ESTADÍSTICA MATEMÁTICA

Iker Caballero Bragagnini

Tabla de contenido

La estadística matemática y la inferencia

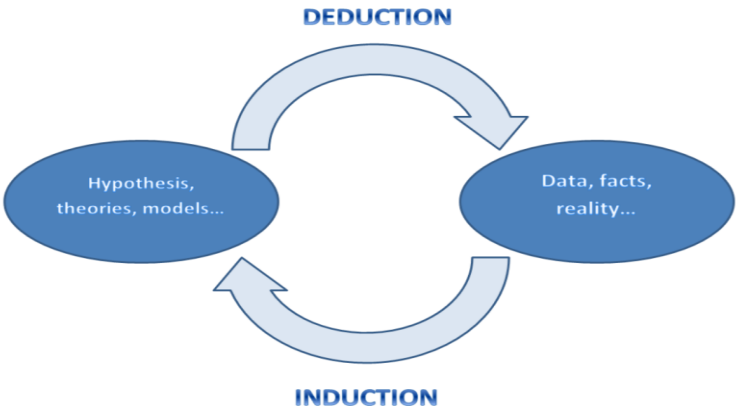
- La estadística es la ciencia que relaciona los datos a cuestiones de interés específicas, lo cual incluye construir métodos para recoger datos relevantes para las cuestiones, métodos para resumir y mostrar los datos, y métodos que permiten obtener respuestas apoyadas por los datos

    - La inferencia estadística permite obtener métodos y herramientas para poder obtener conclusiones generales de los datos aún habiendo incertidumbre en los datos

        - La incertidumbre puede nacer de la selección de elementos que se miden o puede nacer de la variabilidad en el proceso de medida

        - Los métodos usados para el análisis dependen de la manera en que los datos se han recogido, y es de vital importancia que haya un modelo de probabilidad explicando como la incertidumbre se introduce en los datos

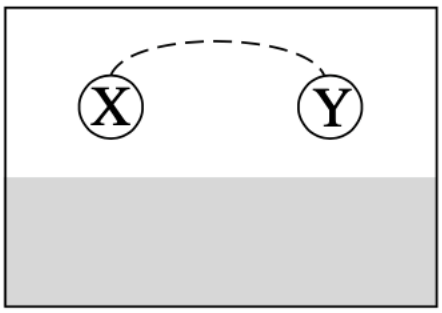
    - Los dos procesos básicos en la inferencia estadística son el proceso inductivo y el proceso deductivo



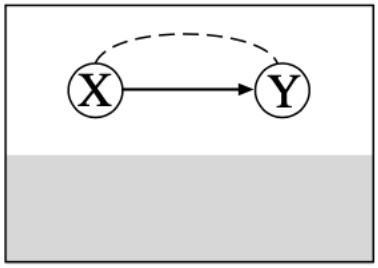
        - El proceso de deducción consiste en establece una hipótesis sobre el mecanismo que genera los datos y, de este, se deducen las probabilidades de los diferentes valores

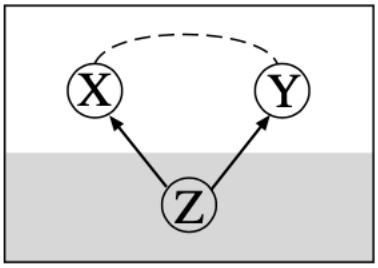
        - El proceso de inducción consiste en que, dado unos datos (o ciertos cálculos sobre ellos) obtenidos de una variable aleatoria, se intenta adivinar los parámetros del modelo estadístico generado por los datos

- Suponiendo que se observan dos variables Y y X, en donde la variable X parece tener asociación con la variable Y (ya sea positiva o negativa), se querría determinar por qué las dos variables están asociadas

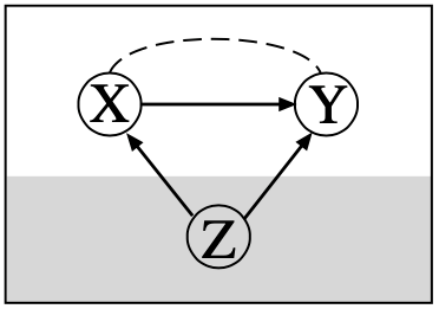


  - Hay varias explicaciones posibles por las cuales dos variables pueden estar asociadas. La relación podría ser causal, de modo que se observa la asociación porque una variable causa la otra, aunque también podría haber una tercera variable (latente) Z no identificada que tiene un efecto causal en estas variables





  - Por lo tanto, también es posible que tanto el efecto causal entre las dos variables y el efecto causal entre ellas y la variable latente puedan estar contribuyendo a esa asociación (los efectos se confunden). Esto hace que haya una necesidad de explicar la asociación a través de relaciones causales y medir el tamaño de las contribuciones de cada relación



- En la edad media, la ciencia se deducía de unos principios fijados hace muchos años por autoridades tales como Aristóteles, pero la idea de que las teorías científicas debían ser comprobadas con datos reales revoluciono el pensamiento científico, creando así el método científico

- El método científico se basa en las siguientes premisas:
    - Una hipótesis científica nunca puede mostrarse con absoluta certeza, pero esta debe ser potencialmente descartable
    - El modelo que se deriva de estas hipótesis es útil hasta que se establece que este no es verdad
    - Lo mejor es siempre ir por la hipótesis más simple, aunque se pueda demostrar su falsedad (la navaja de Ockham)
- Por lo tanto, el método científico dirige a uno a través de una secuencia de modelos que mejoran, dado que los anteriores demuestran ser falsos. El método científico generalmente sigue el procedimiento siguiente:
    - Preguntar o proponer un problema en términos de la hipótesis científica actual
    - Recoger toda la información relevante actualmente disponible (conocimiento actual de los parámetros del modelo)
    - Diseñar una investigación o experimento que se enfoque en la pregunta o problema. El resultado predicho del experimento debería ser una cosa si la hipótesis científica actual es correcta, y otra cosa si esta es falsa
    - Recoger datos del experimento y concluir a partir de los resultados experimentales. Además, se tiene que revisar el conocimiento sobre los parámetros para tener en cuenta los resultados encontrados
- El método científico busca por relaciones causa-efecto entre una variable experimental y una variable de resultado. En otras palabras, se busca como el cambio en una variable experimental hace cambiar la variable resultado
    - El modelaje científico desarrolla modelos matemáticos de estas relaciones. Ambos tipos de modelo necesitan aislar el experimento de factores externos que puedan afectar a los resultados experimentales
    - Todos los factores externos que pueden afectar a los resultados se tienen que controlar, por lo que no hay variables latentes que no se tengan en cuenta. Las otras variables se pueden identificar y controlarse físicamente para mantenerlas constantes, de modo,

en haciendo todo esto, se podría aislar el efecto de la variable experimental sobre la variable de resultado

- Los métodos estadísticos de inferencia se pueden usar cuando hay variabilidad aleatoria en los daros. El modelo de probabilidad para los datos se justifica por el diseño de la investigación o el experimento

    - Esto puede extender el método científico a situaciones en donde los factores externos relevantes no se puedan identificar, ya que en estos casos los factores no se pueden controlar y afectarán a los resultados experimentales

    - La idea estadística de aleatorización se ha desarrolado para lidiar con la posibilidad anterior: los factores no identificados se pueden eliminar al hacer la media si se asigna cada unidad a un grupo de tratamiento o control de manera aleatoria

    - Esto contribuye a la variabilidad de los datos, de modo que las conclusiones estadística siempre tendrán incertidumbre o error debido a esta variabilidad en los datos. Es posible desarrollar un modelo probabilístico de la variabilidad de los datos basado en la aleatorización

    - La aleatorización no solo reduce la incertidumbre debida a factores externos, sino que permite medir la cantidad de incertidumbre que queda usando el modelo de probabilidad. Por lo tanto, la aleatorización permite controlar los factores externos estadísticamente porque estos efectos se eliminan en la media

    - La idea subyacente a la aleatorización es la de población estadística, consistiendo de todos los valores posibles de las observaciones que se podrían dar. Los datos consisten de observaciones tomadas de una muestral de esta población, y las inferencias sobre los parámetros poblaciones se hacen a través de estadísticos muestrales, por lo que la muestra debe de ser representativa de la población

- Hay principalmente dos enfoques filosóficos en estadística: el enfoque frecuentista o clásico y el enfoque bayesiano

    - La mayoría de teoría estadística que se enseña utiliza el enfoque frecuentista para la estadística,

        - Este enfoque se basa en ideas como que los parámetros son constantes fijas desconocidas, que las probabilidades siempre se interpretan como la frecuencia relativa a largo plazo, y que los

procedimientos estadísticos se juzgan según sus rendimiento a largo plazo al utilizar un número infinito de repeticiones del experimento hipotéticas

- Las proposiciones de probabilidad solo se pueden hacer para cantidades aleatorias, por lo que no se pueden hacer para los parámetros. En cambio, se consigue una muestra de la población y se calcula un estadístico muestral, cuya distribución de probabilidad sobre todas las muestras aleatorias posibles se determina (la distribución muestral)

- Un parámetro de la población también tiene que ser un parámetro de la distribución muestral, por lo que la proposición de probabilidad que se podría hacer sobre el estadístico se convierte en una proposición de confianza sobre el parámetro, basado en el comportamiento medio del procedimiento sobre todas las muestras posibles

- La inferencia estadística frecuentista también se puede clasificar en paramétrica o no paramétricos

  - La inferencia estadística paramétrica se da cuando se supone una forma para la distribución de la variable que se estudia y el objetivo es estima los parámetros desconocidos de esta distribución

  - La inferencia estadística no paramétrica se da cuando se considera desconocida la distribución de la variable aleatoria y solo se supone sobre las propiedades generales (en vez de modificar la suposición sobre la distribución)

- Uno se suele centrar en la inferencia paramétrica, por lo que se asume que la medida de probabilidad de una variable aleatoria pertenece a un modelo matemático que depende de un número finito de parámetros

$$f_{\boldsymbol{y}} \in \{f(\boldsymbol{y}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^n\}$$

  - La función $f(\boldsymbol{y}|\boldsymbol{\theta})$ es una familia de distribuciones paramétricas y Θ es el espacio paramétrico

- Thomas Bayes demostró como las probabilidades inversas pueden usarse para calcular la probabilidad de eventos antecedentes a partir de la ocurrencia de un evento consecuente, y en el siglo XX se desarrolló un método de inferencia estadística completo basado en el teorema de Bayes

- Como hay incertidumbre sobre los valores reales de los parámetros, se considera que estos son variables aleatorias

- Las reglas de la probabilidad se usan directamente para hacer inferencias sobre los parámetros

- Las proposiciones de probabilidad sobre los parámetros se interpretan como un grado de creencia. La distribución *a priori* tiene que ser subjetiva, de modo que cada persona tiene su propia distribución que contiene los pesos relativos que se da a cada valor posible de los parámetros (la plausibilidad de que ocurra cada valor antes de que se observen según cada persona)

- Se revisan las creencias sobre los parámetros después de obtener los datos a través del teorema de Bayes. Esto permite obtener una distribución posterior, que da los pesos relativos que se dan a cada parámetro después de analizar los datos, por lo que la distribución posterior proviene de la *a priori* y de los datos

- Esto tiene un número de ventajas sobre el enfoque frecuentista convencional:

    - El teorema de Bayes es la única manera consistente de modificar las creencias sobre los parámetros dados los datos. Esto significa que la inferencia se basa en la ocurrencia real de los datos, no en todos los posibles conjuntos de datos que podrían ocurrir, pero no han ocurrido (como en el enfoque frecuentista)

    - Dejar que los parámetros sean aleatorios permite que se hagan proposiciones de probabilidad sobre estos, después de obtener los datos. Esto contrasta con lo convencional, que se basa en todos los conjuntos de datos que podrían haberse obtenidos dado un parámetro fijo (por lo que solo se pueden hacer proposiciones de confianza basado en lo que podría haber ocurrido)

    - La estadística bayesiana tiene una manera general de lidiar con parámetros molestos (a diferencia de la estadística frecuentista), que son aquellos para los cuales no se quiere inferenciar, pero interfieren con las inferencias hechas sobre los parámetros de interés

    - La estadística bayesiana es predictiva, a diferencia de la estadística convencional, de modo que se puede encontrar una distribución de probabilidad condicional para la siguiente observación de los datos muestrales

- Durante el desarrollo de la teoría sobre la inferencia estadística, se suele utilizar el término "condiciones de regularidad" para denotar suposiciones adecuadas para poder demostrar algunos resultados

    - Estas condiciones de regularidad son técnicas y normalmente se satisfacen en la mayoría de problemas razonables

        - Estas condiciones son lo suficientemente generales para poder aplicarlas en varios contextos, aunque no son las más generales

        - Estas condiciones de regularidad están principalmente relacionadas con la diferenciabilidad de la función de densidad y la capacidad de intercambiar derivación por integración

    - Las siguientes seis suposiciones permiten demostrar varios resultados asintóticos que se verán posteriormente:

        - Se observa $Y_1, Y_2, \dots, Y_n$, donde $Y_i \sim f(y|\theta)$ y son independientes e idénticamente distribuidas

        - El parámetro es identificable, de modo que si $\theta \neq \theta'$, entonces $f(y|\theta) \neq f(y|\theta')$

        - Las densidades $f(y|\theta)$ tienen un espacio muestral o soporte $\mathcal{Y}$ común, y $f(y|\theta)$ es diferenciable en θ

        - El espacio paramétrico Θ contiene un conjunto abierto ω del cual el parámetro θ_0 es un punto interior

        - Se observa $Y_1, Y_2, \dots, Y_n$, donde $Y_i \sim f(y|\theta)$ y son independientes e idénticamente distribuidas

        - El parámetro es identificable, de modo que si $\theta \neq \theta'$, entonces $f(y|\theta) \neq f(y|\theta')$

        - Para cualquier $y \in \mathcal{Y}$, la densidad $f(y|\theta)$ es tres veces diferenciables con respecto a θ, tal que la tercera derivada es continua en θ y $\int f(y|\theta)\, dy$ es tres veces diferenciable bajo el signo de la integral

        - Para cualquier $\theta_0 \in \Theta$ existe un número positivo c y una función $M(y)$ (ambos elementos pueden dependet de θ_0) tal que $E_{\theta_0}[M(Y)] < \infty$ y que se cumple lo siguiente:

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(y|\theta)\right| \leq M(y) \;\; for \;\; \forall y \in \mathcal{Y} \;\; \& \;\; |\theta - \theta_0| < c$$

El muestreo

- La ciencia estadística demuestra que los datos que se usan deben ser relevantes para problemas y preguntas particulares, pero que estos se deben recoger usando la aleatorización. El desarrollo de métodos para recoger datos usando la aleatorización es una de las mayores contribuciones científicas hechas por la estadística

    - La variabilidad en los datos debida al azar puede ser eliminada en la media al incrementar el número de elementos en la muestra, mientras que la variabilidad causada por otros factores no se puede eliminar

        - Los métodos estadísticos han sido desarrollados para recoger datos aleatorios pero relevantes, los cuales se dividen en dos campos: la teoría de la encuesta muestral y el diseño experimental

        - La teoría de la encuesta muestral es el estudio de métodos para el muestro de una población real finita, mientras que el diseño experimental es el estudio de métodos para diseñar experimentos que se enfocan en factores deseados y que no son afectados por otros factores no identificados

    - Las inferencias siempre dependen del modelo de probabilidad que se asume que genera estos datos observados

        - Cuando los datos no se recogen aleatoriamente, hay riesgo de que el patrón observado se deba a las variables latentes no observadas, en vez de ser una reflexión del patrón subyacente. En un diseño experimental, los tratamientos son asignados a sujetos de tal manera que se reducen los efectos de cualquier variable latente

        - Cuando se hacen inferencias de los datos recogidos acorde a un diseño experimental o a una encuesta aleatoria, el modelo de probabilidad para los datos observados se deriva del diseño del experimento o encuesta, y uno puede estar seguro de que es correcto

        - Cuando se hacen inferencias de los datos recogidos provienen de un diseño no aleatorio, no se tiene una justificación subyacente para el modelo de probabilidad, por lo que se tiene que asumir uno y este puede no concordar con la realidad

    - Los tres términos más importantes para la inferencia estadística son la población, la muestra y la inferencia estadística

- La población se define como el grupo entero de objetos o personas sobre el cual el investigador quiere informarse, y se puede considerar la población modelo como el conjunto de números para cada individuo de la población real. El interés es sobretodo informarse sobre los parámetros (números asociados con la distribución de la población)

- La muestra se define como un subconjunto de la población, de modo que el investigador coge una muestra de la población y obtiene información de los individuos de esa muestra. A partir de esta muestra, se pueden calcular los estadísticos muestrales (características numéricas que resumen la distribución muestral), que tiene la misma relación con la muestra que la que tienen los parámetros con la población

- La inferencia estadística consiste en hacer proposiciones sobre los parámetros poblacionales a partir de los estadísticos muestrales. La distribución muestral debe ser parecida a la distribución poblacional, de modo que se pueden hacer buenas inferencias, pero siempre debe evitarse el sesgo muestral (tendencia sistemática de recoger una muestra que no es representativa de la población)

- Aunque uno sea consciente de cosas de la población y se intente representar en una muestra, hay otros factores poblacionales que no se tienen en cuenta y que harían a la muestra no representativa

    - Sorprendentemente, las muestras aleatorias dan más muestras representativas que cualquier otro método no aleatorio. No solo minimizan la cuantía de error en la inferencia, sino que también permite hacer una cuantificación probabilística de los errores que permanecen

- Las variables aleatorias $Y_1, Y_2, \dots, Y_n$ son una muestra aleatoria de tamaño n de la población $f(y)$ si $Y_1, Y_2, \dots, Y_n$ son variables mutuamente independientes y su función de densidad o de masa de probabilidad es la misma función $f(y)$

    - Alternativamente, a las $Y_1, Y_2, \dots, Y_n$ se les llaman variables aleatorias independientes e idénticamente distribuidas, abreviado como i.i.d.

    - El modelo de muestreo aleatorio describe un tipo de situación experimental en la que la variables de interés tiene una distribución descrita por cc, de modo que las observaciones de las variables permiten calcular las probabilidades para Y a través

de $f(y)$. Además, las observaciones se toman de manera que no hay relación entre ellas (son mutuamente independientes)

- Debido a que cada observación es i.i.d., la función conjunta de densidad o de masa de probabilidad (también llamada función de verosimilitud) se da por la siguiente fórmula:

$$f(y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} f(y_i)$$

$$f(y_1, y_2, \dots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

- El modelo de muestra aleatoria definido asume que la población es infinita, dado que observar una variable primero no afecta la distribución de probabilidad de las otras variables (suposición de independencia) y la población de la que se extraen las observaciones es la misma (suposición de distribución idéntica)

  - Cuando el muestreo proviene de una población finita (un número finito de números $\{y_1, \dots, y_N\}$, la definición anterior puede no ser relevante dependiendo de la manera en la que se recogen las observaciones

  - Si se hace un muestreo equiprobable (todas las observaciones tienen la misma probabilidad de ser escogidas) en donde una observación se puede volver a escoger, se dice que se hace un muestreo aleatorio con reemplazo y las condiciones de la definición anterior se mantienen (cada variable aleatoria Y_i puede tomar cualquier valor $y_1, \dots, y_N$)

  - Si se hace un muestreo equiprobable pero sin reemplazo, entonces cada variable Y_i puede tomar todos los valores menos los previamente observados, lo cual hace que la elección de observaciones no sea mutuamente independiente y no se cumpla esta condición de la definición anterior (pero siguen proviniendo de la misma distribución). Este tipo de muestreo se denomina muestreo aleatorio simple

$$Y_1 = y_1 \Rightarrow P(Y_2 = y_1|Y_1 = y_1) = 0$$

$$\Rightarrow P(Y_2 = y_2|Y_1 = y_1) = \frac{1}{N-1}$$

  - Si N es muy grande comparado con el número de observaciones escogidas n, entonces $Y_1, Y_2, \dots, Y_n$ son casi independientes y se

puede asumir que son mutuamente independientes (en el sentido que la distribución condicional de Y_i dado $Y_1, \dots, Y_{i-1}$ no es muy diferente a la distribución marginal de Y_i)

Los estadísticos

- Suponiendo que los datos $Y_1, Y_2, \dots, Y_n$ se sacan de una población cualquiera, los datos observados $Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n$ son números y $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ es el vector de datos observados. Hay veces que los datos $Y_1, Y_2, \dots, Y_n$ son vectores aleatorios

    - Normalmente estos datos tienen una función de masa de probabilidad o una función de densidad de probabilidad conjunta $f(y_1, y_2, \dots, y_n|\boldsymbol{\theta})$, donde $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n)$ es el vector de parámetros desconocidos

        - En la función de masa o densidad de probabilidad conjunta, $y_1, y_2, \dots, y_n$ son variables, no los datos observados

        - Si estos provienen de una muestra aleatoria simple, entonces la función de masa o de densidad de probabilidad conjunta se puede expresar como el producto de las funciones marginales

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

    - Las funciones de variable real $T(y_1, y_2, \dots, y_n) = T(\boldsymbol{y})$ y el vector de este tipo de funciones $\boldsymbol{T}(\boldsymbol{y}) = (T_1(\boldsymbol{y}), T_2(\boldsymbol{y}), \dots, T_n(\boldsymbol{y}))$ son de especial interés en inferencia

        - Estas funciones son la base que permiten definir lo que es un estadístico

    - Un estadístico es una función de los datos que no depende de ningún parámetro desconocido, y la distribución de probabilidad del estadístico se denomina distribución muestral del estadístico

        - Siendo $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ los datos, en donde Y_i es una variable aleatoria, si $T(y_1, y_2, \dots, y_n)$ es una función de variable real cuyo dominio incluye el espacio muestral $\mathcal{Y}$ de $\boldsymbol{Y}$, entonces $W = T(Y_1, Y_2, \dots, Y_n)$ es un estadístico siempre que T no dependa de parámetros desconocidos

        - Los datos provienen de una distribución de probabilidad y el estadístico es una variable aleatoria porque depende de estos datos. Por lo tanto, el estadístico también proviene de una distribución de probabilidad, la cual es la distribución muestral

del estadístico (pero que no es la misma que la de los datos) y se pueden calcular momentos como con otras distribuciones

$$E[T(Y_1, \dots, Y_n)] = \cdots$$

$$Var[T(Y_1, \dots, Y_n)] = E\left[\left(T(Y_1, \dots, Y_n) - E\big(T(Y_1, \dots, Y_n)\big)\right)^2\right] = \cdots$$

…

- Si los datos observados son $Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n$, entonces el valor observado del estadístico es $w = T(y_1, y_2, \dots, y_n)$

- Todo aplica de manera análoga si el estadístico T es una función vectorial y los datos $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \dots, \boldsymbol{Y_n}$ son vectores aleatorios

- Normalmente $Y_1, Y_2, \dots, Y_n$ serán independientes e idénticamente distribuidas (i.i.d), de modo que los estadísticos de la forma $\sum_{i=1}^{n} a_i Y_i$ y $\sum_{i=1}^{n} t(Y_i)$, donde t es una función de variable real, será especialmente importantes

  - Los teoremas anteriormente vistos son útiles para poder encontrar las distribuciones muestrales de algunos de los estadísticos cuando Y_i son i.i.d de una distribución de probabilidad (normalmente una de la familia exponencial)

  - Algunos de los estadísticos más importantes de esta forma son la media muestra, la mediana muestral, la varianza muestral, el máximo y mínimo muestral, los estadísticos del orden, etc.

