaike Information Criterion) y el BIC (Bayesian Information Criterion), se utilizan para seleccionar el modelo que mejor se ajusta a los datos con la menor complejidad posible:

$$\begin{aligned} AIC &= -2 \log L + 2k \\ BIC &= -2 \log L + k \log n \end{aligned}$$

donde L es la función de verosimilitud del modelo, k es el número de parámetros en el modelo y n es el tamaño de la muestra.

Ejemplo de Análisis Multivariado

Consideremos un ejemplo con tres covariables: edad, sexo y tratamiento. Ajustamos un modelo de Cox multivariado y obtenemos los siguientes coeficientes:

$$\hat{\beta}_{edad} = 0{,}03, \quad \hat{\beta}_{sexo} = -0{,}6, \quad \hat{\beta}_{tratamiento} = 1{,}5$$

La función de riesgo ajustada se expresa como:

$$\lambda(t \mid X) = \lambda_0(t) \exp(0{,}03 \cdot \text{edad} - 0{,}6 \cdot \text{sexo} + 1{,}5 \cdot \text{tratamiento})$$

El análisis multivariado de supervivencia permite evaluar el efecto conjunto de múltiples covariables sobre el tiempo hasta el evento. La inclusión de interacciones y efectos no lineales, junto con la selección adecuada de variables, mejora la precisión y la interpretabilidad de los modelos de supervivencia.

Capítulo 14

Supervivencia en Datos Complicados

El análisis de supervivencia en datos complicados se refiere a la evaluación de datos de supervivencia que presentan desafíos adicionales, como la censura por intervalo, datos truncados y datos con múltiples tipos de eventos. Estos escenarios requieren métodos avanzados para un análisis adecuado.

14.1. Censura por Intervalo

La censura por intervalo ocurre cuando el evento de interés se sabe que ocurrió dentro de un intervalo de tiempo, pero no se conoce el momento exacto. Esto es común en estudios donde las observaciones se realizan en puntos de tiempo discretos.

Modelo para Datos Censurados por Intervalo

Para datos censurados por intervalo, la función de verosimilitud se modifica para incluir la probabilidad de que el evento ocurra dentro de un intervalo:

$$L(\beta) = \prod_{i=1}^{n} P(T_i \in [L_i, U_i] \mid X_i; \beta)$$

donde $[L_i, U_i]$ es el intervalo de tiempo durante el cual se sabe que ocurrió el evento para el individuo i.

14.2. Datos Truncados

Los datos truncados ocurren cuando los tiempos de supervivencia están sujetos a un umbral, y solo se observan los individuos cuyos tiempos de supervivencia superan (o están por debajo de) ese umbral. Existen dos tipos principales de truncamiento: truncamiento a la izquierda y truncamiento a la derecha.

Modelo para Datos Truncados

Para datos truncados a la izquierda, la función de verosimilitud se ajusta para considerar solo los individuos que superan el umbral de truncamiento:

$$L(\beta) = \prod_{i=1}^{n} \frac{f(t_i \mid X_i; \beta)}{1 - F(L_i \mid X_i; \beta)}$$

donde L_i es el umbral de truncamiento para el individuo i.

14.3. Análisis de Competing Risks

En estudios donde pueden ocurrir múltiples tipos de eventos (competing risks), es crucial modelar adecuadamente el riesgo asociado con cada tipo de evento. La probabilidad de ocurrencia de cada evento compite con las probabilidades de ocurrencia de otros eventos.

Modelo de Competing Risks

Para un análisis de competing risks, la función de riesgo se descompone en funciones de riesgo específicas para cada tipo de evento:

$$\lambda(t) = \sum_{j=1}^{m} \lambda_j(t)$$

donde $\lambda_j(t)$ es la función de riesgo para el evento j.

14.4. Métodos de Imputación

Los métodos de imputación se utilizan para manejar datos faltantes o censurados en estudios de supervivencia. La imputación múltiple es un enfoque común que crea múltiples conjuntos de datos completos imputando valores faltantes varias veces y luego combina los resultados.

Imputación Múltiple

La imputación múltiple para datos de supervivencia se realiza en tres pasos:

1. Imputar los valores faltantes múltiples veces para crear varios conjuntos de datos completos.
2. Analizar cada conjunto de datos completo por separado utilizando métodos de supervivencia estándar.
3. Combinar los resultados de los análisis separados para obtener estimaciones y varianzas combinadas.

Ejemplo de Análisis con Datos Complicados

Consideremos un estudio con datos censurados por intervalo y competing risks. Ajustamos un modelo para los datos censurados por intervalo y obtenemos los siguientes coeficientes para las covariables edad y tratamiento:

$$\hat{\beta}_{edad} = 0{,}04, \quad \hat{\beta}_{tratamiento} = -0{,}8$$

La función de supervivencia ajustada se expresa como:

$$S(t \mid X) = \exp\left(-\left(\frac{t \exp(0{,}04 \cdot \text{edad} - 0{,}8 \cdot \text{tratamiento})}{\lambda}\right)^k\right)$$

El análisis de supervivencia en datos complicados requiere métodos avanzados para manejar censura por intervalo, datos truncados y competing risks. La aplicación de modelos adecuados y métodos de imputación asegura un análisis preciso y completo de estos datos complejos.