- Para la inferencia paramétrica, la función de masa de probabilidad o la función de densidad de probabilidad de una variable aleatoria Y es $f_{\boldsymbol{\theta}}(y)$, donde $\boldsymbol{\theta} \in \Theta$ es desconocido. Por lo tanto Y proviene de una familia de distribuciones indexadas por $\boldsymbol{\theta}$ y, como estas se especifican completamente por $\boldsymbol{\theta}$ (y sus cantidades), un objetivo importante de la inferencia paramétrica es encontrar buenos estimadores de $\boldsymbol{\theta}$. Para ello, se suele utilizar la noción de estadísticos suficientes

  - La idea básica de un estadístico suficiente $\boldsymbol{T(Y)}$ para $\boldsymbol{\theta}$ es que toda la información necesaria para la inferencia a partir de los datos $Y_1, Y_2, \dots, Y_n$ sobre el parámetro $\boldsymbol{\theta}$ esté contenida en el estadístico $\boldsymbol{T(Y)}$ (principio de suficiencia)

    - La importancia de un estadístico suficiente reside en la reducción de dimensiones. Como el estadístico suficiente $\boldsymbol{T(Y)}$ tiene toda la información a partir de los datos necesaria para inferenciar sobre $\boldsymbol{\theta}$ y este es suele tener unas dimensiones

menores a n, no es necesario entender una n-tupla n-dimensional de datos observados

- Si $(\boldsymbol{Y_1}, \boldsymbol{Y_2}, \dots, \boldsymbol{Y_n})$ tienen una distribución conjunta que depende de un vector de parámetros $\boldsymbol{\theta} \in \Theta$ donde Θ es el espacio de parámetros, entonces un estadístico $\boldsymbol{T}(\boldsymbol{Y_1}, \boldsymbol{Y_2}, \dots, \boldsymbol{Y_n})$ es un estadístico suficiente para $\boldsymbol{\theta}$ si la distribución condicional de $(\boldsymbol{Y_1}, \boldsymbol{Y_2}, \dots, \boldsymbol{Y_n})$ dado un valor del estadístico $\boldsymbol{T} = \boldsymbol{t}$ no depende de $\boldsymbol{\theta}$ para ningún valor de $\boldsymbol{t}$ en el espacio muestral de $\boldsymbol{T}$

  - Lo que esto quiere decir es que, como la distribución condicional de $(Y_1, Y_2, \dots, Y_n)$ dado $\boldsymbol{T} = \boldsymbol{t}$ no depende de $\boldsymbol{\theta}$, entonces $\boldsymbol{T}(Y_1, Y_2, \dots, Y_n)$ permite obtener toda la información de $(Y_1, Y_2, \dots, Y_n)$ sin necesitar $\boldsymbol{\theta}$ (el estadístico es suficiente) y no se necesita el parámetro para calcular las probabilidades para los valores de $\boldsymbol{Y}$

  - Matemáticamente, esto se traduce en que $P_{\boldsymbol{\theta}}\big(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big)$ no dependerá del parámetro $P_{\boldsymbol{\theta}}$, por lo que se cumple la siguiente igualdad:

$$P\big(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big) = P_{\boldsymbol{\theta}}\big(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big)$$

$$for\ \forall \boldsymbol{\theta} \in \Theta$$

  - Normalmente $\boldsymbol{T}$ e $\boldsymbol{Y_i}$ son funciones de variable real

- Siendo $f(\boldsymbol{y}|\boldsymbol{\theta})$ para $\boldsymbol{\theta} \in \Theta$ una familia de funciones de densidad de probabilidad o funciones de masa de probabilidad para $\boldsymbol{Y}$ y g y h funciones no negativas, y asumiendo que $f(\boldsymbol{y}|\boldsymbol{\theta})$ cumple una condición de regularidad, un estadístico $\boldsymbol{T}(\boldsymbol{y})$ es suficiente para $\boldsymbol{\theta}$ si, y solo si, para todos los puntos muestrales $\boldsymbol{y}$ y para todo $\boldsymbol{\theta} \in \Theta$ se cumple la siguiente igualdad:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})h(\boldsymbol{y})$$

  - A este teorema se le denomina criterio de factorización de Neyman-Fisher

  - La condición de regularidad mencionada expresa que existe un conjunto $\{\boldsymbol{y_i}\}_{i=1}^{\infty}$ que no depende de $\boldsymbol{\theta} \in \Theta$ tal que $\sum_{i=1}^{\infty} f(\boldsymbol{y_i}|\boldsymbol{\theta}) = 1$ para toda $\boldsymbol{\theta} \in \Theta$ para la familia de funciones de masa. Esta se suele satisfacer, por ejemplo, cuando el espacio muestral $\mathcal{Y}$ no tiene ningún $\boldsymbol{\theta}$ o si $\boldsymbol{y} = (y_1, \dots, y_n)$ y y_i toman valores como $y_i \in \{1,2, \dots, \theta\}$ para $\theta \in \{1,2,3, \dots\}$

- Se utiliza $\boldsymbol{T}(\boldsymbol{y})$ y $\boldsymbol{\theta}$ porque el criterio de factorización se puede aplicar para múltiples parámetros que actúen como variables y múltiples estadísticos que los intenten estimar

- En este caso, la función h no depende de $\boldsymbol{\theta}$ y la función g solo depende de $\boldsymbol{y}$ a través de $\boldsymbol{T}(\boldsymbol{y})$. Por lo tanto, también se puede decir que un estadístico $\boldsymbol{T}(\boldsymbol{y})$ es suficiente si, y solo si, el cociente $f(\boldsymbol{y}|\boldsymbol{\theta})/g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$). Esto ocurre porque al separar $f(\boldsymbol{y}|\boldsymbol{\theta})$ en una parte $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$ y otra $h(\boldsymbol{y})$, dividir $f(\boldsymbol{y}|\boldsymbol{\theta})$ entre $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$ eliminará esa parte que depende del parámetro (dado que serán equivalentes si la igualdad se cumple) y hará que la razón solo dependa de las observaciones $\boldsymbol{y}$

$$\frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})} = h(\boldsymbol{y})$$

- Debido a que $\{\boldsymbol{Y} = \boldsymbol{y}\} \subseteq \{\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\}$, entonces $P(\boldsymbol{Y} = \boldsymbol{y}) = P\big(\boldsymbol{Y} = \boldsymbol{y} \cap \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big)$, por lo que se pueden obtener las siguientes equivalencias:

$$\{\boldsymbol{Y} = \boldsymbol{y}\} \subseteq \{\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\}$$

$$\Rightarrow P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}\big(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big)$$

$$= P_{\boldsymbol{\theta}}\big(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big)P\big(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big)$$

$$P_{\boldsymbol{\theta}}\big(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big) \equiv g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta}) \;\&\; P\big(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big) \equiv h(\boldsymbol{y})$$

$$\Rightarrow g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})h(\boldsymbol{y}) \;\; for \; \forall \boldsymbol{\theta} \in \Theta$$

- Finalmente, a partir de esas equivalencias, se puede obtener una expresión para g y h a través de la definición de función de masa de probabilidad y de la probabilidad condicional (también se puede demostrar en el caso continuo):

$$P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t}) = \sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} f(\boldsymbol{y}|\boldsymbol{\theta}) = g(\boldsymbol{t}|\boldsymbol{\theta}) \sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} h(\boldsymbol{y})$$

$$P_{\boldsymbol{\theta}}\big(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big) = \frac{P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t})}{P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t})} =$$

$$= \frac{P_{\boldsymbol{\theta}}\big(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big)P\big(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\big)}{P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t})} = \frac{P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})}{P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t})} =$$

$$= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{g(\boldsymbol{t}|\boldsymbol{\theta})\sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} h(\boldsymbol{y})} = \frac{g(\boldsymbol{t}|\boldsymbol{\theta})h(\boldsymbol{y})}{g(\boldsymbol{t}|\boldsymbol{\theta})\sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} h(\boldsymbol{y})} =$$

$$= \frac{h(\boldsymbol{y})}{\sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} h(\boldsymbol{y})} \Rightarrow \boldsymbol{T}(\boldsymbol{y}) \; is \; suff.$$

- A los teoremas anteriores se le pueden añadir varias observaciones útiles:

  - Se puede considerar que la función que depende de los valores muestrales $h(\boldsymbol{y})$ es constante (normalmente 1)

  - Suponiendo que $Y_1, Y_2, \dots, Y_n$ son i.i.d. de una distribución con espacio muestral $\mathcal{Y}^* \equiv \mathcal{Y}_i$ y función de densidad de probabilidad o función de masa de probabilidad $f(y|\boldsymbol{\theta}) = k(y|\boldsymbol{\theta})\boldsymbol{I}(y \in \mathcal{Y}^*)$, entonces $f(y|\boldsymbol{\theta}) = \prod_{i=1}^{n} k(y_i|\boldsymbol{\theta}) \prod_{i=1}^{n} \boldsymbol{I}(y_i \in \mathcal{Y}^*)$. Pero, como el espacio muestral de $\mathcal{Y}$ es $\mathcal{Y} = \mathcal{Y}^* \times \dots \times \mathcal{Y}^*$ y eso hace que $\boldsymbol{I}(\boldsymbol{y} \in \mathcal{Y}) = \prod_{i=1}^{n} \boldsymbol{I}(y_i \in \mathcal{Y}^*) = \boldsymbol{I}(all \; y_i \in \mathcal{Y}^*)$, entonces se puede ver que $f(y|\boldsymbol{\theta}) = \prod_{i=1}^{n} k(y_i|\boldsymbol{\theta})\, \boldsymbol{I}(all \; y_i \in \mathcal{Y}^*)$

$$f(y|\boldsymbol{\theta}) = \prod_{i=1}^{n} k(y_i|\boldsymbol{\theta}) \prod_{i=1}^{n} \boldsymbol{I}(y_i \in \mathcal{Y}^*) = \prod_{i=1}^{n} k(y_i|\boldsymbol{\theta})\, \boldsymbol{I}(all \; y_i \in \mathcal{Y}^*)$$

  - Si $\mathcal{Y}^*$ no depende de $\boldsymbol{\theta}$, entonces $\boldsymbol{I}(all \; y_i \in \mathcal{Y}^*)$ es parte de $h(\boldsymbol{y})$, pero si depende de $\boldsymbol{\theta}$, entonces $\boldsymbol{I}(all \; y_i \in \mathcal{Y}^*)$ se puede poner en $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$. Normalmente $\mathcal{Y}^*$ es un intervalo con extremos a y b (no necesariamente finito). Por lo tanto, para funciones de densidad de probabilidad se da la siguiente igualdad, la cual permite dividir $\prod_{i=1}^{n} \boldsymbol{I}(y_i \in [a, b])$ y dar una intuición de en dónde poner cada función:

$$\prod_{i=1}^{n} \boldsymbol{I}(y_i \in [a, b]) = \boldsymbol{I}\big(a \le y_{(1)} < y_{(n)} \le b\big) =$$

$$= \boldsymbol{I}\big[a \le y_{(1)}\big]\boldsymbol{I}\big[y_{(n)} \le b\big]$$

If θ_1 or θ_2 is known, then the above factorization works, but it is better to make the dimension of the sufficient statistic as small as possible. If θ_1 is known, then

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{(\theta_2 - \theta_1)^n} I(y_{(n)} \le \theta_2) I(\theta_1 \le y_{(1)})$$

where the first two terms are $g(T(\boldsymbol{y})|\theta_2)$ and the third term is $h(\boldsymbol{y})$. Hence $T(\boldsymbol{Y}) = Y_{(n)}$ is a sufficient statistic for θ_2 by factorization. If θ_2 is known, then

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 \le y_{(1)}) I(y_{(n)} \le \theta_2)$$

- En los casos especiales en el que los intervalos sean $(-\infty, b)$ o $[a, \infty)$ (ejemplos para intervalos abiertos y semiabiertos), se dan las siguientes identidades:

$$\prod_{i=1}^{n} \boldsymbol{I}\big(y_i \in (-\infty, b)\big) = \boldsymbol{I}\big[y_{(n)} \leq b\big]$$

$$\prod_{i=1}^{n} \boldsymbol{I}\big(y_i \in [a, \infty)\big) = \boldsymbol{I}\big[a \leq y_{(1)}\big]$$

- En el caso en el que $y_i \in A_j$ donde A_j es una partición, una igualdad obvia pero útil es la siguiente:

$$\prod_{j=1}^{k} \boldsymbol{I}\big(\boldsymbol{y} \in A_j\big) = \boldsymbol{I}\left(\boldsymbol{y} \in \bigcap_{j=1}^{k} A_j\right)$$

- Suponiendo que $Y_1, Y_2, \ldots, Y_n$ tiene una distribución conjunta que depende de un vector de parámetros $\boldsymbol{\theta}$ para $\boldsymbol{\theta} \in \Theta$, un estadístico suficiente $\boldsymbol{T}(\boldsymbol{Y})$ para $\boldsymbol{\theta}$ es un estadístico suficiente mínimo para $\boldsymbol{\theta}$ si $\boldsymbol{T}(\boldsymbol{Y})$ es una función de $\boldsymbol{S}(\boldsymbol{Y})$ para cualquier otro estadístico suficiente $\boldsymbol{S}(\boldsymbol{Y})$ para $\boldsymbol{\theta}$

  - Por lo tanto, un estadístico suficiente mínimo se puede entender como una función $\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{g_S}\big(\boldsymbol{S}(\boldsymbol{Y})\big)$ para una función $\boldsymbol{g_S}$ en donde $\boldsymbol{S}(\boldsymbol{Y})$ es un estadístico suficiente

  - Si $\boldsymbol{S}(\boldsymbol{Y})$ no es un estadístico suficiente mínimo, entonces $\boldsymbol{S}(\boldsymbol{Y})$ no es una función del estadístico suficiente mínimo $\boldsymbol{T}(\boldsymbol{Y})$ (un estadístico que no es suficiente mínimo no puede ser función de un estadístico suficiente mínimo). Si $\boldsymbol{S}(\boldsymbol{Y})$ no es mínimo, existe un estadístico suficiente $\boldsymbol{W}(\boldsymbol{Y})$ tal que $\boldsymbol{S}(\boldsymbol{Y})$ no es una función de $\boldsymbol{W}(\boldsymbol{Y})$, por lo que suponiendo que $\boldsymbol{S}(\boldsymbol{Y}) = \boldsymbol{h}[\boldsymbol{T}(\boldsymbol{Y})]$ para alguna función $\boldsymbol{h}$, $\boldsymbol{S}(\boldsymbol{Y}) = \boldsymbol{h}\big[\boldsymbol{g_W}\big(\boldsymbol{W}(\boldsymbol{Y})\big)\big]$ es una función de $\boldsymbol{W}(\boldsymbol{Y})$ y se llega a una contradicción, por lo que no puede darse que $\boldsymbol{S}(\boldsymbol{Y}) = \boldsymbol{h}[\boldsymbol{T}(\boldsymbol{Y})]$

  - Si $\boldsymbol{T_1}$ y $\boldsymbol{T_2}$ son estadísticos suficientes mínimos, entonces $\boldsymbol{T_1} = \boldsymbol{g}(\boldsymbol{T_2})$ y $\boldsymbol{T_2} = \boldsymbol{h}(\boldsymbol{T_1})$. Eso quiere decir que $\boldsymbol{g}\big(\boldsymbol{h}(\boldsymbol{T_1})\big) = \boldsymbol{T_1}$ y $\boldsymbol{h}\big(\boldsymbol{g}(\boldsymbol{T_2})\big) = \boldsymbol{T_2}$, por lo que $\boldsymbol{h}$ y $\boldsymbol{g}$ son funciones inversas que son biyectivas y $\boldsymbol{T_1}$ y $\boldsymbol{T_2}$ son estadísticos equivalentes

  - Si el estadístico suficiente mínimo $\boldsymbol{T}(\boldsymbol{Y})$ es igual a $\boldsymbol{g_S}\big(\boldsymbol{S}(\boldsymbol{Y})\big)$ donde $\boldsymbol{g_S}$ no es biyectiva, entonces $\boldsymbol{T}(\boldsymbol{Y})$ proporciona una mayor reducción de los datos que el estadístico suficiente $\boldsymbol{S}(\boldsymbol{Y})$.

Por lo tanto, los estadísticos suficientes mínimos proporcionan la mayor reducción posible de datos

- Suponiendo que $\boldsymbol{T}(\boldsymbol{Y})$ tiene una función de masa de probabilidad o una función de densidad de probabilidad $f(\boldsymbol{t}|\boldsymbol{\theta})$, entonces $\boldsymbol{T}(\boldsymbol{Y})$ es un estadístico suficiente completo para $\boldsymbol{\theta}$ si $E_{\boldsymbol{\theta}}[\boldsymbol{g}(\boldsymbol{T}(\boldsymbol{Y}))] = 0$ para toda $\boldsymbol{\theta}$ implica que $P_{\boldsymbol{\theta}}[\boldsymbol{g}(\boldsymbol{T}(\boldsymbol{Y})) = 0] = 1$ para toda $\boldsymbol{\theta}$. La función g no puede depender de ningún parámetro desconocido

  - El estadístico $\boldsymbol{T}(\boldsymbol{Y})$ tiene una distribución muestral que depende de n en $\boldsymbol{\theta} \in \Theta$, por lo que la propiedad de ser un estadístico suficiente completo depende de la familia de distribuciones con función de masa o densidad de probabilidad $f(\boldsymbol{t}|\boldsymbol{\theta})$

  - El criterio usado para demostrar la complitud de un estadístico pone una restricción más fuerte en g, y cuanto más grande sea la familia de distribuciones, más grande es la restricción en g

  - Las familias exponenciales regulares tienen un estadístico suficiente completo $\boldsymbol{T}(\boldsymbol{Y})$ (para cada uno de los parámetros en $\boldsymbol{\theta}$ a estimar)

- Además del principio de suficiencia, otros principios que se pueden utilizar para los estadísticos son el principio de verosimilitud y el principio de equivarianza

  - El principio de verosimilitud expresa que si $\boldsymbol{x}$ e $\boldsymbol{y}$ son dos muestras de la misma población tales que la función de verosimilitud $L(\boldsymbol{\theta}|\boldsymbol{x})$ es proporcional a $L(\boldsymbol{\theta}|\boldsymbol{y})$, entonces las conclusiones extraídas de las muestras $\boldsymbol{x}$ e $\boldsymbol{y}$ deben ser idénticas

    - Siendo $f(\boldsymbol{y}|\boldsymbol{\theta})$ la función de masa de probabilidad o de densidad de probabilidad de una muestra $\boldsymbol{Y}$ con un espacio paramétrico Θ , si $\boldsymbol{Y} = \boldsymbol{y}$ se observa, entonces la función de verosimilitud es $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$

    - De manera no formal, el principio expresa que, dado un modelo estadístico, toda la evidencia relevante para la modelización de los parámetros de una muestra está contenida en la función de verosimilitud

    - De manera, formal si $\boldsymbol{x}$ e $\boldsymbol{y}$ son dos muestras de la misma población, entonces existe una constante $c(\boldsymbol{x}, \boldsymbol{y})$ (la constante depende de las muestras) tal que se cumpla la siguiente igualdad:

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = c(\boldsymbol{x}, \boldsymbol{y})L(\boldsymbol{\theta}|\boldsymbol{y})$$

- Debido a que la función de verosimilitud se usa para comparar la plausibilidad de varios valores de los parámetros, como $c_{\boldsymbol{x},\boldsymbol{y}}$ depende de $\boldsymbol{x}$ e $\boldsymbol{y}$, entonces da igual que muestra se escoja, $c(\boldsymbol{x},\boldsymbol{y})$ no varía y la relación entre la verosimilitud de dos vectores de parámetros diferentes debe de ser la misma y, por tanto, la plausibilidad relativa de un vector se mantiene

$$L(\boldsymbol{\theta_1}|\boldsymbol{x}) = c_{\boldsymbol{x},\boldsymbol{y}} L(\boldsymbol{\theta_2}|\boldsymbol{x})$$

$$L(\boldsymbol{\theta_1}|\boldsymbol{y}) = c_{\boldsymbol{x},\boldsymbol{y}} L(\boldsymbol{\theta_2}|\boldsymbol{y})$$

- Se utiliza el término plausibilidad porque si se usara el término de probabilidad, se tendría que garantizar que $L(\boldsymbol{\theta}|\boldsymbol{y})$ es una función de densidad y eso no siempre se puede garantizar

Los estadísticos para muestras aleatorias

- Como se ha mencionado antes, los estadísticos más importantes son la media muestral, la varianza muestral y la cuasi-varianza muestral, dado que estos proporcionan un buen resumen de la información de la muestra

    - Las definiciones de la media, la varianza y la cuasi-varianza muestral son las siguientes:

        - La media muestral es la media aritmética de los valores en la muestra aleatoria, y se define de la siguiente manera:

$$\bar{Y} \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i$$

        - La varianza muestral es el estadístico definido de la siguiente manera:

$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

        - La desviación estándar muestral o error estándar no es más que la raíz cuadrada de la varianza muestral

        - La cuasi-varianza muestral es el estadístico definido de la siguiente manera:

$$\hat{S}^2 \equiv \frac{n-1}{n} S^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Siendo $y_1, y_2, \dots, y_n$ números cualesquiera y $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, entonces se cumplen las siguientes dos proposiciones:

  - El valor a que minimiza $\sum_{i=1}^{n}(y_i - a)^2$ es $a = \bar{y}$. Esto se puede demostrar a través de expandir la expresión y sumando y restando $\bar{y}$

$$\sum_{i=1}^{n}(y_i - a)^2 = \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - a)^2 =$$

$$= \sum_{i=1}^{n}\big((y_i - \bar{y}) + (\bar{y} - a)\big)^2 =$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - a)\sum_{i=1}^{n}(y_i - \bar{y}) + n(\bar{y} - a)^2 =$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - a)^2$$

  - El estadístico $(n-1)S^2$ es equivalente a $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$. Esto se puede demostrar a través de expandir la expresión

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\left(ny_i - \sum_{i=1}^{n} y_i\right)^2 =$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} y_i^2 - 2n\left(\sum_{i=1}^{n} y_i\right)^2 + \sum_{i=1}^{n}\left(\sum_{i=1}^{n} y_i\right)^2\right] =$$

$$= \frac{1}{n^2}\left[n^2\sum_{i=1}^{n} y_i^2 - 2n\left(\sum_{i=1}^{n} y_i\right)^2 + n\left(\sum_{i=1}^{n} y_i\right)^2\right] =$$

$$= \frac{1}{n^2}\left[n^2\sum_{i=1}^{n} y_i^2 - n\left(\sum_{i=1}^{n} y_i\right)^2\right] = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} =$$

$$= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

- Siendo $Y_1, Y_2, \dots, Y_n$ una muestra aleatoria de una población y $g(y)$ una función tal que $E[g(Y_1)]$ y $Var[g(Y_1)]$ existen, entonces se cumplen las siguientes identidades:

$$E\left[\sum_{i=1}^{n} g(Y_i)\right] = nE[g(Y_1)] \qquad Var\left[\sum_{i=1}^{n} g(Y_i)\right] = nVar[g(Y_1)]$$

  - Estas identidades se pueden demostrar a través de las propiedades lineales de la esperanza, de la varianza de la suma de variables y de la independencia de las variables aleatorias:

$$E\left[\sum_{i=1}^{n} g(Y_i)\right] = \sum_{i=1}^{n} E[g(Y_i)] = nE[g(Y_1)]$$

$$Var\left[\sum_{i=1}^{n} g(Y_i)\right] = \sum_{i=1}^{n} Var[g(Y_i)] + 2\sum_{i<j}^{n} Cov\left[g(Y_i), g\left(Y_j\right)\right]$$

$$= \sum_{i=1}^{n} Var[g(Y_i)] = \sum_{i=1}^{n} Var[g(Y_1)] = nVar[g(Y_1)]$$

  - No obstante, si hay restas dentro de $g(Y_i)$, estas se convierten en sumas gracias a la propiedad de la varianza $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$. Por lo tanto, se puede modelar este hecho a través de una segunda función $h(Y_i)$:

$$Var\left[\sum_{i=1}^{n} g(Y_i) - \sum_{i=1}^{n} h(Y_i)\right] =$$

$$= \sum_{i=1}^{n} Var[g(Y_i)] + \sum_{i=1}^{n} Var[h(Y_i)] - 2\sum_{i<j}^{n} Cov\left[g(Y_i), g\left(Y_j\right)\right] =$$

$$= \sum_{i=1}^{n} Var[g(Y_i)] + \sum_{i=1}^{n} Var[h(Y_i)]$$

- Siendo $Y_1, Y_2, \dots, Y_n$ una muestra aleatoria de una población con media μ y varianza $\sigma^2 < \infty$, entonces se cumplen las siguientes identidades:

$$\begin{cases} E(\bar{Y}) = \mu \\ Var(\bar{Y}) = \dfrac{\sigma^2}{n} \end{cases} \qquad \begin{cases} E(S^2) = \sigma^2 \\ Var(S^2)^* = \dfrac{2\sigma^4}{n-1} \end{cases} \qquad \begin{cases} E(S^2) = \left(\dfrac{n-1}{n}\right)\sigma^2 \\ Var(S^2)^* = \dfrac{2\sigma^4}{n} \end{cases}$$

$$^* solo\ aplica\ si\ Y_i \sim N(\mu, \sigma^2)$$

- Las identidades para la media muestral se pueden demostrar utilizando las propiedades de la esperanza y la varianza

$$E(\bar{Y}) = E\left(\frac{\sum_{i=1}^n Y_i}{n}\right) = \frac{1}{n}\sum_{i=1}^n E(Y_i) = \frac{n\mu}{n} = \mu$$

$$Var(\bar{Y}) = Var\left(\frac{\sum_{i=1}^n Y_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n Var(Y_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- Las identidades para la varianza muestral se pueden demostrar utilizando las propiedades de la esperanza y la varianza y a partir de que $(n-1)S^2/\sigma^2$ sigue una distribución χ^2_{n-1} y su varianza es $2(n-1)$ (pero esto último solo aplica si las variables se distribuyen de manera normal)

$$E(S^2) = E\left[\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}\right] = \frac{1}{n-1}\sum_{i=1}^n E[(Y_i - \bar{Y})^2] =$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^n (Y_i - \mu)^2 - n(\bar{Y} - \mu)^2\right] = \frac{\sum_{i=1}^n \sigma^2 - nE[(\bar{Y} - \mu)^2]}{n-1}$$

$$= \frac{n\sigma^2 - \sigma^2}{n-1} = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

$$Var(S^2) = Var\left[\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}\right] = Var\left[\frac{\sigma^2}{n-1}\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2}\right]$$

$$= \frac{\sigma^4}{(n-1)^2} Var\left[\sum_{i=1}^n \left(\frac{Y_i - \bar{Y}}{\sigma^2}\right)^2\right] = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{(n-1)}$$

- Las identidades para la cuasi-varianza muestral se pueden demostrar utilizando las propiedades de la varianza muestral (pero esto último solo aplica si las variables se distribuyen de manera normal)

$$E(\hat{S}^2) = \frac{n-1}{n} E(S^2) = \left(\frac{n-1}{n}\right)\sigma^2$$

$$Var(\hat{S}^2) = \left(\frac{n-1}{n}\right)^2 Var(S^2) = \frac{2\sigma^4(n-1)}{n^2}$$

- También existen estadísticos que permiten medir la asociación lineal que hay entre dos variables, tales como la covarianza muestral y el coeficiente de correlación muestral de Pearson

    - Considerando n pares de medidas para dos variables p y m (porque la el individuo i se observa para ambas variables), la covarianza muestral mide la asociación lineal entre las medidas de las variables

$$S_{pm} = \frac{1}{n}\sum_{i=1}^{n}(Y_{ip} - \bar{Y}_p)(Y_{im} - \bar{Y}_m) \quad for \quad p, m = 1,2, \dots, k$$

        - La covarianza muestral es la media del producto entre las desviaciones de las variables de sus respectivas medias. Si valores grandes de una variable se observan conjuntamente con otros grandes valores para la otra variable, será positiva, pero si valores grandes de una ocurren con valores pequeños de otra, la covarianza será negativa. Si no hay relación lineal particular entre ambas, esta será aproximadamente cero

        - Cuando $p = m$, la covarianza muestral se reduce a la varianza muestral. Además, $S_{pm} = S_{mp}$ para toda p y m

    - El último estadístico descriptivo a considerar es el coeficiente de correlación lineal muestral de Pearson. Esta mide la asociación lineal entre dos variables sin depender de las unidades de medida, definida de la siguiente manera:

$$r_{pm} = \frac{S_{pm}}{\sqrt{S_{pp}}\sqrt{S_{mm}}} = \frac{\sum_{i=1}^{n}(Y_{ip} - \bar{Y}_p)(Y_{im} - \bar{Y}_m)}{\sqrt{\sum_{i=1}^{n}(Y_{ip} - \bar{Y}_p)^2}\sqrt{\sum_{i=1}^{n}(Y_{im} - \bar{Y}_m)^2}}$$

$$for \quad p, m = 1,2, \dots, k$$

        - Igual que con la covarianza, $r_{pm} = r_{mp}$ para toda p y m. Además, el valor de r_{pm} no se ve afectado por el uso de n o $n-1$ en el denominador de la varianza muestral

        - La correlación muestral es la versión estandarizada de la covarianza muestral, donde el producto de las desviaciones estándar muestrales permite hacer esta estandarización

- El coeficiente de correlación muestral también se puede ver como una covarianza muestral para valores estandarizados, los cuales son conmensurables porque están centrados en cero y expresados en unidades de la desviación estándar muestral

- Aunque los signos de la correlación y la covarianza muestral sean los mismos, la correlación es más fácil de interpretar porque su magnitud está acotada y cumple con las siguientes propiedades:

  - El valor de r debe estar acotado entre los valores -1 y +1 (ambos incluidos)

  - El valor de r mide la fuerza de la asociación lineal. Si $r = 0$, entonces hay una falta de asociación lineal entre componentes, y de otro modo, se indicaría la dirección de la asociación: $r < 0$ implica una tendencia de un valor en el par a ser mayor que su media cuando el otro es más pequeño que su media, mientras que $r > 0$ implica que ambos valores son mayores o menores a su media conjuntamente

  - El valor de r_{pm} permanece igual si las medidas de la variable p se cambian a $X_{pi} = aY_{pi} + b$ para $i = 1,2, \dots, n$ y los valores de la variable m se cambian a $X_{mi} = cY_{mi} + d$ para $i = 1,2, \dots, n$, siempre que a y c tengan el mismo signo

- Las cantidades S_{pm} y r_{pm} no proporcionan todo lo que se debe saber sobre la asociación entre dos variables, ya que asociaciones no lineales pueden existir y no revelarse en estos estadísticos porque proporcionan solo medidas de asociación lineal

  - Además, estas cantidades son muy sensibles a observaciones anormales o extremas, por lo que pueden indicar una asociación cuando, en verdad, existe poca

  - Sin embargo, la covarianza y el coeficiente de correlación lineal se calculan y se analizan regularmente. Estos estadísticos proporcionan resúmenes numéricos convincentes de asociación cuando los datos no muestran patrones de asociación no lineales obvios y cuando no hay presentes observaciones extremas

  - Las observaciones sospechosas deben tenerse en cuenta corrigiendo errores de medición obvios y tomando acciones consistentes con las causas identificadas de los errores. Los valores de S_{pm} y r_{pm} deberían presentarse con y sin estas observaciones

- Cuando las cantidades muestrales se obtienen de una población distribuida normalmente, entonces se obtienen propiedades útiles para los estadísticos muestrales y de varias distribuciones muestrales famosas

    - Siendo $Y_1, Y_2, \dots, Y_n$ una muestra aleatoria de una población con función generadora de momentos $M_Y(t)$, entonces la función generadora de momentos de la media muestral es la siguiente:

$$M_{\bar{Y}}(t) = \left[M_Y\left(\frac{t}{n}\right)\right]^n$$

    - Siendo $Y_1, Y_2, Y_3, \dots, Y_n$ i.i.d. de una distribución normal $N(\mu, \sigma^2)$, se cumplen las siguientes propiedades:

        - La media muestral $\overline{Y}$ y la varianza muestral S^2 son independientes (también lo es la cuasi-varianza muestral), lo cual se puede ver a través de expresar las variables como restas $Y_i - \overline{Y}$ para $i = 2,3, \dots, n$ (fijando $Y_i = Y_1$), de modo que como se puede hacer que haya diferentes observaciones en cada uno de los estadísticos. Este es un resultado del Teorema de Fisher

        - La media muestral $\overline{Y}$ sigue una distribución $N(\mu, \sigma^2/n)$. Esto se puede demostrar a través de la función generadora de momentos para la media muestral

$$M_{\bar{Y}}(t) = \left[\exp\left(\mu\frac{t}{n} + \frac{\sigma^2\left(\frac{t}{n}\right)^2}{2}\right)\right]^n = \exp\left(n\left(\mu\frac{t}{n} + \frac{\sigma^2\left(\frac{t}{n}\right)^2}{2}\right)\right) =$$

$$= \exp\left(\mu t + \frac{\frac{\sigma^2}{n}t^2}{2}\right) \Rightarrow \overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

        - El estadístico $(n-1)S^2/\sigma^2$ (que es igual a $n\hat{S}^2/\sigma^2$) sigue una distribución $\mathcal{X}^2_{n-1}$, ya que el estadístico $(Y_i - \mu)/\sigma \sim N(0,1)$ y eso permite obtener estadísticos con una distribución $\mathcal{X}^2_1$ y $\mathcal{X}^2_n$ que, a través de una descomposición de $\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$, permiten obtener la función generadora de momentos de $(n-1)S^2/\sigma^2$. Esto es uno de los resultados del Teorema de Fisher

$$\frac{(Y_i - \mu)}{\sigma} \sim N(0,1) \Rightarrow \frac{(Y_i - \mu)^2}{\sigma^2} \sim \mathcal{X}^2_1 \Rightarrow \frac{\sum_{i=1}^n (Y_i - \mu)^2}{\sigma^2} \sim \mathcal{X}^2_n$$

$$\sum_{i=1}^{n}(Y_i-\mu)^2=\sum_{i=1}^{n}\left(Y_i-\overline{Y}+\overline{Y}-\mu\right)^2=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\left(\overline{Y}-\mu\right)^2$$

$$=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2+n\left(\overline{Y}-\mu\right)^2$$

$$\Rightarrow\ W=\frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{\sigma^2}=\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{\sigma^2}+\left(\frac{\overline{Y}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2=U+V$$

$$W\sim\mathcal{X}_n^2\ \&\ V\sim\mathcal{X}_1^2\ \ are\ independent\ \Rightarrow\ m_W(t)=m_U(t)m_V(t)$$

$$\Rightarrow\ M_U(t)=\frac{M_W(t)}{M_V(t)}=\frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}}=(1-2t)^{-(n-1)/2}$$

$$\Rightarrow\ \frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{\sigma^2}\sim\mathcal{X}_{n-1}^2$$

- En muchos, la varianza de las variables aleatorias es desconocida y se utiliza la varianza muestral, lo cual permite obtener unas distribuciones llamadas t-Student y F de Fisher-Snedecor. Estos son dos resultados del Teorema de Fisher

  - Siendo $Y_1, Y_2, \dots, Y_n$ una muestra aleatoria de una población distribuida $N(\mu, \sigma^2)$, entonces la variable aleatoria $T=\frac{\bar{Y}-\mu}{S/\sqrt{n}}$ tiene una distribución t-Student con $n-1$ grados de libertad. Equivalentemente, una variable aleatoria T tiene una distribución t-Student con p grados de libertad si tiene la siguiente función de densidad de probabilidad:

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}}\sim t_{n-1}$$

$$f_T(t)=\frac{\Gamma\left(\frac{p-1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)}\frac{1}{(p\pi)^{\frac{1}{2}}}\frac{1}{\left(1+\frac{t^2}{p}\right)^{\frac{p+1}{2}}}\quad for\ -\infty<t<\infty$$

  - Siendo $X_1, X_2, \dots, X_n$ una muestra aleatoria de una población distribuida $N(\mu_X, \sigma_X^2)$ y $Y_1, Y_2, \dots, Y_m$ una muestra aleatoria de una población distribuida $N(\mu_Y, \sigma_Y^2)$, las cuales son independientes entre sí, entonces el siguiente estadístico tiene una distribución t-Student con $n+m-2$ grados de libertad:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\frac{(n-1)S_X^2}{\sigma_X^2} + \frac{(m-1)S_Y^2}{\sigma_Y^2}} \sqrt{\frac{n+m-2}{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim t_{n+m-2}$$

- Siendo $X_1, X_2, \dots, X_n$ una muestra aleatoria de una población distribuida $N(\mu_X, \sigma_X^2)$ y $Y_1, Y_2, \dots, Y_m$ una muestra aleatoria de una población distribuida $N(\mu_Y, \sigma_Y^2)$, las cuales son independientes entre sí, entonces la variable aleatoria $F = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ tiene una distribución F de Fisher-Snedecor con $n-1$ y $m-1$ grados de libertad. Equivalentemente, una variable aleatoria F tiene una distribución F de Fisher-Snedecor con p y q grados de libertad si tiene la siguiente función de densidad de probabilidad:

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}$$

$$f_T(t) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2} \frac{x^{(p/2)-1}}{(1+(p/q)x)^{(p+q)/2}} \quad for\ 0 < x < \infty$$

- Si $Y \sim F_{p,q}$, entonces $1/X \sim F_{q,p}$, y si $X \sim t_q$, entonces $X^2 \sim F_{1,q}$

- Para poder formar variables que sigan estas dos distribuciones, es necesario que $\bar{X}$ y S_X^2 sean independientes entre sí, dado que se garantiza que $\bar{X}$ y S_X^2 son independientes por el teorema de Fisher y esta es la razón de que se den ambas distribuciones (pero se tiene que garantizar esta independencia en la forma característica)

- Cuando los grados de libertad tienden a infinito, estas distribuciones se aproximan a la distribución normal, por lo que se simetrizan y se pueden inferir probabilidades de que se den valores con lo que se llama su "distribución aproximada" (la cual será normal)

- Valores muestrales como la observación más pequeña, la observación más grande o la que está en medio pueden proporcionar información adicional de una muestral aleatoria. Estos estadísticos se denominan estadísticos de orden, y se pueden tratar también como variables aleatorias

  - Los estadísticos de orden de una muestra aleatoria $Y_1, Y_2, \dots, Y_n$ son los valores muestrales posicionados en orden ascendente, denotados por $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$

    - Estos son variables aleatorias que satisfacen $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$

$$Y_{(1)} = \min_{1 \le i \le n} Y_i$$

$$Y_{(2)} = second\ smallest\ Y_i$$

...

$$Y_{(n)} = \max_{1 \le i \le n} Y_i$$

- Como son variables aleatorias, se pueden discutir las probabilidades de que tomen diversos valores. Estas probabilidades se calculan a partir de las funciones de densidad o de masa de probabilidad de estos estadísticos de orden

- Antes de estudiar las funciones de distribución y de densidad, es útil definir dos estadísticos que se pueden definir en términos de los estadísticos de orden

  - El rango muestral, denotado por $R = Y_{(n)} - Y_{(1)}$, es la distancia entre la observación más grande y la más pequeña

  - La mediana muestral, denotada por M, es un número tal que aproximadamente la mitad de las observaciones son menores a M y la otra mitad es mayor. En términos de estadísticos de orden, esta se puede definir de la siguiente manera:

$$M = \begin{cases} Y_{\left(\frac{n+1}{2}\right)} & if\ n\ is\ odd \\ \left(Y_{\left(\frac{n}{2}\right)} + Y_{\left(\frac{n}{2}+1\right)}\right)/2 & if\ n\ is\ even \end{cases}$$

  - La mediana es una medida de localización que se puede considerar alternativa al valor esperado, dado que la primera no se ve tan afectada por valore extremos como la última. La mediana, en consecuencia, da una indicación mejor de los valores típicos de la distribución

  - Alternativamente, la mediana se interpreta como el valor M tal que el $P(Y \le M) = 0.5$ y $P(Y > M) = 0.5$ (divide la distribución en dos partes con un 50% de probabilidad en cada lado)

- Para cualquier número p entre 0 y 1, el percentil $100p$ es la observación tal que aproximadamente np de las observaciones sean menores a esta observación y $n(1-p)$ de las observaciones sean mayores. Definiendo el percentil en términos de estadísticos de orden, se puede usar la siguiente expresión:

$$y^{perc} = \begin{cases} Y_{(np)} & if\ \frac{1}{2n} < p < \frac{1}{2} \\ Y_{(n+1-\{n(1-p)\})} & if\ \frac{1}{2} < p < 1 - \frac{1}{2n} \end{cases}$$

$$\{n(1-p)\} = nearest\ integer\ near\ n(1-p)$$

- De manera más formal, la notación $\{b\}$ se define como el número b redondeado al entero más cercano (de la manera usual). Si i es un entero y $i - 0.5 \leq b < i + 0.5$, entonces $\{b\} = i$

- La mediana se puede definir, por tanto, como el percentil 50 (con $p = 0.5$). Otros percentiles importantes son el cuartil inferior (con $p = 0.25$) y el cuartil superior (con $p = 0.75$)

- Por lo tanto, el percentil y^{perc} es aquel valor tal que el $P(Y \leq y^{perc}) = p$ y $P(Y > y^{perc}) = 1 - p$ (divide la distribución en dos partes que tienen que tener una probabilidad en función de p)

- Otra medida de dispersión muy utilizada se basa en la diferencia entre el cuartil superior y el inferior, llamada rango intercuartílico

- Siendo $Y_1, Y_2, \dots, Y_n$ una muestra aleatoria de una distribución discreta con función de masa $f_Y(y_i) = p_i$, donde $y_1 < y_2 < \cdots$ son posibles valores de Y en orden ascendente, $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ estadísticos de orden de la muestra, y definiendo $P_i \equiv P(Y \leq y_i) \equiv \sum_{i=1}^{n} p_i$, entonces se obtienen los siguientes resultados:

$$P\left(Y_{(j)} \leq y_i\right) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

$$P\left(Y_{(j)} = y_i\right) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}\right]$$

  - Utilizando una variable Z que cuente el número de $Y_1, Y_2, \dots, Y_n$ que son menores o iguales a y_i, y se puede interpretar que para cada $Y_1, Y_2, \dots, Y_n$, el evento $\{Y_j \leq y_i\}$ es un éxito y $\{Y_j > y_i\}$ un fracaso (con una probabilidad de éxito igual a P_i para cada $Y_1, Y_2, \dots, Y_n$. Como la muestra es aleatoria y simple, las variables son independientes e idénticamente, Z sigue una distribución binomial, y como el evento $\{Y_{(j)} \leq y_i\}$ es igual al evento $\{Z \geq j\}$, se puede obtener la siguiente igualdad:

$$P\left(Y_{(j)} \leq y_i\right) = P(Z \geq j)$$

- La función de masa de probabilidad para $Y_{(j)}$ nace de la diferencia entre $P\left(Y_{(j)} \leq y_i\right)$ y $P\left(Y_{(j)} \leq y_{i-1}\right)$

$$P\left(Y_{(j)} \leq y_i\right) - P\left(Y_{(j)} \leq y_{i-1}\right) =$$

$$= \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k} - \sum_{k=j}^{n} \binom{n}{k} P_{i-1}^k (1 - P_{i-1})^{n-k} =$$

$$= \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}\right]$$

- Siendo $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ los estadísticos de orden de una muestra aleatoria $Y_1, Y_2, \dots, Y_n$ de una población continua con función de distribución $F_Y(y)$ y función de densidad de probabilidad $f_Y(y)$, la función de densidad de $Y_{(j)}$ es la siguiente:

$$f_{Y_{(j)}} = \frac{n!}{(j-1)!\,(n-j)!} f_Y(y)[F_Y(y)]^{j-1}[1 - F_Y(y)]^{n-j}$$

  - Este resultado se puede demostrar a través de derivar la función de distribución encontrada anteriormente

- Siendo $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ los estadísticos de orden de una muestra aleatoria $Y_1, Y_2, \dots, Y_n$ de una población continua con función de distribución $F_Y(y)$ y función de densidad de probabilidad $f_Y(y)$, la función de densidad conjunta de $Y_{(i)}$ y $Y_{(j)}$ para $1 \leq i < j \leq n$es la siguiente:

$$f_{Y_{(i)},Y_{(j)}}(u, v) =$$

$$= \frac{n!}{(i-1)!\,(j-1-i)!\,(n-j)!} f_Y(u) f_Y(v)[F_Y(u)]^{j-1}[F_Y(v) - F_Y(u)]^{j-1-i}[1 - F_Y(u)]^{n-j}$$

$$for \;\; -\infty < u < v < \infty$$

  - La función de densidad conjunta de tres o más estadísticos de orden puede derivarse utilizando argumentos similares, pero más enrevesados

  - Una función de densidad conjunta importante es la de todos los estadísticos de orden, la cual es la siguiente:

$$f_{Y_{(1)},\dots,Y_{(n)}}(y_1, \dots, y_n) =$$

$$= \begin{cases} n!\, f_Y(y_1) \dots f_Y(y_n) & if\ -\infty < y_1 < \cdots < y_n < \infty \\ 0 & otherwise \end{cases}$$

- La $n!$ aparece en la formula porque para cualquier conjunto de valores $y_1, y_2, \dots, y_n$, existen $n!$ asignaciones igualmente probables de valores a $Y_1, Y_2, \dots, Y_n$ que permiten obtener el mismo valor para los estadísticos de orden

La estimación puntual: métodos, sesgo, información y FCRLB

- Un estimador puntual da un único valor como estimación de un parámetro, mientras que un estimador interválico da un rango (L_n, U_n) de valores razonables para un parámetro (como los intervalos de confianza). Uno de los estimadores puntuales más usados son los estimadores de máxima verosimilitud

  - Siendo $f(\boldsymbol{y}|\boldsymbol{\theta})$ la función de masa de probabilidad o de densidad de probabilidad de una muestra $\boldsymbol{Y}$ con un espacio paramétrico Θ , si $\boldsymbol{Y} = \boldsymbol{y}$ se observa, entonces la función de verosimilitud es $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$

    - Para cualquier punto muestral $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$, siendo $\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) \in \Theta$ un valor del parámetro en el que $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y})$ llega a un máximo como una función de $\boldsymbol{\theta}$ con $\boldsymbol{y}$ siendo fija, se dice que el estimador de máxima verosimilitud (MLE) del parámetro $\boldsymbol{\theta}$ basado en la muestra $\boldsymbol{Y}$ es $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$

    - Es crucial ver que la función de verosimilitud es una función de $\boldsymbol{\theta}$, por lo que $y_1, y_2, \dots, y_n$ son constantes, y que la función de densidad o de masa de probabilidad $f(\boldsymbol{y}|\boldsymbol{\theta})$ es una función de n variables mientras que la función $L(\boldsymbol{\theta})$ es una función de k variables (si $\boldsymbol{\theta}$ es un vector $1 \times k$)

    - Si $Y_1, Y_2, \dots, Y_n$ es una muestra independiente de la población con función de densidad o de masa de probabilidad $f(\boldsymbol{y}|\boldsymbol{\theta})$ (la muestra es independiente e idénticamente distribuida), entonces la función de verosimilitud es la siguiente:

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

    - Si $Y_1, Y_2, \dots, Y_n$ es una muestra independiente de la población con función de densidad o de masa de probabilidad diferente, entonces la función de verosimilitud es la siguiente:

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})$$

- Si el estimador de máxima verosimilitud $\widehat{\boldsymbol{\theta}}$ existe, entonces $\widehat{\boldsymbol{\theta}} \in \Theta$. Por lo tanto, si $\widehat{\boldsymbol{\theta}} \notin \Theta$, entonces $\widehat{\boldsymbol{\theta}}$ no es un estimador de máxima verosimilitud de $\boldsymbol{\theta}$

- Si el estimador de máxima verosimilitud existe, este será función de un estadístico suficiente (o será él mismo un estadístico suficiente). Además, el estimador de máxima verosimilitud es único, entonces es una función del estadístico suficiente mínimo

- Siendo $Y_1, Y_2, \dots, Y_n$ son los datos y suponiendo que el parámetro $\boldsymbol{\theta}$ tiene componentes $(\theta_1, \theta_2, \dots, \theta_k)$, $\hat{\theta}_i$ será el estimador de máxima verosimilitud de θ_i

  - Sin falta de generalidad, si se asume que $\boldsymbol{\theta} = (\theta_1, \theta_2)$, que el estimador de máxima verosimilitud es $\widehat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2\right)$ y que $\hat{\theta}_2$ es conocido, entonces la función de verosimilitud de perfil es $L(\theta_1) = L\left(\theta_1, \hat{\theta}_2(\boldsymbol{y})\right)$ con dominio $\left\{\theta_1 : \left(\theta_1, \hat{\theta}_2\right) \in \Theta\right\}$

  - Como $L(\theta_1, \theta_2)$ se maximiza sobre Θ en $\left(\hat{\theta}_1, \hat{\theta}_2\right)$, el maximizado de la función de verosimilitud de perfil y de su transformación logarítmica es $\hat{\theta}_1$

- Uno de los teoremas más útiles a la hora de obtener el estimador de máxima verosimilitud es el teorema del principio de invariabilidad

  - Si $\widehat{\boldsymbol{\theta}}$ es el estimador de máxima verosimilitud de $\boldsymbol{\theta}$, entonces $h\left(\widehat{\boldsymbol{\theta}}\right)$ es el estimador de máxima verosimilitud de $h(\boldsymbol{\theta})$ donde h es una función con dominio Θ

- Existen cuatro técnicas usadas comúnmente para encontrar el estimador de máxima verosimilitud:

  - Los candidatos potenciales se pueden encontrar diferenciando $\log L(\boldsymbol{\theta})$, llamada la función de verosimilitud logarítmica o *log-likelihood function*

  - Los candidatos potenciales se pueden encontrar diferenciando $L(\boldsymbol{\theta})$ directamente

  - Hay veces que se puede encontrar a través de la maximización directa de la verosimilitud $L(\boldsymbol{\theta})$

- A través del principio de invariabilidad, si $\widehat{\boldsymbol{\theta}}$ es el estimador de máxima verosimilitud de $\boldsymbol{\theta}$, entonces $h(\widehat{\boldsymbol{\theta}})$ es el estimador de máxima verosimilitud de $h(\boldsymbol{\theta})$

- Para poder obtener el estimador de máxima verosimilitud, primero se tiene que obtener la función de verosimilitud, de modo que utilizar propiedades del productorio y de la sumatoria que simplifiquen estos resultados es útil

  - En el caso en que $Y_1, Y_2, \dots, Y_n$ es una muestra independiente de la población con función de densidad o de masa de probabilidad $f(\boldsymbol{y}|\boldsymbol{\theta})$, entonces la función de verosimilitud se puede obtener como el productorio de esta para cada i y se puede simplificar (el caso más común)

  - En el caso en el que se involucren exponenciales, se puede aplicar la siguiente propiedad para eliminar el productorio de las exponenciales:

$$\prod_{i=1}^{n} e^{y_i} = e^{\sum_{i=1}^{n} y_i}$$

  - Cuando hay otros términos (constantes o parámetros normalmente) que no dependen de x_i, entonces estos se pueden extraer del productorio multiplicándolos las n veces que indica el productorio (propiedad del productorio)

$$\prod_{i=1}^{n} c y_i = c^n \prod_{i=1}^{n} y_i$$

  - En el caso en el que haya un factorial de las observaciones y_i, el productorio no se puede simplificar (dado que el factorial será diferente para cada y_i)

$$\prod_{i=1}^{n} \frac{c}{y_i!} = c^n \frac{1}{\prod_{i=1}^{n} y_i!}$$

  - En el caso en el que haya un producto de las observaciones y_i por una función de las observaciones, el productorio se puede simplificar dividiendo el productorio

$$\prod_{i=1}^{n} y_i \, f(y_i) = \prod_{i=1}^{n} y_i \prod_{i=1}^{n} f(y_i)$$

  - Si hay funciones con muchos términos exponenciales o productorios, lo más fácil es hacer una transformación

logarítmica y así trabajar con la función logarítmica de máxima verosimilitud

- Existen casos para los cuales no es sencillo obtener la función de máxima verosimilitud debido a que los valores de y_i están censurados o truncados, porque los valores de θ están acotados, o porque no se puede obtener un estimador de máxima verosimilitud (como en el caso de la distribución uniforme, por ejemplo)

  - En el caso en el que los valores de y_i estén truncados (solo se observan valores en un intervalo concreto), entonces lo que se tiene que hacer es dividir la función de densidad o de masa de probabilidad entre la probabilidad y_i esté en ese intervalo, dado que de este modo se condiciona la función solo a esos valores y permite que la probabilidad sumada sea 1 (se escala)

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{f_i(y_i|\boldsymbol{\theta})}{P(y_i \in I)} \quad where\ I\ is\ an\ interval$$

  - En el caso en el que los valores de y_i estén censurados (los valores para un intervalo I del espacio muestral $\mathcal{Y}$ se fijan a un valor a concreto), entonces lo que se tiene que hacer es utilizar la función de densidad o de masa de probabilidad para los valores que estén en $\mathcal{Y} - I$ y la función de densidad o de masa de probabilidad para los valores que estén en I

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}) \prod_{i=1}^{n} P(y_i \in I) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}) \prod_{i=1}^{n} P(y_i = a)$$

  - En el caso en el que los parámetros estén acotados, se tiene que comprobar que los estimadores encontrados existen para todos los valores que pueden tomar los parámetros. Esto se puede hacer comprobando los extremos de la acotación y viendo si existe alguna singularidad dentro del estimador

  - En el caso en el que no se puede obtener un estimador por máxima verosimilitud, se tiene que ver cuales son los valores que hacen que se maximice la función de verosimilitud para el parámetro (se supone un estimador, normalmente de orden, de modo que se obtenga el mayor valor posible de $L(\theta|\boldsymbol{y})$)

$$\hat{\theta} = \arg\max L(\theta|\boldsymbol{y}) \Rightarrow \begin{cases} \hat{\theta} = Y_{(j)} \\ \hat{\theta} = g(Y_i) \ \ for\ some\ interval \\ \ldots \end{cases}$$

- Una vez obtenida la función de verosimilitud, es necesario obtener el gradiente de primer orden para obtener los candidatos a máximo y el gradiente de segundo orden evaluado en $\widehat{\boldsymbol{\theta}}$ para comprobar que es negativo (y confirmar que $\widehat{\boldsymbol{\theta}}$ es un máximo)

$$\nabla L(\boldsymbol{\theta}) = \mathbf{0} \qquad \boldsymbol{x}' \boldsymbol{H}_{L(\widehat{\boldsymbol{\theta}})} \boldsymbol{x} < 0 \quad for\ \forall \boldsymbol{x} \in \mathbb{R}^n$$

  - En el caso de varios parámetros, es necesario utilizar la matriz Hessiana, la cual contiene las segundas derivadas parciales y las derivadas parciales cruzadas

  - Si la función $L(\boldsymbol{\theta})$ es continua en un intervalo $[a, b]$, entonces tanto el máximo como el mínimo de L existen, y si además es diferenciable en (a, b), entonces se pueden encontrar los puntos críticos y evaluar $L(\boldsymbol{\theta})$ en los puntos a, b y en los críticos para comprobar el máximo y el mínimo

  - Si $L(\boldsymbol{\theta})$ es estrictamente cóncava para todo $\forall \boldsymbol{\theta} \in \Theta$, cualquier máximo local de L es un máximo global, ya que un máximo local se obtiene si la función es cóncava ($\boldsymbol{x}' \boldsymbol{H}_{L(\widehat{\boldsymbol{\theta}})} \boldsymbol{x} \leq 0$)

$$\boldsymbol{x}' \boldsymbol{H}_{L(\widehat{\boldsymbol{\theta}})} \boldsymbol{x} < 0 \quad for\ \ \forall \boldsymbol{x} \in \mathbb{R}^n - \{\mathbf{0}\}\ \ \&\ \ \forall \boldsymbol{\theta} \in \Theta$$

  - Para comprobar que la Hessiana es una definida negativa, se puede utilizar el criterio de Sylvester y comprobar si las submatrices principales de orden par son positivas y si las de orden impar son negativas

$$\left| \frac{\partial^2}{\partial \theta_1^2} L(\widehat{\boldsymbol{\theta}}) \right| < 0 \qquad \left| \begin{matrix} \frac{\partial^2}{\partial \theta_1^2} L(\widehat{\boldsymbol{\theta}}) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} L(\widehat{\boldsymbol{\theta}}) \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} L(\widehat{\boldsymbol{\theta}}) & \frac{\partial^2}{\partial \theta_2^2} L(\widehat{\boldsymbol{\theta}}) \end{matrix} \right| > 0$$

$$1st\ submatrix \qquad\qquad 2nd\ submatrix$$

  - Si la función $L(\boldsymbol{\theta})$ es continua en un intervalo $[a, b]$ y diferenciable en el intervalo (a, b) y el punto crítico es único, entonces el punto crítico es el estimador de máxima verosimilitud si es un máximo local (dado que también sería un máximo global)

- El método de momentos es otro enfoque útil para obtener estimadores puntuales, el cual se basa en resolver un sistema de ecuaciones que iguala los momentos poblacionales con los momentos muestrales

- Siendo $Y_1, Y_2, \dots, Y_n$ una muestra aleatoria i.i.d. y definiendo $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i^j$ y $\mu_j \equiv \mu_j(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(Y^j)$ (de modo que $\hat{\mu}_j$ es el momento muestral j y μ_j es el momento poblacional), se puede fijar un número k y asumir que $\mu_j = \mu_j(\theta_1, \theta_2, \dots \theta_k)$ para resolver el siguiente sistema para $\widetilde{\boldsymbol{\theta}}$:

$$\begin{cases} \hat{\mu}_1 = \mu_1(\theta_1, \theta_2, \dots \theta_k) \\ \dots \\ \hat{\mu}_k = \mu_k(\theta_1, \theta_2, \dots \theta_k) \end{cases} \Rightarrow \begin{cases} \frac{1}{n}\sum_{i=1}^{n} Y_i^1 = E_{\boldsymbol{\theta}}(Y^1) \\ \dots \\ \frac{1}{n}\sum_{i=1}^{n} Y_i^k = E_{\boldsymbol{\theta}}(Y^k) \end{cases}$$

  - La solución $\widetilde{\boldsymbol{\theta}} = \left(\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k\right)$ (a veces es denotada por $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}$) es el estimador del método de momentos de $\boldsymbol{\theta}$. Si g es una función continua de los primeros k momentos y $h(\boldsymbol{\theta}) = g\left(\mu_1(\boldsymbol{\theta}), \mu_2(\boldsymbol{\theta}), \dots, \mu_k(\boldsymbol{\theta})\right)$, entonces el estimador del método de momentos de $h(\boldsymbol{\theta})$ es $g(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_k)$

  - Esta última definición es similar al principio de invariabilidad, pero la función g debe ser continua y debe, además, ser una función de los momentos muestrales para $k \geq 1$

  - No siempre existe en una forma cerrada de los estimadores del método de momentos y se tendrían que utilizar métodos numéricos. Además, puede ser que exista más de un estimador del método de momentos

  - Si hay distribuciones en las que un mismo parámetro puede corresponder a más de un momento poblacional, entonces se escoge el estimador del método de momentos de menor orden

- Siendo $Y_1, Y_2, \dots, Y_n$ una muestra i.i.d. de una distribución con una función de densidad o de masa $f(y|\boldsymbol{\theta})$, se cumplen las siguientes proposiciones:

  - Si $E(Y) = h(\theta)$, entonces $\hat{\theta}_{MM} = h^{-1}(\theta)$ **.** Esta es una consecuencia lógica del método explicado, dado que el estimado serán los valores de los parámetros tal que $\mu_j(\theta) = \hat{\mu}_j(\theta)$

  - El estimador del método de momentos de $E(Y) = \mu_1$ es $\hat{\mu}_1 = \bar{Y}$. Esto se puede demostrar a partir de la definición dada de $\hat{\mu}_1$

  - El estimador del método de momentos de la varianza $Var_\theta(Y) = \mu_2(\theta) - [\mu_1(\theta)]^2$ es la siguiente expresión:

$$\hat{\sigma}^2_{MM} = \hat{\mu}_2(\theta) - [\hat{\mu}_1(\theta)]^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \equiv S_M^2$$

- Siendo $S_M^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ y suponiendo que $E(Y) = h_1(\theta_1, \theta_2)$ y $Var(Y) = h_2(\theta_1, \theta_2)$, entonces se puede obtener un estimado del método de momentos $\widetilde{\boldsymbol{\theta}}$

$$\begin{cases} \bar{Y} = h_1(\theta_1, \theta_2) \\ S_M^2 = h_2(\theta_1, \theta_2) \end{cases}$$

  - Esto se puede demostrar a través de la definición de la varianza y de las definiciones anteriores:

$$\mu_1 = E(Y) = h_1(\theta_1, \theta_2) = \mu_1(\theta_1, \theta_2)$$

$$\Rightarrow \; Var(Y) = \mu_2 - \mu_1^2 = h_2(\theta_1, \theta_2)$$

$$\Rightarrow \; \mu_2 = h_2(\theta_1, \theta_2) + [\mu_1(\theta_1, \theta_2)]^2 = \mu_2(\theta_1, \theta_2)$$

$$\Rightarrow \begin{cases} \frac{1}{n}\sum_{i=1}^{n} Y_i = \mu_1(\theta_1, \theta_2) \\ \frac{1}{n}\sum_{i=1}^{n} Y_i^2 = h_2(\theta_1, \theta_2) + [\mu_1(\theta_1, \theta_2)]^2 \end{cases}$$

$$\Rightarrow \begin{cases} \bar{Y} = h_1(\theta_1, \theta_2) \\ S_M^2 = h_2(\theta_1, \theta_2) \end{cases}$$

- Una de las características más importantes de los estimadores es el sesgo que tienen, el cual también está relacionado con el error cuadrático medio. Además, también se puede introducir el concepto de UMVUE y de FCRLB, que permiten evaluar los estimadores con o sin sesgo

  - Siendo $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ una muestra donde $\boldsymbol{Y}$ tiene una función de densidad o de masa de probabilidad $f(\boldsymbol{y}|\boldsymbol{\theta})$ para $\boldsymbol{\theta} \in \Theta$, una función de valores reales $\tau(\boldsymbol{\theta})$ de $\boldsymbol{\theta}$, $T \equiv T(Y_1, Y_2, \dots, Y_n)$ un estimador de $\tau(\boldsymbol{\theta})$, y asumiendo que todas las esperanzas relevantes existen, se pueden dar las siguientes definiciones:

    - El sesgo o *bias* del estimador T para $\tau(\boldsymbol{\theta})$ se define de la siguiente manera:

$$Bias_{\tau(\boldsymbol{\theta})}(T) \equiv E_{\boldsymbol{\theta}}(T) - \tau(\boldsymbol{\theta})$$

- Un estimador es un estimador no sesgado para $\tau(\boldsymbol{\theta})$ si se cumple la siguiente igualdad:

$$E_{\boldsymbol{\theta}}(T) = \tau(\boldsymbol{\theta}) \;\Leftrightarrow\; Bias_{\tau(\boldsymbol{\theta})}(T) = 0 \quad for\ \forall \boldsymbol{\theta} \in \Theta$$

- A partir de estas definiciones, es posible definir el sesgo asintótico. Por tanto, se dice que un estimador es un estimador no sesgado asintóticamente para $\tau(\boldsymbol{\theta})$ si se cumple la siguiente condición

$$E_{\boldsymbol{\theta}}(T) \to \tau(\boldsymbol{\theta}) \quad as\ \ n \to \infty$$

- El error cuadrático medio o *mean square error* (MSE) estimador T para $\tau(\boldsymbol{\theta})$ se define de la siguiente manera:

$$MSE_{\tau(\boldsymbol{\theta})}(T) \equiv E_{\boldsymbol{\theta}}\left[\left(T - \tau(\boldsymbol{\theta})\right)^2\right] = Var_{\boldsymbol{\theta}}(T) + \left[Bias_{\tau(\boldsymbol{\theta})}(T)\right]^2$$

- Algunas características que caben mencionar sobre los conceptos anteriores son las siguientes:

  - En general, si el momento poblacional $E(Y^k)$ existe, entonces el momento muestral k no está sesgado para $E(Y^k)$

  - Si T es un estimador no sesgado para $\boldsymbol{\theta}$, en general, $g(T)$ no es un estimador no sesgado de $g(\boldsymbol{\theta})$ a no ser que g sea una función lineal. El ejemplo más claro es el del estimador de la varianza no sesgado S^2, debido a que S tiene sesgo para σ

  - Hay veces que no existen estimadores no sesgados para un parámetro o conjunto de parámetros concreto, o si existen, pueden ser absurdos

  - Tanto el sesgo como el MSE son funciones de $\boldsymbol{\theta}$ para $\boldsymbol{\theta} \in \Theta$ y $\boldsymbol{\theta}$ suele tomar valores reales

  - Si $MSE_{\tau(\boldsymbol{\theta})}(T_1) < MSE_{\tau(\boldsymbol{\theta})}(T_2)$ para toda $\boldsymbol{\theta} \in \Theta$, entonces T_1 es un mejor estimador de $\tau(\boldsymbol{\theta})$ que T_2, en el sentido de que un menor MSE es mejor que un mayor MSE. Por lo tanto, también se puede utilizar para evaluar dos estimadores sin importar sus características (con sesgo o sin sesgo)

  - Un problema común suele ser considerar una clase de estimadores $T_k(\boldsymbol{Y})$ para $\tau(\boldsymbol{\theta})$ donde $k \in \Lambda$ e intentar encontrar el MSE como función de k y encontrar el valor $k_o \in \Lambda$ que sea el minimizador global de $MSE(k) \equiv MSE(T_k)$

- Siendo $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ una muestra donde $\boldsymbol{Y}$ tiene una función de densidad o de masa de probabilidad $f(\boldsymbol{y}|\theta)$ para $\theta \in \Theta$, una función de valores reales $\tau(\theta)$ de θ, $U \equiv U(Y_1, Y_2, \dots, Y_n)$ un estimador de $\tau(\theta)$, y asumiendo que todas las esperanzas relevantes existen, se puede definir el siguiente concepto:

    - Un estimador U es el estimador de varianza mínima uniforme no sesgado o *uniformly mínimum variance unbiased estimator* (UMVUE) para $\tau(\theta)$ si U es un estimador no sesgado de $\tau(\theta)$ y $Var_\theta(U) \leq Var_\theta(W)$ para toda $\theta \in \Theta$ donde W es cualquier otro estimador no sesgado para $\tau(\theta)$

- Si $T(\boldsymbol{Y})$ es un estadístico completo suficiente para $\boldsymbol{\theta}$, entonces $U = g\big(T(\boldsymbol{Y})\big)$ es el UMVUE de su valor esperado $E_{\boldsymbol{\theta}}(U) = E_{\boldsymbol{\theta}}\big[g\big(T(\boldsymbol{Y})\big)\big]$. En particular, si $W(\boldsymbol{Y})$ es cualquier estimador no sesgado para $\tau(\theta)$, entonces el UMVUE de $\tau(\theta)$ es el siguiente:

$$U \equiv g\big(T(\boldsymbol{Y})\big) = E[W(\boldsymbol{Y})|T(\boldsymbol{Y})]$$

    - Si $Var_\theta(U) < \infty$ para toda $\theta \in \Theta$, entonces U es el único UMVUE para $\tau(\theta) = E_{\boldsymbol{\theta}}\big[g\big(T(\boldsymbol{Y})\big)\big]$

    - A este teorema se le llama teorema UMVUE de Lehmann-Scheffé (o *LSU theorem*), y este es el método más útil para encontrar UMVUE gracias a que si $Y_1, Y_2, \dots, Y_n$ son i.i.d. de una 1P-REF con función de densidad $f(\boldsymbol{y}|\theta) = h(\boldsymbol{y})c(\theta)\exp\big(w(\theta)t(\boldsymbol{y})\big)$ para $w(\theta) \in \Omega = (a, b)$, entonces $T(\boldsymbol{Y}) = \sum_{i=1}^{n} t(Y_i)$ es un estadístico suficiente completo

- Siendo $W \equiv W(\boldsymbol{Y})$ un estimador no sesgado de $\tau(\theta)$ y $T \equiv T(\boldsymbol{Y})$ un estadístico suficiente para θ, el estimador $\phi(T) = E(W|T)$ es un estimador no sesgado de $\tau(\theta)$ y $Var_\theta[\phi(T)] \leq Var_\theta[W]$ para toda $\theta \in \Theta$

    - A este teorema se le denomina teorema de Rao-Blackwell, y el proceso de encontrar un estimador como el descrito se suele llamar Rao-Blackwellización

    - El estimador $\phi(T)$ no depende de θ porque es un estadístico suficiente, y como $\tau(\theta) = E_\theta(W) = E_\theta\big(E(W|T)\big) = E_\theta\big(\phi(T)\big)$ por la ley de esperanzas iteradas, se puede ver como $\phi(T)$ no tiene sesgo

    - Por la fórmula de Steiner, se puede ver que la varianza de W viene dada por la siguiente fórmula:

$$Var(W) = E_\theta\big(Var(W|T)\big) + Var_\theta\big(E(W|T)\big) \geq Var_\theta\big(E(W|T)\big) = Var_\theta\big(\phi(T)\big)$$

- Algunos consejos para encontrar el UMVUE de un parámetro son los siguientes:

  - A partir del teorema LSU, si $T(\boldsymbol{Y})$ es un estadístico suficiente completo y $g\big(T(\boldsymbol{Y})\big)$ es una función que toma valores reales, entonces $U = g\big(T(\boldsymbol{Y})\big)$ es el UMVUE de su esperanza $E_{\boldsymbol{\theta}}\big[g\big(T(\boldsymbol{Y})\big)\big]$

  - Dado un estadístico suficiente completo $T(\boldsymbol{Y})$, el primer método para encontrar el UMVUE de $\tau(\theta)$ es adivinar una función g y demostrar que $E_{\boldsymbol{\theta}}\big[g\big(T(\boldsymbol{Y})\big)\big] = \tau(\theta)$ para toda $\theta \in \Theta$

  - Si $T(\boldsymbol{Y})$ es un estadístico suficiente completo, el segundo método es encontrar un estimador sin sesgo cualquiera $W(\boldsymbol{Y})$ de c. Entonces, $U(\boldsymbol{Y}) = E\big(W(\boldsymbol{Y})\big|T(\boldsymbol{Y})\big)$ es el UMVUE para $\tau(\theta)$

  - El problema con estos métodos es que es muy difícil adivinar g o encontrar un estimador no sesgado W y calcular $E\big(W(\boldsymbol{Y})\big|T(\boldsymbol{Y})\big)$ puede se difícil. Para ello, normalmente se pueden escribir ambos métodos para encontrar UMVUE y simplificar $E\big(W(\boldsymbol{Y})\big|T(\boldsymbol{Y})\big)$ o se puede encontrar información sobre estimadores no sesgados o sobre la esperanza condicional en el problema

- Siendo $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ una muestra donde $\boldsymbol{Y}$ tiene una función de densidad o de masa de probabilidad $f(\boldsymbol{y}|\theta)$ para $\theta \in \Theta$, el número de información o la información de Fisher para el parámetro θ se define de la siguiente manera:

$$I(\theta) \equiv E_\theta\left(\left[\frac{\partial}{\partial\theta}\log[f(\boldsymbol{Y}|\theta)]\right]^2\right)$$

  - Es una medida de la cantidad de información que una variable lleva sobre el parámetro θ. Si la variable contiene mucha información sobre el parámetro, entonces se pueden obtener estimaciones más precisas del parámetro a partir del estadístico, mientras que, si es muy pequeña, se tendrá menor precisión

  - Si $\eta = \tau(\theta)$, en donde $\tau'(\theta) \neq 0$, entonces se obtiene la siguiente expresión para la información de Fisher:

$$I(\eta) \equiv I\big(\tau(\theta)\big) = \frac{I(\theta)}{[\tau'(\theta)]^2} = \frac{E_\theta\left(\left[\frac{\partial}{\partial\theta}\log[f(\boldsymbol{Y}|\theta)]\right]^2\right)}{[\tau'(\theta)]^2}$$

- Las dos ecuaciones propuestas coinciden si $\tau'(\theta)$ es continua, $\tau'(\theta) \neq 0$ y además $\tau(\theta)$ es una función biyectiva, de modo que existe una función inversa $\theta = \tau^{-1}(\eta)$

- Si $f(\boldsymbol{y}|\theta)$ es una distribución de probabilidad que satisface $\frac{\partial}{\partial\theta} E_\theta\left(\frac{\partial}{\partial\theta}\log[f(\boldsymbol{Y}|\theta)]\right) = \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log[f(\boldsymbol{y}|\theta)]\right) f(\boldsymbol{y}|\theta)\right] d\boldsymbol{y}$ (lo cual se cumple para familias exponenciales), entonces se da la siguiente igualdad:

$$E_\theta\left(\left[\frac{\partial}{\partial\theta}\log[f(\boldsymbol{Y}|\theta)]\right]^2\right) = -E_\theta\left(\frac{\partial^2}{\partial\theta^2}\log[f(\boldsymbol{Y}|\theta)]\right)$$

- En el caso en donde las n observaciones son i.i.d, cada observación actúa como una variable aleatoria independiente y se cumple $I_n(\theta) = nI_1(\theta)$. Esto se puede demostrar de la siguiente manera:

$$E_\theta\left(\left[\frac{\partial}{\partial\theta}\log[f(\boldsymbol{Y}|\theta)]\right]^2\right) = E_\theta\left(\left[\frac{\partial}{\partial\theta}\log\left[\prod_{i=1}^{n} f(Y_i|\theta)\right]\right]^2\right) =$$

$$= E_\theta\left(\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log[f(Y_i|\theta)]\right]^2\right) =$$

$$= \sum_{i=1}^{n} E_\theta\left[\left(\frac{\partial}{\partial\theta}\log[f(Y_i|\theta)]\right)^2\right] + \sum_{i\neq j} E_\theta\left[\frac{\partial}{\partial\theta}\log[f(Y_i|\theta)]\frac{\partial}{\partial\theta}\log[f(Y_j|\theta)]\right]$$

$$\Rightarrow E_\theta\left[\frac{\partial}{\partial\theta}\log[f(Y_i|\theta)]\frac{\partial}{\partial\theta}\log[f(Y_j|\theta)]\right] =$$

$$= E_\theta\left[\frac{\partial}{\partial\theta}\log[f(Y_i|\theta)]\right] E_\theta\left[\frac{\partial}{\partial\theta}\log[f(Y_j|\theta)]\right] = 0$$

$$\Rightarrow \sum_{i=1}^{n} E_\theta\left[\left(\frac{\partial}{\partial\theta}\log[f(Y_i|\theta)]\right)^2\right] = nE_\theta\left[\left(\frac{\partial}{\partial\theta}\log[f(Y_i|\theta)]\right)^2\right]$$

$$as\ distributions\ are\ identical$$

- Si $Y_1, Y_2, \ldots, Y_n$ son i.i.d. de una distribución 1P-REF, si $\tau'(\theta)$ existe y es continua y si $\tau'(\theta) \neq 0$, entonces se cumple la siguiente igualdad:

$$I_n\big(\tau(\theta)\big) = \frac{nI_1(\theta)}{[\tau'(\theta)]^2}$$

- Siendo $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ una muestra y considerando una función del parámetro $\tau(\theta)$, en donde $\tau'(\theta) \neq 0$, la cota inferior de Fréchet-Cramér-Rao (FCRLB) para la varianza de los estimadores de $\tau(\theta)$ es la siguiente:

$$FCRLB\big(\tau(\theta)\big) = \frac{1}{I\big(\tau(\theta)\big)} = \frac{[\tau'(\theta)]^2}{E_\theta\left(\left[\frac{\partial}{\partial\theta}\log[f(\boldsymbol{Y}|\theta)]\right]^2\right)}$$

  - En particular, si $\tau(\theta) = \theta$, entonces $FCRLB(\theta) = 1/I(\theta)$

  - La interpretación sobre la información de Fisher se puede entender mejor al analizar como varia el $FCRLB\big(\tau(\theta)\big)$, dado que, si la información es mayor, la varianza del estimador es menor y, por tanto, la estimación es más precisa

  - Esta definición es más general, pero sigue siendo para estimadores no sesgados, dado que la esperanza del estimador sigue siendo $\tau(\theta)$ (el cual se considera el parámetro)

- Si $Y_1, Y_2, \dots, Y_n$ son i.i.d. de una distribución cualquiera con una función de densidad o de masa de probabilidad $f(y|\theta)$ y $W(Y_1, Y_2, \dots, Y_n) = W(\boldsymbol{Y})$ es un estimador cualquiera para $\tau(\theta) = E_\theta[W(\boldsymbol{Y})]$, entonces se cumple la siguiente desigualdad:

$$Var_\theta\big(W(\boldsymbol{Y})\big) \geq FCRLB_n\big(\tau(\theta)\big) = \frac{[\tau'(\theta)]^2}{I_n(\theta)} = \frac{\left(\frac{\partial}{\partial\theta}E_\theta[W(\boldsymbol{Y})]\right)^2}{nE_\theta\left(\left[\frac{\partial}{\partial\theta}\log[f(y|\theta)]\right]^2\right)}$$

  - El suscrito n tanto en el $FCRLB$ como en la información de Fisher indica que se utilizan sus equivalentes muestrales (se suma la información de Fisher para las n observaciones). Por lo tanto, lo único que cambia para el $FCRLB_n\big(\tau(\theta)\big)$ es que este se multiplica por $1/n$ (porque $I_n(\theta) = nI_1(\theta)$)

  - Si $Y_1, Y_2, \dots, Y_n$ son i.i.d. de una distribución 1P-REF con una función de densidad o de masa de probabilidad $f(y|\theta)$ y $W(Y_1, Y_2, \dots, Y_n) = W(\boldsymbol{Y})$ es un estimador cualquiera para $\tau(\theta) = E_\theta[W(\boldsymbol{Y})]$, entonces se cumple la siguiente desigualdad:

$$Var_\theta\big(W(\boldsymbol{Y})\big) \geq FCRLB_n\big(\tau(\theta)\big) = \frac{[\tau'(\theta)]^2}{I_n(\theta)} = \frac{[\tau'(\theta)]^2}{nI_1(\theta)}$$

        - Esta definición es más general, pero sigue siendo para estimadores no sesgados, dado que la esperanza del estimador sigue siendo $\tau(\theta)$ (el cual se considera el parámetro)

    - Los estimadores de máxima verosimilitud guardan una relación estrecha con la cota inferior de la varianza

        - Si un estimador para $\tau(\theta)$ alcanza el $FCRLB_n\big(\tau(\theta)\big)$, entonces el problema de maximización de la función de verosimilitud $L(\theta|\boldsymbol{y})$ tiene una única solución θ que maximiza la función

        - Para ciertas condiciones de regularidad, existe un estimador de máxima verosimilitud para $\tau(\theta)$ que es consistente, eficiente (tiene la mínima varianza) y asintóticamente normal

$$T(Y_i) \sim N\left(\tau(\theta), \tau'(\theta)\sqrt{\frac{1}{nI_1(\theta)}}\right)$$

El contraste de hipótesis estadísticas

- El método de inferencia anteriormente visto, la estimación puntual, no es el único método de inferencia, sino que existen muchos otros, como el contraste de hipótesis

    - El objetivo de un contraste de hipótesis es decidir, basándose en la muestra de la población estudiada, cuál de dos hipótesis complementarias son ciertas: la hipótesis nula y la alternativa

        - Una hipótesis es una proposición sobre un parámetro poblacional. Aunque esta definición es muy general, lo importante es que es una proposición sobre una característica de la población

        - Las dos hipótesis complementarias en un problema de contraste de hipótesis se denominan hipótesis nula, denotada por H_0, y la hipótesis alternativa, denotada por H_1

        - En un problema de contraste de hipótesis, después de observar la muestra, el investigador tiene que decidir si rechazar o no rechazar la hipótesis nula y aceptar la alternativa, por lo que tiene que decidir qué hipótesis es más consistente con la evidencia empírica

    - En estadística, las hipótesis se formulan en términos paramétricos y la evidencia empírica proviene de los datos de la muestra a través de la probabilidad

- Una hipótesis paramétrica es una proposición sobre los parámetros desconocidos $\boldsymbol{\theta}$ de la distribución de una variable aleatoria $Y \sim f(y|\boldsymbol{\theta})$ en donde $\boldsymbol{\theta} \in \Theta$

- Existen otros tipos de hipótesis, tales como hipótesis sobre la igualdad de dos o más parámetros de dos o más distribuciones o sobre la forma de la función de distribución de Y (en donde se usan otros tipos de contrastes, como los de bondad de ajuste)

- Las hipótesis se pueden clasificar entre simples, que son hipótesis que especifican la distribución poblacional completamente, y compuestas, que no especifican completamente la distribución poblacional, tales como aquellas que no son una igualdad (como las hipótesis alternativas para contrastes de una cola o de dos colas)

$$(simple) \quad \theta = \theta_0$$

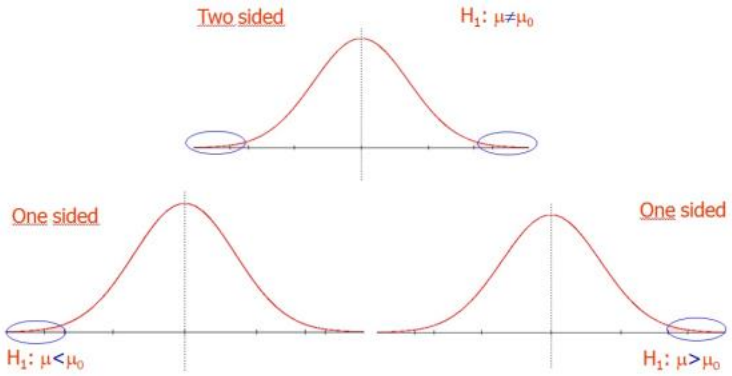
$$(composite) \quad \theta \neq \theta_0 \; ; \; \theta > \theta_0 \; ; \; \theta < \theta_0$$

- Si θ denota un parámetro poblacional, el formato general de la hipótesis nula y la hipótesis alternativa es $H_0: \theta \in \Theta_0$ y $H_1: \theta \in \Theta_0^c$, en donde $\Theta_0, \Theta_0^c \subseteq \Theta$

- Un procedimiento de contraste de hipótesis o contraste de hipótesis es una regla que especifica para qué valores muestrales la decisión es no rechazar H_0 y para qué valores muestrales la decisión es rechazar la H_0 y aceptar la alternativa H_1

  - Normalmente, un contraste de hipótesis está especificado en términos de un estadístico de contraste $W(Y_1, Y_2, \dots, Y_n) = W(\boldsymbol{Y})$, que es una función de la muestra

  - El subconjunto del espacio muestral para el cual H_0 se rechazaría se denomina región de rechazo o región crítica R. El complemento de la región crítica se denomina región de aceptación R^c, y los valores que delimitan ambas regiones son los valores críticos, los cuales son los valores tal que $P(W(\boldsymbol{Y}) \in R | H_0 \; is \; true) = \alpha$



- En los contrastes de dos colas, el nivel de significación α se divide entre dos, de modo que la región crítica se divide entre las dos colas y los valores críticos son más grandes en valor absoluto

- Si los datos son consistentes con la hipótesis nula, entonces la hipótesis nula no se rechaza, pero en ningún caso se demuestra que la hipótesis nula es cierta

- Los dos enfoques más famosos para contrastar hipótesis fueron desarrollados por Fisher y por Neyman y Pearson

  - El enfoque de Fisher consistía en construir una hipótesis nula H_0, escoger un estadístico con distribución muestral conocida que mida la desviación respecto a la H_0, obtener datos de las poblaciones y comparar el valor de estadístico de las muestras con la distribución muestral. Una vez hecho esto, se determinaba un p-valor o *p-value*, que es la probabilidad asociada a obtener el valor muestral del estadístico (o uno más extremo) si H_0 es verdad, y se rechaza o no esta hipótesis dependiendo del valor

  - El enfoque de Neyman-Pearson se basa en escoger un nivel de significación predeterminado y después obtener los datos. Además, se incorpora explícitamente una hipótesis alternativa H_1 en su esquema, la cual es la hipótesis que debe ser verdad si H_0 no lo es, y se desarrolla el concepto de error de tipo I y error de tipo II, derivando al concepto de poder de un contraste

- Consecuentemente, se puede desarrollar un esquema para el contraste de hipótesis, en donde se siguen los siguientes pasos:

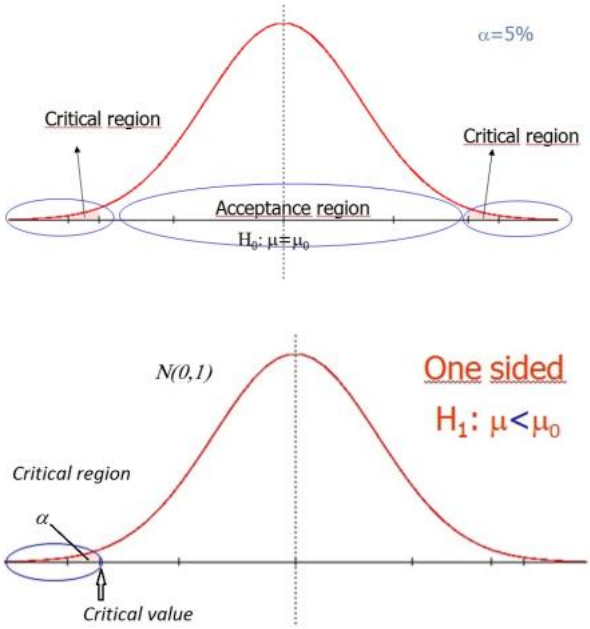
  - Se tiene que especificar H_0, H_1 y un estadístico de contraste apropiado

- Se especifica (a priori) un nivel de significación, el cual será la frecuencia asintótica de errores de tipo I que se está dispuesto a aceptar

- Se recogen datos de una o más muestras aleatorias de las poblaciones de interés y se calcula el estadístico de contraste a partir de los datos muestrales

- Se compara el valor del estadístico de contraste a su distribución muestral, asumiendo que H_0 es cierta. Esto se hace determinando si el valor del estadístico está dentro de la región crítica R o a través de comparar el *p-value* con el nivel de significación

- Si el estadístico de contraste está dentro de R o su *p-value* es menor al nivel de significación, se rechaza H_0 (es un resultado significativo)



- Si el estadístico está fuera de la región crítica R o el *p-value* es mayor al nivel de significación, no hay evidencia suficiente para poder rechazar H_0

- El método de la razón de verosimilitud para el contraste de hipótesis está relacionado con los estimadores de máxima verosimilitud y son aplicables en varios contextos

    - El estadístico de contraste de razón de verosimilitud o *likelihood ratio test statistic* para contrastar $H_0: \theta \in \Theta_0$ contra $H_1: \theta \in \Theta_0^c$ se puede definir de la siguiente manera:

$$\lambda(\boldsymbol{y}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\boldsymbol{y})}{\sup_{\theta \in \Theta} L(\theta|\boldsymbol{y})}$$

- Un contraste de razón de verosimilitud o *likelihood ratio test* (LRT) es cualquier contraste que tenga un área de rechazo de la forma $R = \{\boldsymbol{y} : \lambda(\boldsymbol{y}) \leq k_\alpha\}$, donde k_α es cualquier número satisfaciendo $0 \leq k_\alpha \leq 1$ y $P(\lambda(\boldsymbol{y}) \leq k_\alpha | H_0) = \alpha$

- El numerador es la máxima probabilidad de haber obtenido la muestra observada bajo los parámetros de la hipótesis nula Θ_0, mientras que el denominador es la probabilidad máxima de haber obtenido la muestra observada considerando todos los parámetros posibles Θ

- La correspondencia entre los estimadores máximo verosímiles y los LRT es más clara si se considera la maximización sobre el espacio paramétrico. Suponiendo que $\hat{\theta}$ es el MLE de θ que se obtiene maximizando sobre todo el espacio paramétrico, y el parámetro $\hat{\theta}_0$ es el MLE bajo el espacio paramétrico restringido, entonces el estadístico del LRT es el siguiente:

$$\lambda(\boldsymbol{y}) = \frac{L(\hat{\theta}_0|\boldsymbol{y})}{L(\hat{\theta}|\boldsymbol{y})}$$

- Por lo tanto, siempre se tiene que hacer una estimación máxima verosímil en el subespacio paramétrico de la hipótesis nula y en el espacio paramétrico de la hipótesis alternativa. De este modo, solo se tiene que sustituir el valor dado en H_0 si la hipótesis nula es simple, pero se tiene que hacer una estimación cuando se involucra más de un parámetro (comparación de igualdad de parámetros) o cuando no se especifica otro parámetro necesario (como cuando la desviación estándar o la media son desconocidas)

- Obviamente, este estadístico solo toma valores entre cero y uno (al ser una razón). Valores cercanos a uno muestran que la evidencia apoya no rechazar la hipótesis nula (la verosimilitud bajo la hipótesis nula es mucho mayor a la alternativa), mientras que valores cercanos a cero apoya el rechazo de la hipótesis (la verosimilitud bajo la hipótesis alternativa es mucho mayor a la nula)

- Como se ha visto, el LRT depende de k_α, pero este depende de la distribución de $\lambda(\boldsymbol{y})$. No obstante, se pueden considerar otras alternativas a la de encontrar la distribución exacta de $\lambda(\boldsymbol{y})$

- Si $T(\boldsymbol{Y})$ es un estadístico suficiente para θ y que $\lambda^*(\boldsymbol{y})$ y $\lambda(\boldsymbol{y})$ son los estadísticos LRT basados en T y $\boldsymbol{Y}$, respectivamente, entonces $\lambda^*\big(T(\boldsymbol{y})\big) = \lambda(\boldsymbol{y})$ para toda $\boldsymbol{y}$ en el espacio muestral

- El sentido del teorema anterior es que como $T(\boldsymbol{y})$ contiene toda la información sobre θ en $\boldsymbol{y}$, un contraste basado en T debe ser igual de bueno que uno basado en la muestra completa $\boldsymbol{Y}$. Por lo tanto, se puede utilizar la distribución de $\lambda^*(\boldsymbol{y})$ en vez de la de $\lambda(\boldsymbol{y})$

- En la mayoría de casos, la distribución exacta de la razón es muy difícil de determinar, pero el teorema de Wilks permite obtener una distribución para cualquier estadístico de razón de verosimilitud

- Siendo k el número de parámetros independientes en Θ y r el número de parámetros independientes en Θ_0, bajo algunas condiciones de regularidad generales y para grandes muestras (asintóticamente), si H_0 es cierta, entonces el siguiente estadístico sigue una distribución chi cuadrada con $k - r$ grados de libertad:

$$-2\ln\lambda(\boldsymbol{y}) \sim \chi^2_{k-r}$$

- El criterio de decisión bajo este tipo de contraste es que, si $-2\ln\lambda(\boldsymbol{y}) < \chi^2_{k-r,\alpha}$, entonces no se rechaza la hipótesis nula, mientras que si $-2\ln\lambda(\boldsymbol{y}) \geq \chi^2_{k-r,\alpha}$, se rechaza la hipótesis nula. En este caso, $\chi^2_{k-r,\alpha}$ es un valor crítico tal que se cumple $P\big(\chi^2_{k-r} > \chi^2_{k-r,\alpha}\big) = \alpha$

- Al decidir si rechazar o no rechazar una hipótesis nula, uno puede equivocarse a la hora de decidir, por lo que normalmente se evalúan los contrastes de hipótesis a través de la probabilidad de cometer estos errores

  - Un contraste de hipótesis para $H_0: \theta \in \Theta_0$ contra $H_1: \theta \in \Theta_0^c$ puede cometer dos tipos de errores: el error de tipo I y el error de tipo II

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

- Si $\theta \in \Theta_0$ pero el contraste de hipótesis rechaza incorrectamente la hipótesis nula H_0, entonces el contraste ha cometido un error tipo I (conocido también como falso positivo). En un contraste con significación α, la probabilidad de cometer el error será α (el máximo nivel de error de tipo I que se quiere asumir)

$$P(\ Reject\ H_0 \mid H_0\ is\ true\ ) = \alpha$$

- Si, en cambio, $\theta \in \Theta_0^c$ pero el contraste decide no rechazar H_0, entonces el contraste ha cometido un error tipo II (conocido también como falso negativo). En un contraste de nivel de significación α, la probabilidad de cometer el error será β

$$P(\ Not\ reject\ H_0 \mid H_0\ is\ false\ ) = \beta$$

- Como se puede ver, al ser probabilidades condicionales, las probabilidades de cometer un error no son el complemento de las probabilidades de cometer el otro tipo, sino que son diferentes

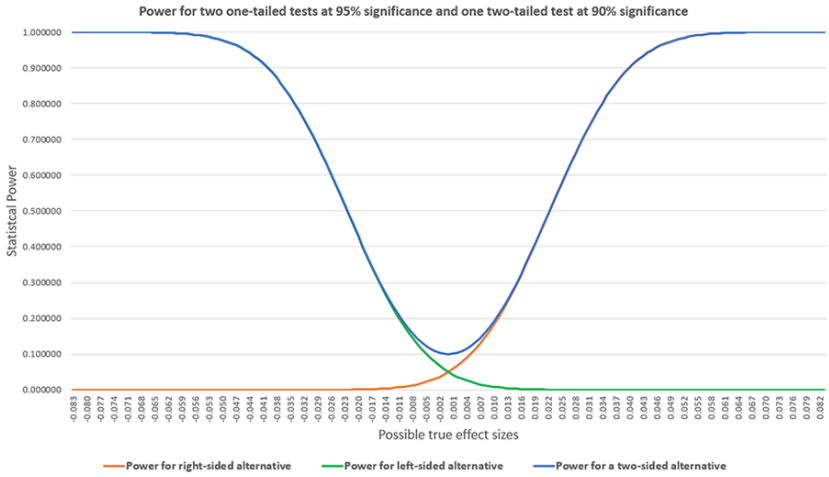
- Suponiendo que R denota la zona de rechazo para un contraste de hipótesis, entonces para $\theta \in \Theta_0$, el contraste cometerá un error si $\boldsymbol{x} \in R$ (porque se rechazaría), por lo que la probabilidad de error de tipo I es $P_\theta(\boldsymbol{x} \in R)$. Además, para $\theta \in \Theta_0^c$, la probabilidad de que se cometa un error de tipo II es $P_\theta(\boldsymbol{x} \in R^c)$ (donde R^c es la región de aceptación)

  - Como $P_\theta(\boldsymbol{x} \in R^c) = 1 - P_\theta(\boldsymbol{x} \in R)$, entonces la función de θ, $P_\theta(\boldsymbol{x} \in R)$, contiene toda la información sobre el contraste con región de rechazo R. Esta función se puede expresar de la siguiente manera:

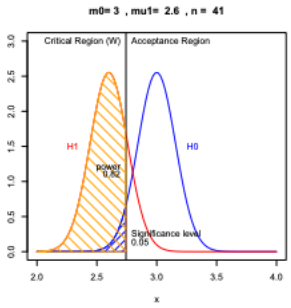
$$P_\theta(\boldsymbol{x} \in R) = \begin{cases} P_\theta(\boldsymbol{x} \in R) & if\ \theta \in \Theta_0 \\ 1 - P_\theta(\boldsymbol{x} \in R^c) & if\ \theta \in \Theta_0^c \end{cases}$$

  - La función de de un contraste de hipótesis con región de rechazo R es la función de θ definida por $\beta(\theta) = P_\theta(\boldsymbol{x} \in R)$. Esta función depende de si la hipótesis alternativa es simple o compuesta, teniendo una menor potencia en la región en la que no se rechazaría la hipótesis nula

potencia



- Suponiendo que se sabe el parámetro de la distribución real (bajo la hipótesis alternativa), la potencia de un contraste se puede calcular a través del valor crítico de la distribución muestral bajo la hipótesis nula. Esta será la probabilidad acumulada del valor crítico en la distribución muestral bajo la hipótesis alternativa (que sería $1 - P_\theta(\boldsymbol{x} \in R)$)




```
> mu0<-3
> mu1<-2.6
> sd<-1
> n<-41
> sd0<-sd/sqrt(n)
> cv=-1.645*sd0+mu0
> cv

[1] 2.743094

> pnorm(cv,mu1,sd0,lower.tail=T)

[1] 0.820232
```

- Idealmente, se quiere que $P_\theta(\boldsymbol{x} \in R) = 0$ para $\theta \in \Theta_0$ y que $P_\theta(\boldsymbol{x} \in R) = 1$ para $\theta \in \Theta_0^c$, pero esto es imposible excepto en situaciones triviales, por lo que se tiene que aproximar a esos valores ideales. Por tanto, para una muestra específica, ambos errores están vinculados (reducir las probabilidades de rechazar cuando la hipótesis nula es cierta aumenta las de no rechazar cuando la hipótesis nula es falsa)

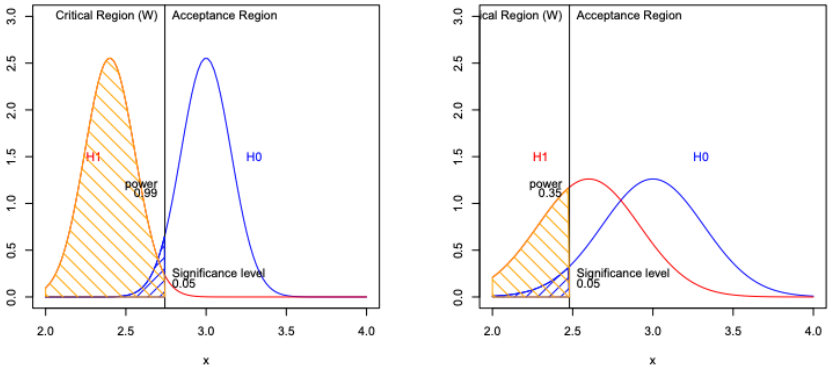
- Típicamente, la función de poder de un contraste dependerá del tamaño de la muestra n. Si n puede ser escogido por el experimentador, considerar la función de poder puede ayudar a determinar el tamaño muestral apropiado para el experimento, dado que aumentar el tamaño muestral puede reducir las probabilidades de cometer ambos tipos de errores a la vez

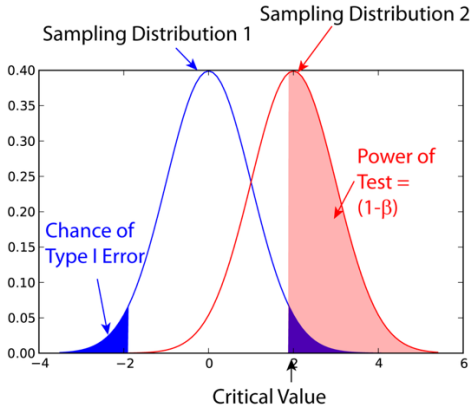
- Se puede dudar sobre la existencia de problemas de potencia en un contraste cuando no se rechaza la hipótesis nula, dado que la

potencia es la probabilidad de rechazar la hipótesis nula cuando esta es falsa

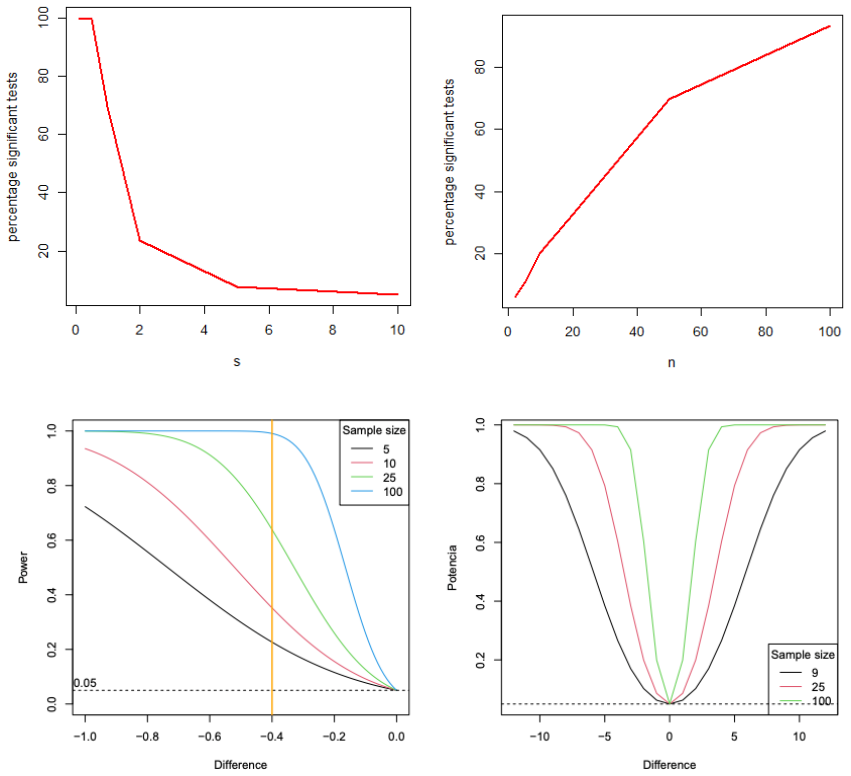
- Los parámetros que afectan a la función de potencia de un contraste son el valor del parámetro real (o el diferencial sobre el parámetro de H_0), el error estándar y el nivel de significación

  - Cuando el parámetro real (aquel parámetro que está incluido en el complementario del espacio paramétrico de la hipótesis nula) está muy alejado del parámetro bajo la hipótesis nula, la potencia del contraste incrementa (es más probable rechazar la hipótesis nula cuando es falsa debido a la distribución real). Por lo tanto, un diferencial más grande en valor absoluto hace que aumente el poder del contraste y que el tamaño de la muestra lo reduzca





  - Cuando el error estándar es mayor, el valor del contraste de hipótesis será menor, y por tanto se tenderá a rechazar más la hipótesis nula, incrementando así la probabilidad de rechazar la hipótesis nula cuando esta es falsa. Por lo tanto, un incremento de la desviación estándar (ya sea poblacional o estimada) disminuye la potencia del contraste (crea valores del estadístico más bajos), mientras que un incremento en el tamaño muestral la aumenta (crea valores del estadístico más altos)



- Cuando el nivel de significación es mayor, la probabilidad de rechazar la hipótesis nula cuando esta es falsa incrementa, de modo que se observa un nivel de potencia mayor para los valores diferentes a la media

- Para un tamaño muestral fijo, es normalmente imposible hacer arbitrariamente pequeños las probabilidades de los dos tipos de error, por lo que, al buscar un buen contraste, es común restringir la consideración a contrastes que controlan la probabilidad del error de tipo I a un nivel especificado

  - Dentro de esta clase de contrastes, lo que se hace después es buscar aquel con las probabilidades de cometer un error de tipo II lo más bajas posible

  - Normalmente se escoge un nivel de significación de $\alpha = 0.01$, $\alpha = 0.05$ o $\alpha = 0.1$. Cuando se fija una α pequeña, el experimentador se asegura de decir que la evidencia empírica no permite rechazar la hipótesis nula cuando es falsa

  - No obstante, cuando se fija el nivel de un contraste, se controlan solo las probabilidades del error de tipo I y no las del error de tipo II, de modo que el experimentador debe especificar las hipótesis de una manera tal que controlar el error de tipo I sea lo más importante (haciendo que la hipótesis alternativa sea la esperada)

- Incrementando el tamaño de la muestra, sin embargo, se pueden reducir ambos tipos de errores (dado que se incrementa la potencia de los contrastes)

- Se pueden definir dos términos relacionados con los contrastes de hipótesis al discutir como controlar probabilidades de error del tipo I: el tamaño y el nivel de un contraste

  - Para $0 \leq \alpha \leq 1$, un contraste con una función de poder $\beta(\theta)$ es un contraste de tamaño α si $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$

  - Para $0 \leq \alpha \leq 1$, un contraste con una función de poder $\beta(\theta)$ es un contraste de nivel α si $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$

  - Muchos autores no hacen distinción entre ambas nociones y se pueden intercambiar, pero según las definiciones hechas, el conjunto de contrastes de nivel α contiene un conjunto de contrastes de tamaño α (puede ser de ese tamaño o menor)

  - Además, esta distinción entre nociones es útil en situaciones complejas, en donde es difícil construir un contraste con un tamaño α. En estas situaciones, un experimentador estaría satisfecho con un contraste de nivel α, sabiendo que se tendrán que hacer algunos compromisos

- A parte de los niveles α, hay otras características de un contraste que son interesantes, como su sesgo. Es interesante saber si un contraste es más propenso a rechazar H_0 si $\theta \in \Theta_0^c$ que si $\theta \in \Theta_0$

  - Un contraste con función de poder $\beta(\theta)$ no está sesgado si $\beta(\theta') \geq \beta(\theta'')$ para cualquier $\theta' \in \Theta_0^c$ y $\theta'' \in \Theta_0$. En otras palabras, un contraste no está sesgado si la probabilidad de rechazar H_0 es mayor o igual cuando H_0 es falsa que cuando es cierta

  - En la mayoría de problemas se tienen contrastes no sesgados. Por lo tanto, hay muchos contrastes con tamaño α, contrastes de razón de verosimilitud, etc. No obstante, hay criterios para seleccionar qué tipo de contraste usar dependiendo en su función de poder

- Después de que una hipótesis se haya realizado, las conclusiones se tienen que informar de una manera estadísticamente significativa. Una manera de poder informar los resultados es a través de un valor de un cierto tipo de estadístico de contraste llamado *p-value*

- El p-valor o *p-value* $p(\boldsymbol{Y})$ es un estadístico de contraste que satisface $0 \leq p(\boldsymbol{y}) \leq 1$ para cualquier punto muestral $\boldsymbol{y}$

    - Los valores pequeños de $p(\boldsymbol{Y})$ aportan evidencia de que H_0 se puede rechazar, mientras que los grandes aportan de lo contrario

    - El p-valor es la probabilidad de obtener los datos $\boldsymbol{y}$ (o el valor del estadístico en valor absoluto, en consecuencia) bajo la hipótesis nula, pero no la probabilidad de que la hipótesis nula sea cierta habiendo obtenido los datos que se tienen

    - Una ventaja de informar el *p-value* es que el lector puede escoger la α del contraste que considere apropiado y compararla con el $p(\boldsymbol{y})$ para saber si loss datos comportan el rechazo o no de la hipótesis nula

    - Mientras menor sea el $p(\boldsymbol{Y})$, mayor es la evidencia para rechazar $p(\boldsymbol{Y})$, por lo que el *p-value* permite explicar más que solo el informar si se rechaza o no la hipótesis nula

- Un *p-value* es válido si, para toda $\theta \in \Theta_0$ y cualquier $0 \leq \alpha \leq 1$, se cumple que $P_\theta(p(\boldsymbol{Y}) \leq \alpha) \leq \alpha$

    - Si el *p-value* es válido, entonces es fácil construir un contraste de nivel α basado en $p(\boldsymbol{Y})$: el contraste que rechace H_0 si, y solo si, $p(\boldsymbol{Y}) \leq \alpha$ es un contraste de nivel α porque $P_\theta(p(\boldsymbol{Y}) \leq \alpha) \leq \alpha$

    - Siendo $W(\boldsymbol{y})$ un estadístico de contraste tal que valores grandes de W den evidencia de que H_0 no se puede rechazar, para cada punto muestral $\boldsymbol{y}$, $p(\boldsymbol{y})$ es válido si se define de la siguiente manera:

$$p(\boldsymbol{y}) = \sup_{\theta \in \Theta_0} P_\theta\big(W(\boldsymbol{Y}) \geq W(\boldsymbol{y})\big)$$

- Otro método para definir un *p-value* válido requiere condicionar el estadístico suficiente. Suponiendo que $S(\boldsymbol{Y})$ es un estadístico suficiente para el modelo $\{f(\boldsymbol{y}|\boldsymbol{\theta}) : \theta \in \Theta_0\}$, si la hipótesis nula es cierta, la distribución condicional de $\boldsymbol{Y}$ dado $S = s$ no dependerá de θ

    - Siendo $W(\boldsymbol{Y})$ un estadístico de contraste para el cual grandes valores evidencian que H_0 se puede rechazar, se define $p(\boldsymbol{y})$ para cada punto muestral de la siguiente manera:

$$p(\boldsymbol{y}) = P\big(W(\boldsymbol{Y}) \geq W(\boldsymbol{y}) \big| \, S = S(\boldsymbol{y})\big)$$

- Considerando solo la distribución única que sea la distribución condicional de $\boldsymbol{Y}$ dado $S = s$ (la cual no depende de los parámetros), se puede ver que para toda $0 \leq \alpha \leq 1$ se cumple la siguiente desigualdad:

$$P(p(\boldsymbol{Y}) \leq \alpha | \, S = s) \leq \alpha$$

- Entonces, para cualquier $\theta \in \Theta_0$, se tiene que la probabilidad incondicional (que depende de los parámetros) respeta la siguiente desigualdad:

$$P_\theta(p(\boldsymbol{Y}) \leq \alpha) = \int_{\boldsymbol{y}} P(p(\boldsymbol{Y}) \leq \alpha | \, S = s) P_\theta(S = s)$$
$$\leq \int_{\boldsymbol{y}} \alpha P_\theta(S = s) = \alpha \int_{\boldsymbol{y}} P_\theta(S = s) = \alpha$$

- Los resultados muestran que este $p(\boldsymbol{y})$ es un *p-value* válido, y aunque se ha asumido una S continua, normalmente este método se aplica con S discreta (se sustituye la integral por la sumatoria)

- El *p-value* $p(\boldsymbol{Y})$ de un contraste se puede calcular de diferentes maneras dependiendo de la hipótesis alternativa

  - El *p-value* de un contraste de una cola se puede calcular como la probabilidad acumulada a la izquierda del valor del estadístico si la cola es la inferior, o como la probabilidad acumulada a la derecha del estadístico si la cola es la superior

$$p(\boldsymbol{y}) = P_\theta(W(\boldsymbol{y}) \leq c_\alpha) \quad (one - sided\ left)$$

$$p(\boldsymbol{y}) = P_\theta(W(\boldsymbol{y}) \geq c_\alpha) \quad (one - sided\ right)$$

  - El *p-value* de un contraste de dos colas se puede calcular como la probabilidad acumulada a la izquierda y a la derecha del valor absoluto del valor del estadístico (dado que la región crítica se divide en ambas colas y se tiene que medir la probabilidad en ambas)

$$p(\boldsymbol{y}) = 2P_\theta\big(W(\boldsymbol{y}) \geq c_{1-\alpha/2}\big) \quad if \;\; W(\boldsymbol{y}) > 0$$

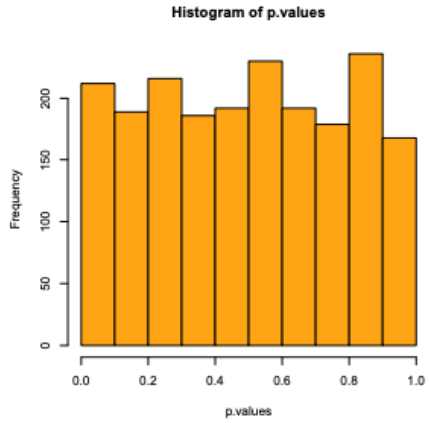
$$p(\boldsymbol{y}) = 2P_\theta\big(W(\boldsymbol{y}) \leq c_{\alpha/2}\big) \quad if \;\; W(\boldsymbol{y}) < 0$$

$$(two - sided)$$

  - Como se puede ver, en los contrastes de dos colas, el valor $p(\boldsymbol{y})$ será más grande en valor absoluto que en los contrastes de una

cola, debido a que se divide el nivel de significación α entre las dos colas, por lo que $c_{\alpha/2} < c_\alpha$ y $c_{1-\alpha/2} > c_{1-\alpha}$

- Si la hipótesis nula es cierta, cualquier *p-value* tiene la misma probabilidad de obtenerse (sigue una distribución uniforme estándar), y la proporción de *p-values* que rechazarían esta hipótesis nula cierta serían α



  - Siendo $F^{\leftarrow}(y) \equiv \inf\{x : F(x) \geq y\}$ la inversa generalizada de una función de distribución y W un estadístico con función de distribución F bajo la hipótesis nula, el *p-value* $p(\boldsymbol{y}) = F\big(W(\boldsymbol{y})\big)$ tiene una función de distribución uniforme

$$P\big(p(\boldsymbol{Y}) < p(\boldsymbol{y})\big) = P\left(F\big(W(\boldsymbol{Y})\big) < F\big(W(\boldsymbol{y})\big)\right) =$$

$$= P\left(F^{\leftarrow}\big(p(\boldsymbol{Y})\big) < F^{\leftarrow}\big(p(\boldsymbol{y})\big)\right) = P\big(W(\boldsymbol{Y}) < W(\boldsymbol{y})\big) =$$

$$= 1 - P\big(W(\boldsymbol{Y}) \geq W(\boldsymbol{y})\big) = 1 - p(\boldsymbol{y}) \in [0,1]$$

  - No obstante, no pasa nada, dado que el *p-value* no expresa que tan probable es H_0, sino que se puede considerar solo un mecanismo para rechazar a un cierto nivel α

- Cuando se realizan múltiples contrastes de hipótesis, la inferencia sobre estos debe corregirse debido al efecto de realizar varios contrastes en relación a los errores al concluir sobre los resultados

  - En general, cuando se hacen m contrastes de hipótesis independientes y no hay diferencias entre los grupos de la población para los cuales se realizan los contrastes, la probabilidad de que haya al menor un error de tipo I es de $1 - (1 - \alpha)^m$, lo que se suele llamar tasa de error de tipo I familiar o *family-wise type I error rate* (FWER)

- La probabilidad es $1-(1-\alpha)^m$ porque $(1-\alpha)$ es la probabilidad de no cometer el error de tipo I y $(1-\alpha)^m$ es la probabilidad de no cometer ningún error de tipo I en todos los contrastes de hipótesis, por lo que $1-(1-\alpha)^m$ es la probabilidad de cometer el error de tipo I en todos estos contrastes

$$FWER = P\left(\bigcup_{i=1}^{m} A_i\right) = 1 - P\left[\left(\bigcup_{i=1}^{m} A_i\right)^C\right] =$$

$$= 1 - P\left(\bigcap_{i=1}^{m} A_i^C\right) = 1 - (1-\alpha)^m$$

$$\Rightarrow\ FWER = 1 - (1-\alpha)^m$$

- Como se puede ver, este error incrementaría cuantos más contrastes se realicen si no son perfectamente idénticos (dependencia perfecta positiva), y eso causaría que hubiera una alta probabilidad (FWER) de rechazar incorrectamente una hipótesis nula. Por lo tanto, es necesario hacer una corrección de los *p-value* que se obtienen de estos contrastes con tal de reducir esta probabilidad de error de tipo I para múltiples contrastes

- Aún si no se asume que los contrastes son independientes, igualmente se puede utilizar la desigualdad de Boole para ver que el error familiar (la probabilidad de que se rechace incorrectamente al menos un contraste) es menor o igual a suma de los errores individuales de cada contraste o comparación. Asumiendo que el evento A es el evento de que se rechace incorrectamente, entonces:

$$FWER = P\left(\bigcup_{i=1}^{m} A_i\right) \leq \sum_{i=1}^{m} P(A_i) = \sum_{i=1}^{m} \alpha = m\alpha$$

$$\Rightarrow\ FWER \leq m\alpha$$

- Los tipos de correcciones más utilizados son el control de la FWER y el control de la tasa de error de descubrimiento falso o *false discovery rate* (FDR)

- Controlando el FWER en el *p-value* se controla la probabilidad de cometer uno o más errores de tipo I cuando se realizan múltiples contrastes de hipótesis y no hay diferencias entre los grupos de la población para los cuales se realizan estos. Las correcciones más

importantes que controlan el FWER son la corrección de Bonferroni y la corrección de Hölm

- La corrección de Bonferroni se basa en ajustar el nivel de significación de cada contraste a α/m, de modo que es equivalente a multiplicar los *p-values* por el número de contrastes hechos (m)

$$p - value_{Bonferroni} = p - value \times m$$

$$\Rightarrow FWER_{corrected} = 1 - \left(1 - \frac{\alpha}{m}\right)^{m} \leq \alpha$$

- Este método es muy conservador, dado que al hacer esto se reduce mucho el número de contrastes significativos (diferencias entre los grupos de la población más sutiles no serán significativas), pero potencialmente incrementa el número de contrastes no significativos que tendrían que serlo (error de tipo II). Este método solo se tendría que usar cuando se quiere evitar más la aparición de falsos positivos que de falsos negativos, cuando se tienen muchísimas observaciones

- La corrección de Hölm es muy similar a la de Bonferroni, solo que en vez de reducir el nivel de significación de cada contraste a α/m, la reduce a $\alpha/(m+1-k)$ para $k = 1,2,\dots,m$, por lo que es equivalente a multiplicar los *p-values* ordenados de manera ascendente por esta cantidad ajustada $m+1-k$. Este método es menos conservador que el de Bonferroni, pero sigue siendo conservador debido a la tasa de errores de tipo II

- Controlando el FDR se controla la probabilidad de que dentro de los contrastes estadísticamente significativos haya una tasa de α contrastes significativos por pura aleatoriedad. Para este procedimiento, se asume que los contrastes son independientes (aunque también funciona para formas concretas de dependencia positiva)

  - Este FDR será el valor esperado del número de falsos positivos entre los positivos totales (no sobre todos los contrastes realizados)

  - Usando el procedimiento de Benjamini-Hochberg los *p-values* de menor a mayor y multiplicar el vector de *p-values* (ordenados de manera ascendente) por m/i para $i = 1,2,\dots,m$ y reinstaurar el orden original

$$p - value_{FDR} = p - value \times \frac{c}{i} \quad for \;\; i = 1,2,\dots,m$$

$$p_{FDR_i} = \min(p - value_{FDR_i}, p - value_{FDR_{i+1}})$$

- La lógica de este tipo de corrección es que, si muchas variables son significativas, entonces debe de haber algunos efectos verdaderos identificados por la significación del contraste (hay diferencias entre los grupos de la misma población). La probabilidad de un falso positivo (de un error de tipo I) debe ser, por tanto, mucho más pequeña de lo que la FWER sugiere. Esto quiere decir que los *p-values* deben de multiplicarse por una fracción de m (no por m)

- Como el *p-value* más significativo tiene un rango 1, entonces este se multiplicará por m, de modo que, como mucho, se puede obtener un porcentaje de contrastes significativos igual al de la corrección de Bonferroni dejando más contrastes significativos

- Esta corrección es menos conservadora que las vistas para el FWER, dado que permite reducir el error de tipo II que se obtendría al realizar los ajustes anteriores y a la vez reduce el error de tipo I. También se suele utilizar este procedimiento cuando los efectos de interés no son muy grandes o consistentes (son sutiles) o cuando se tiene una muestra limitada. Si los contrastes están correlacionados, no obstante, entonces este procedimiento no se puede utilizar porque se asumen contrastes independientes, mientras que no se asume

- Para que los contrastes de hipótesis paramétricos puedan producir resultados precisos, las suposiciones subyacentes deben satisfacerse, pero estas rara vez se satisfacen en los datos y, aunque muchos contrastes son robustos a violaciones de las suposiciones, esta robustez es limitada

    - Por lo tanto, contrastes robustos no paramétricos permiten aliviar los problemas inherentes de usar métodos paramétricos cuando se violan las suposiciones

        - Aunque hay una gran variedad de contrastes no paramétricos basados en rango (el orden de los datos muestrales según su valor), los dos más importantes son el contraste de Mann-Whitney-Wilcoxon para muestras independientes y el contraste de rango con signo de Wilcoxon para muestras pareadas

        - Se dice que el contraste de Mann-Whitney-Wilcoxon tiene una mayor eficiencia que el contraste t para distribuciones no normales y que es casi tan eficiente como este en el caso de distribuciones normales

- No obstante, los estadísticos clásicos no paramétricos no son robustos cuando se utilizan para analizar datos heteroscedásticos, y solo son apropiados para analizar datos específicos. La heterocedasticidad significa que la varianza de los errores no es constante entre observaciones

- En ambos casos, las hipótesis nulas y las hipótesis alternativas son equivalentes:

  - La hipótesis nula dice que, para dos muestras que vienen de poblaciones idénticamente distribuidas, la probabilidad de que una observación de una población exceda una observación de la segunda es igual a la probabilidad de que ocurra justo lo inverso (la probabilidad de ambos eventos es la misma)

  - La hipótesis alternativa dice que las muestras provienen de poblaciones que difieren en su localización, de modo que la probabilidad de que una observación exceda a una observación de la otra no es la misma que la probabilidad de que ocurra lo inverso

- La transformación de rango fue propuesta por Conover e Iman en su investigación de 1981, y esta se basa en convertir los datos en rangos y en realizar un análisis paramétrico estándar en esos datos transformados

  - Teóricamente se puede aplicar a cualquier análisis en el que hay un contraste paramétrico, pero la suposición de varianzas iguales se mantiene

  - La transformación de rango tiene un buen rendimiento, pero en muchas circunstancias no es robusto y puede tener menos poder que un contraste paramétrico clásico y que métodos no paramétricos, por lo que el consenso en la literatura es que no debería usarse

- En situaciones donde se comparan grupos, es posible aplicar contrastes de aleatorización, en donde la hipótesis nula es que no hay diferencias entre las medias de los grupos y la alternativa es que si las hay

  - Un contraste de permutación permite calcular de manera simple la distribución muestral para cualquier estadístico de contraste bajo la hipótesis nula

  - Este contraste no está limitado a la diferencia de medias, sino que también se pueden usar otro tipo de estadísticos

        - Este tipo de contrastes existen para cualquier estadístico de contraste, independientemente de si la distribución es conocida o no, por lo que siempre se puede escoger el estadístico que discrimine mejor

        - Si la hipótesis nula es cierta, entonces cualquier agrupación aleatoria de observaciones es igualmente probable

        - El posicionamiento o *ranking* del estadístico de contraste real entre los estadísticos de las permutaciones permite obtener un *p-value*. Cuando la hipótesis alternativa tiene solo una cola, entonces la manera de calcular el estadístico cambia de manera acorde

    - Asumiendo que se quieren comparar dos poblaciones con tamaño muestral n_1 para la primera población y n_2 para la segunda población, se pueden seguir los siguientes pasos para realizar un contraste de permutación:

        - Primero se calculan las diferencias entre las medias de ambos grupos, la cual se llama D_0

        - De manera aleatoria, se reasignan las $n_1 + n_2$ observaciones de modo que n_1 observaciones estén en el primer grupo y n_2 estén en el segundo y se calcula la diferencia de medias entre ambos grupos nuevos. Este paso se repite un gran número de veces, de modo que se obtiene D_i para $i = 1,2, \dots, m$ iteraciones

        - Finalmente, se calcula la proporción de todos los D_i que superan o son iguales en valor absoluto a D_0, la cual se interpretará como el *p-value*. Si el contraste es de una cola, entonces no se utiliza el valor absoluto, sino que se usa la magnitud original y se adapta la medición dependiendo de la cola

Los intervalos de confianza

- Los estimadores puntuales daban una única estimación razonable de θ, pero los intervalos de confianza (y más generalmente, los conjuntos de confianza) dan un conjunto de valores $\theta \in C = C(\boldsymbol{y}) \subset \Theta$ que es determinado por los datos observados $\boldsymbol{Y} = \boldsymbol{y}$

    - Una estimación interválica o estimación de intervalo de un parámetro $\theta \in \mathbb{R}$ es cualquier par de funciones $L(y_1, y_2, \dots, y_n)$ y $U(y_1, y_2, \dots, y_n)$ de una muestra que satisface $L(y_1, y_2, \dots, y_n) \leq U(y_1, y_2, \dots, y_n)$ para toda $\boldsymbol{y} \in \mathcal{Y}$. Si $\boldsymbol{Y} = \boldsymbol{y}$ es observada, la inferencia se realiza. El estimador interválico es el intervalo aleatorio $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$

- Se usan las convenciones anteriormente utilizadas para escribir $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ como un estimador interválico para θ basado en la muestra aleatoria $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ y $[L(\boldsymbol{y}), U(\boldsymbol{y})]$ como el valor realizado del intervalo

- Aunque en la mayoría de casos se trabaje con valores finitos para $L(\boldsymbol{y})$ y $U(\boldsymbol{y})$, hay casos en donde es interesante trabajar con intervalos de un solo lado (intervalos de la forma $[L(\boldsymbol{y}), \infty)$ o $(-\infty, U(\boldsymbol{y})]$)

- Aún usando un intervalo cerrado en la definición, hay veces que es más natural utilizar un intervalo abierto $\big(L(\boldsymbol{y}), U(\boldsymbol{y})\big)$ o hasta un intervalo medio-abierto o medio-cerrado, pero la preferencia sigue siendo por el intervalo cerrado

- La estimación interválica permite ganar más seguridad en la estimación, en el sentido que permite que haya alguna garantía de que se captura el parámetro de interés. Para cuantificar esta certidumbre, se utiliza el concepto de probabilidad de cobertura y el nivel de confianza

  - Comparada con la estimación puntual, se pierde precisión en la estimación, pero esta es mucho menos segura, dado que la probabilidad de que el estimador puntual sea igual al valor poblacional del parámetro es infinitesimal

  - Para un estimador interválico $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ de un parámetro θ, la probabilidad de cobertura $P_\theta(\theta \in [L(\boldsymbol{Y}), U(\boldsymbol{Y})])$ es la probabilidad de que el intervalo aleatorio $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ cubra el parámetro verdadero θ. Es decir, es la "probabilidad de cobertura real" de que el parámetro verdadero esté contenido en el intervalo de confianza

  - Para un estimador interválico $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ de un parámetro θ, el coeficiente o nivel de confianza $1-\alpha$ de $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ es la ínfima de la probabilidad de cobertura. Es decir, es la "probabilidad de cobertura nominal" del intervalo (la probabilidad mínima de que

$$1 - \alpha = \inf_\theta P_\theta(\theta \in [L(\boldsymbol{Y}), U(\boldsymbol{Y})])$$

- Hay algunos aspectos importantes que considerar sobre las definiciones anteriores:

  - El intervalo es la cantidad aleatoria, no el parámetro, por lo que cuando se escriben las probabilidades $P_\theta(\theta \in [L(\boldsymbol{Y}), U(\boldsymbol{Y})])$, la probabilidad a la que se refiere es a la de $\boldsymbol{Y}$, no a la de θ (que es una cantidad fija). En otras palabras, la probabilidad $P_\theta(\theta \in$

$[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$) es algebraicamente equivalente a $P_\theta(L(\boldsymbol{Y}) \leq \theta, U(\boldsymbol{Y}) \geq \theta)$ y, por tanto, es una proposición para $\boldsymbol{Y}$

- Como no se sabe el verdadero valor de θ, solo se puede garantizar una probabilidad igual al ínfimo, que está definido como el coeficiente de confianza $1-\alpha$. En muchos casos esto no importa porque la probabilidad de cobertura es una función constante de θ (siempre es la misma independientemente de θ), pero hay veces en donde es una función variable

- La interpretación del intervalo de confianza construido es que el $100(1-\alpha)\%$ de las veces se contendrá el valor real del parámetro

- Los estimadores interválicos junto con el coeficiente de confianza se conocen como intervalos de confianza. Asumiendo que los datos $Y_1, Y_2, \dots, Y_n$ tienen una función de densidad o de masa conjunta $f(\boldsymbol{y}|\theta)$ con espacio paramétrico Θ y espacio muestral $\mathcal{Y}$, y siendo $L_n(\boldsymbol{Y})$ y $U_n(\boldsymbol{Y})$ estadísticos tal que $L_n(\boldsymbol{y}) \leq U_n(\boldsymbol{y})$ para toda $\boldsymbol{y} \in \mathcal{Y}$, entonces $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ es un $100(1-\alpha)\%$ intervalo de confianza para θ si se cumple la siguiente igualdad:

$$P_\theta\big(L_n(\boldsymbol{Y}) < \theta < U_n(\boldsymbol{Y})\big) = 1-\alpha \quad for\ \forall\theta \in \Theta$$

  - El intervalo $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ es un $100(1-\alpha)\%$ IC de muestra grande para θ si se da la siguiente tendencia cuando $n \to \infty$:

$$\lim_{n\to\infty} P_\theta\big(L_n(\boldsymbol{Y}) < \theta < U_n(\boldsymbol{Y})\big) = 1-\alpha \quad for\ \forall\theta \in \Theta$$

  - Un intervalo de confianza también puede ser de un solo lado, de modo que solo $L_n(\boldsymbol{Y})$ o $U_n(\boldsymbol{Y})$ es de interés y el otro extremo se fija como infinito. Además, se puede considerar también un intervalo abierto o semi-abierto o semi-cerrado a conveniencia

  - Aunque lo más importante son los intervalos, hay veces que se trabajan con conjuntos más generales, los cuales se denominan conjuntos de confianza y tienen un coeficiente de confianza asociado $1-\alpha$

- Suponiendo que un IC de nivel de confianza $100(1-\alpha)\%$ para θ tiene un nivel de confianza real de $1-\delta$ (de modo que $P_\theta\big(L(\boldsymbol{Y}) < \theta < U(\boldsymbol{Y})\big) = 1-\delta$ para toda $\theta \in \Theta$, entonces si $1-\delta > 1-\alpha$, el IC es conservador, mientras que si $1-\delta < 1-\alpha$, entonces el IC es liberal

  - Se prefieren los IC conservadores a los IC liberales porque es seguro que estos incluyes las estimaciones incluidas en el liberal, pero no al revés

- Suponiendo que un IC de muestra grande para θ tiene un nivel de confianza real de $1-\delta_n$ donde $\delta_n \to \delta$ cuando $n \to \infty$ para toda $\theta \in \Theta$, entonces si $1-\delta > 1-\alpha$ el IC es asintóticamente conservador, mientras que si $1-\delta < 1-\alpha$, el IC es asintóticamente liberal

- Es posible que $\delta \equiv \delta(\theta)$ dependa de θ y que el IC sea conservador o liberal para diferentes valores de θ (también en el sentido asintótico

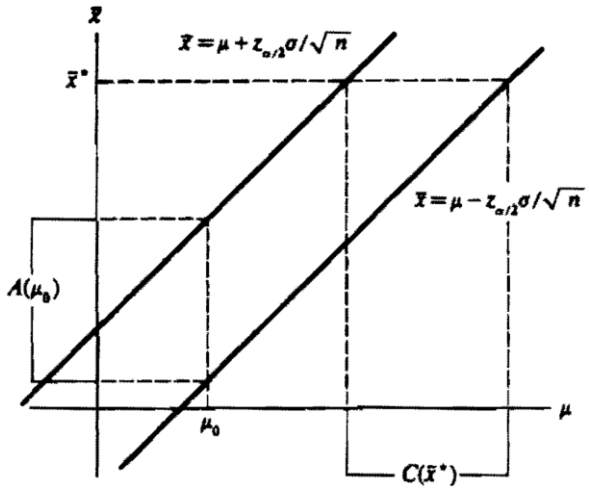
- Existen métodos para poder encontrar estimadores interválicos, los cuales, aunque parezcan diferentes métodos, en realidad son operacionalmente lo mismo: todo se basa en invertir un estadístico de contraste de hipótesis

  - Existe una fuerte correspondencia entre el contraste de hipótesis y la estimación interválica. En general, cualquier conjunto de confianza corresponde a un contraste de hipótesis y viceversa

    - La región de aceptación del contraste de hipótesis $A(\theta_0)$ es el subconjunto del espacio muestral $\mathcal{Y}$ para la cual la hipótesis nula $H_0: \theta = \theta_0$, mientras que el intervalo de confianza $C(\boldsymbol{y})$ es el subconjunto del espacio paramétrico Θ con valores plausibles θ

    - Estos conjuntos están conectados por la siguiente tautología:

$$(y_1, y_2, \dots, y_n) \in A(\theta) \;\Leftrightarrow\; \theta \in C(y_1, y_2, \dots, y_n)$$



  - Tanto los intervalos como los contrastes de hipótesis realizan la misma pregunta, ya que ambos procedimientos buscan consistencia entre los estadísticos muestrales y los parámetros de la población

    - Los contrastes de hipótesis fijan el parámetro y preguntan si los valores muestrales (la región de aceptación) son consistentes

con el valor fijado, mientras que los conjuntos de confianza fijan el valor muestral y preguntan qué valores paramétricos (el intervalo de confianza) hacen esta muestra más plausible

- De este modo, solo hace faltar cambiar el valor θ_0 por θ en la región de aceptación del contraste y aislar θ en la desigualdad doble. De manera conversa, se cambia el valor θ por θ_0 y se aisla el estimador de θ para poder obtener una región de aceptación del contraste

$$A(\mu_0) = \left\{(x_1, \ldots, x_n) : \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

$$C(x_1, \ldots, x_n) = \left\{\mu : \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

- Para toda $\theta_0 \in \Theta$ y para toda $\boldsymbol{y} \in \mathcal{Y}$, siendo $A(\theta_0)$ la región de aceptación para un contraste $H_0: \theta = \theta_0$ y un nivel α y $C(\boldsymbol{y}) \equiv \{\theta_0 : \boldsymbol{y} \in A(\theta_0)\}$ un conjunto en el espacio paramétrico, el conjunto aleatorio $C(\boldsymbol{Y})$ es un conjunto de confianza $1 - \alpha$. De manera conversa, siendo $C(\boldsymbol{Y})$ un conjunto de confianza $1 - \alpha$ y siendo $A(\theta_0) \equiv \{\boldsymbol{y} : \theta_0 \in C(\boldsymbol{y})\}$ un conjunto en el espacio muestral $\boldsymbol{y}$, el conjunto $A(\theta_0)$ es una región de aceptación para un contraste $H_0: \theta = \theta_0$ para un nivel α

  - La demostración de la primera parte de este teorema se basa en que, como $A(\theta)$ es la región de aceptación para un contraste a un nivel α, la probabilidad de rechazo es de α, y, como la probabilidad de que $\theta \in C(\boldsymbol{Y})$ es la misma que $\boldsymbol{Y} \in A(\theta)$ (por la tautología anterior), se puede ver como $C(\boldsymbol{Y})$ es un conjunto de confianza a un nivel $1 - \alpha$

$$P_{\theta_0}\big(\boldsymbol{Y} \notin A(\theta_0)\big) \leq \alpha \Rightarrow P_{\theta_0}\big(\boldsymbol{Y} \in A(\theta_0)\big) \geq 1 - \alpha \quad for\ \forall \theta_0 \in \Theta$$

$$\Rightarrow P_{\theta}\big(\theta \in C(\boldsymbol{Y})\big) = P_{\theta}\big(\boldsymbol{Y} \in A(\theta)\big) \geq 1 - \alpha$$

  - La demostración de la segunda parte se basa en que, como $A(\theta)$ es la región de aceptación para un contraste a un nivel α, la probabilidad de rechazo es de α, y, como la probabilidad de que $\theta \notin C(\boldsymbol{Y})$ es la misma que $\boldsymbol{Y} \notin A(\theta)$ (por la tautología anterior), se puede ver como la región $A(\theta_0)$ pertenece a un contraste con un nivel α

$$P_{\theta_0}\big(\boldsymbol{Y} \notin A(\theta_0)\big) = P_{\theta_0}\big(\theta_0 \notin C(\boldsymbol{Y})\big) \leq \alpha$$

  - Aunque se dice que se invierte un contraste de hipótesis para obtener un conjunto de confianza, en verdad se invierte una

familia de contrastes (uno para cada valor $\theta_0 \in \Theta$ posible) para obtener un conjunto de confianza

- En el teorema anterior solo se especificaba que el contraste de hipótesis debía tener una hipótesis nula $H_0: \theta = \theta_0$, ya que solo se necesita que la región de aceptación sea $P_\theta\big(\boldsymbol{Y} \in A(\theta)\big) \geq 1 - \alpha$. No obstante, en la práctica, también se tiene que tener en mente la hipótesis alternativa H_1, dado que esta determinará la forma razonable de $A(\theta)$ y, por tanto, la de $C(\boldsymbol{Y})$

  - En el teorema, se ha utilizado la palabra conjunto y no intervalo porque no se puede garantizar que el conjunto de confianza obtenido al invertir un contraste de hipótesis sea un intervalo

  - Las propiedades del contraste invertido también se trasladan al conjunto de confianza (su sesgo, por ejemplo) y, debido a que se puede enfocar la atención solo a estadísticos suficientes al buscar un buen contraste de hipótesis, solo es necesario enfocarse en los mismos para conjuntos de confianza

- Un caso importante a estudiar es el de inversión de un contraste de hipótesis de *ratio* de verosimilitud (LRT), ya que funciona de manera diferente a otros

  - La región obtenida invirtiendo el LRT de $H_0: \theta = \theta_0$ contra $H_1: \theta \neq \theta_0$ y la región de confianza resultante tienen la siguiente forma para alguna función $k'(\boldsymbol{y}, \theta)$ que de un nivel de confianza $1 - \alpha$

$$A(\theta_0) = \left\{(y_1, y_2, \dots, y_n) : \frac{L(\theta_0|\boldsymbol{y})}{L\big(\hat{\theta}|\boldsymbol{y}\big)} \leq k(\theta_0)\right\}$$

$$C(\theta) = \{\theta : L(\theta|\boldsymbol{y}) \leq k'(\boldsymbol{y}, \theta)\}$$

  - En algunos casos (tales como la distribución gamma y la distribución normal) la función k' no depende de θ, y en estos casos justamente, la interpretación de la región de aceptación es la de la región conteniendo los valores de θ que maximizan la verosimilitud

- Otro método muy utilizado para la construcción de conjuntos de confianza es el de cantidades pivotales, resultando en lo que se denomina inferencia pivotal (término acuñado por Barnard)

  - Una variable aleatoria $Q(\boldsymbol{Y}, \theta) = Q(Y_1, Y_2, \dots, Y_n, \theta)$ es una cantidad pivotal o pivote si la distribución de $Q(\boldsymbol{Y}, \theta)$ es independiente de todos

los parámetros. Es decir, siendo $\boldsymbol{Y} \sim F(\boldsymbol{Y}|\theta)$, entonces $Q(\boldsymbol{Y}, \theta)$ tiene la misma distribución para cualquier valor de θ

- La función $Q(\boldsymbol{Y}, \theta)$ comúnmente contendrá de manera explícita los parámetros y los estadísticos, pero para cualquier conjunto $\mathcal{A}$, la condición general es que la probabilidad $P_\theta(Q(\boldsymbol{Y}, \theta) \in \mathcal{A})$ no puede depender de θ. La técnica de construir conjuntos de confianza a partir de pivotes se apoya en ser capaz de encontrar un pivote $Q(\boldsymbol{Y}, \theta)$ y un conjunto $\mathcal{A}$ tal que el conjunto $\{\theta: Q(\boldsymbol{Y}, \theta) \in \mathcal{A}\}$ sea una estimación de conjunto de θ

- Algunos ejemplos de funciones de distribución que cumplen estas características son las versiones estándar de distribuciones como la normal, la uniforme u otras o la distribución chi-cuadrada (que depende solo de n)

- Hay veces en las que, para obtener una cantidad pivotal, solo hace falta mirar la forma de la función de densidad de probabilidad y ver si existe un pivote

  - Suponiendo que la función de densidad $f(T|\theta)$ de probabilidad de un estadístico T se puede expresar como $f(T|\theta) = f_X\big(Q(t,\theta)\big)\left|\frac{\partial}{\partial t}Q(t,\theta)\right|$ para alguna función f_X y alguna función monótona Q (en t para cada θ), entonces se puede demostrar que $Q(t,\theta)$ es una cantidad pivotal

$$f(T|\theta) = f_X\big(Q(t,\theta)\big)\left|\frac{\partial}{\partial t}Q(t,\theta)\right|$$

$$\Rightarrow F_T(\theta) = \int f(t|\theta)\, dt = F_X\big(Q(t,\theta)\big)$$

$$\Rightarrow f(T|\theta)\ doesn't\ depend\ on\ parameter\ \theta$$

  - Una vez se tiene el pivote se puede construir un conjunto de confianza a través de definir una región de aceptación para un contraste de hipótesis al α

- Una manera genérica de definir una cantidad pivotal para parámetros $\boldsymbol{\theta}$ reales es a través de la observación de que, si una variable aleatoria continua Y tiene una función de distribución $F(Y,\theta)$ continua y monótona para un parámetro θ y una muestra $\boldsymbol{Y}$, entonces $U_i \equiv F(Y_i,\theta)$ sigue una distribución uniforme $U(0,1)$

  - De darse estas condiciones, entonces $X_i = -\log U_i \sim exp(1)$, de modo que la distribución de esta variable aleatoria nueva no depende de los parámetros de la distribución, y por tanto la

distribución de su suma $\sum_{i=1}^{n} X_i \sim Gamma(n,1)$ no depende tampoco de los parámetros (si la muestra es aleatoria). Finalmente, se puede ver que $2\sum_{i=1}^{n} X_i \sim \chi^2_{2n}$ (si la muestra es aleatoria), lo cual permitirá obtener una cantidad pivotal

- Si Y tiene una función de distribución $F(Y,\theta)$ continua y monótona para un parámetro θ y una muestra $\boldsymbol{Y}$, entonces la siguiente cantidad es siempre un pivote:

$$Q(\boldsymbol{Y},\theta) = -2\sum_{i=1}^{n} \log F(Y_i,\theta) \sim \chi^2_{2n}$$

- También es posible definir la cantidad pivotal a partir de $U_i = 1 - F(Y_i,\theta)$ en el teorema anterior, dado que también se distribuiría $U(0,1)$. De este modo, dependiendo de la función de la distribución, es posible utilizar una u otra definición con tal se simplificar los cálculos u obtener una expresión ideal

$$Q(\boldsymbol{Y},\theta) = -2\sum_{i=1}^{n} \log[1 - F(Y_i,\theta)] \sim \chi^2_{2n}$$

- Combinando el método de la cantidad pivotal con los resultados de la teoría asintótica es posible obtener conjuntos que sean asintóticamente válidos
  - Si $Y_1, Y_2, \dots, Y_n$ son observaciones independientes e idénticamente distribuidas de una distribución $f(\boldsymbol{y}|\theta)$ y $\hat{\theta}$ es el estimador de máxima verosimilitud de θ, entonces se puede usar el siguiente estadístico como cantidad pivotal:

$$\frac{\tau(\hat{\theta}) - \tau(\theta)}{\sqrt{FCRLB_n(\tau(\theta))}} \sim N(0,1)$$

- Para muestras grandes (y cuyas observaciones son independientes e idénticamente distribuidas), la varianza de una función $\tau(\hat{\theta})$ se puede aproximar a través de la cota inferior de Fréchet-Crámer-Rao (debido a que los estimadores de máxima verosimilitud son asintóticamente eficientes)

$$\widehat{Var(\tau(\theta))} = FCRLB_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_n(\theta)} = \frac{[\tau'(\theta)]^2}{nI_1(\theta)}$$

- Por lo tanto, utilizando un estadístico de estandarización, se puede ver como este estimador sigue una distribución normal

estándar, haciendo que se pueda obtener un conjunto de confianza de la siguiente forma:

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

$$\Rightarrow P\left(\tau(\hat{\theta}) - z_{\alpha/2}S < -\tau(\theta) < \tau(\hat{\theta}) + z_{1-\alpha/2}S\right) = 1 - \alpha$$

$$where\ S \equiv \sqrt{FCRLB_n(\tau(\theta))}$$

- Como normalmente $FCRLB_n(\tau(\theta))$ depende de $\tau(\theta)$
- Este intervalo, no obstante, solamente tiene un nivel de confianza aproximado y es un enfoque aceptable solo si la muestra es lo suficientemente grande

- Algunos de los intervalos de confianza obtenidos a partir de cantidades pivotales y que son más utilizados en la práctica son los siguientes:
  - En los casos de distribuciones pertenecientes a familias de localización y escala hay muchas cantidades pivotales para la media μ desconocida. Siendo $Y_1, Y_2, \ldots, Y_n$ una muestra aleatoria de las funciones de densidad de probabilidad indicadas, y siendo $\bar{Y}$ y S^2 los estimadores de la media y la varianza, se puede demostrar que las siguientes cantidades son pivotes:

Form of pdf	Type of pdf	Pivotal quantity
$f(x-\mu)$	Location	$\bar{X}-\mu$
$\frac{1}{\sigma}f(\frac{x}{\sigma})$	Scale	$\frac{\bar{X}}{\sigma}$
$\frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$	Location–scale	$\frac{\bar{X}-\mu}{S}$

    - Como se puede ver, distribuciones pertenecientes a estas familias (la gamma, la exponencial, la Poisson, la normal, etc.) tienen pivotes similares, lo cual simplifica mucho su obtención
    - En particular, si $Y_1, Y_2, \ldots, Y_n$ es una muestra aleatoria de una variable normal $N(\mu, \sigma^2)$ y su desviación típica σ es conocida, entonces se puede estandarizar la variable $\bar{Y}$ para que esta se distribuya como una $z \sim N(0,1)$ y el estadístico $Z = \sqrt{n}(\bar{Y} - \mu)/\sigma$ será una cantidad pivotal

$$P\left(z_{\alpha/2} < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2} \right) = 1 - \alpha$$

$$\Rightarrow P\left(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{Y} - \mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\bar{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- Si, en cambio, $Y_1, Y_2, \ldots, Y_n$ es una muestra aleatoria de una variable normal $N(\mu, \sigma^2)$ y su desviación típica σ es desconocida, entonces no se puede estandarizar la variable $\bar{Y}$, pero si se puede utilizar el estadístico t para que la variable $\bar{Y}$ se distribuya como una $t \sim t_{n-1}$ y el estadístico $t = \sqrt{n}(\bar{Y} - \mu)/\sqrt{S^2}$ será una cantidad pivotal

$$P\left(t_{n-1,\alpha/2} < \frac{\bar{Y} - \mu}{\sqrt{S^2}/\sqrt{n}} < t_{n-1,1-\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(t_{n-1,\alpha/2}\frac{\sqrt{S^2}}{\sqrt{n}} < \bar{Y} - \mu < t_{n-1,1-\alpha/2}\frac{\sqrt{S^2}}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\bar{Y} + t_{n-1,\alpha/2}\frac{\sqrt{S^2}}{\sqrt{n}} < \mu < \bar{Y} + t_{n-1,1-\alpha/2}\frac{\sqrt{S^2}}{\sqrt{n}}\right) = 1-\alpha$$

- En el primer caso, la longitud del intervalo o del conjunto estaría fijada por el margen de error, el cual no es aleatorio porque σ es conocida. No obstante, si σ es desconocida, el uso del estimador S^2 hace que el margen de error sea una variable aleatoria y que, por tanto, la longitud de este conjunto o intervalo sea aleatorio. De media, la longitud del segundo es mayor a la del primer intervalo o conjunto porque los valores críticos para la distribución t-Student son mayores que los de la normal

- En los casos donde $X_1, X_2, \ldots, X_n$ y $Y_1, Y_2, \ldots, Y_n$ sean muestras aleatorias de distribuciones normales $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, se puede demostrar que las siguientes cantidades son pivotes para uno y dos parámetros de varianza (en donde cada varianza proviene de una variable normal):

  - Sabiendo que el estadístico $(n-1)S_n^2/\sigma^2$ sigue una distribución χ^2_{n-1} y que este contiene σ^2 en el denominador (sin depender de ningún otro parámetro más), entonces el estadístico será una cantidad pivotal

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

$$\Rightarrow P\left(\frac{\chi^2_{n-1,\frac{\alpha}{2}}}{(n-1)S_n^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{n-1,1-\frac{\alpha}{2}}}{(n-1)S_n^2}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{(n-1)S_n^2}{\chi^2_{n-1,\frac{\alpha}{2}}} > \sigma^2 > \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{(n-1)S_n^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S_n^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = 1-\alpha$$

- Sabiendo que el estadístico $(S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ sigue una distribución $F_{n-1,m-1}$, entonces el estadístico será una cantidad pivotal para la *ratio* de dos varianzas

$$P\left(F_{n-1,m-1,\alpha/2} < \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} < F_{n-1,m-1,1-\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(F_{n-1,m-1,\alpha/2} < \frac{\sigma_Y^2/\sigma_X^2}{S_Y^2/S_X^2} < F_{n-1,m-1,1-\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{S_Y^2}{S_X^2}F_{n-1,m-1,\frac{\alpha}{2}} < \frac{\sigma_Y^2}{\sigma_X^2} < \frac{S_Y^2}{S_X^2}F_{n-1,m-1,1-\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{S_X^2}{S_Y^2}\frac{1}{F_{n-1,m-1,1-\alpha/2}} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{S_X^2}{S_Y^2}\frac{1}{F_{n-1,m-1,\alpha/2}}\right) = 1-\alpha$$

- En los casos donde $X_1, X_2, \dots, X_n$ y $Y_1, Y_2, \dots, Y_n$ sean muestras aleatorias de distribuciones normales $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, se puede demostrar que las siguientes cantidades son pivotes para la diferencia de medias

Difference between two means (Unknown equal vars)	$(\bar{x}_A - \bar{x}_B) \pm t_{n_A+n_B-2,\alpha/2} s \sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$ where $s = \sqrt{\frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2}}$
Difference between two means (Unknown but different vars)	$(\bar{x}_A - \bar{x}_B) \pm t_{g,\alpha/2}\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}$ with $g = \frac{[s_A^2/n_A+s_B^2/n_B]}{(s_A^2/n_A)^2((n_A-1)+(s_B^2/n_B)^2((n_B-1)}$
Difference between two means (Unknown but different vars)	$(\bar{x}_A - \bar{x}_B) \pm z_{\alpha/2}\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}$ with large sample size n_A and n_B

- Debido a que se miran las diferencias entre medias y ambas variables aleatorias son independientes, se puede ver como la diferencia sigue una distribución $N\left(\mu_{\bar{X}-\bar{Y}}, \sigma^2_{\bar{X}-\bar{Y}}\right)$. Los parámetros se pueden obtener aplicando la esperanza y la varianza:

$$\mu_{\bar{X}-\bar{Y}} = E(\bar{X}-\bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$

$$\sigma^2_{\bar{X}-\bar{Y}} = Var(\bar{X}-\bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

- En el caso de que las varianzas sean desconocidas pero iguales, se puede utilizar el estimador $S^2_{\bar{X}-\bar{Y}}$ para la varianza de la diferencia de medias, ajustándose para los grados de libertad. De este modo, se puede ver que $\frac{\bar{X}-\bar{Y}-\mu_{\bar{X}-\bar{Y}}}{\sqrt{S^2_{\bar{X}-\bar{Y}}/(n_X+n_Y)}}$ es una cantidad pivotal que sigue una distribución t-Student con $n_x + n_Y - 2$ grados de libertad (al ser la suma de dos distribuciones t-Student independientes con misma varianza):

$$P\left(t_{n^*,\alpha/2} < \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{S^2_{\bar{X}-\bar{Y}}\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} < t_{n^*,1-\alpha/2} \right) = 1-\alpha$$

$$S^2_{\bar{X}-\bar{Y}} \equiv \frac{(n_X-1)S_X^2 + (n_X-1)S_Y^2}{n^*} \quad where\ \ n^* = n_X + n_Y - 2$$

$$\Rightarrow\ P\left(\bar{X}-\bar{Y}+t_{n^*,\frac{\alpha}{2}}S < \mu_X-\mu_Y\ < \bar{X}-\bar{Y}+t_{n^*,1-\frac{\alpha}{2}}S\right) = 1-\alpha$$

$$where\ \ S \equiv \sqrt{S^2_{\bar{X}-\bar{Y}}\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}$$

- En el caso de que las varianzas sean desconocidas pero diferentes, ya no se puede utilizar el estimador $S^2_{\bar{X}-\bar{Y}}$ para la varianza de la diferencia de medias, sino que se tiene que hacer un ajuste para los grados de libertad de la distribución t-Student y aplicar la varianza de la diferencia:

$$P\left(t_{g,\alpha/2} < \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{(S_X^2/n_X)+(S_Y^2/n_Y)}} < t_{g,1-\alpha/2} \right) = 1-\alpha$$

$$\Rightarrow\ P\left(\bar{X}-\bar{Y}+t_{g,\frac{\alpha}{2}}S < \mu_X-\mu_Y\ < \bar{X}-\bar{Y}+t_{g,1-\frac{\alpha}{2}}S\right) = 1-\alpha$$

$$where\ \ g = \frac{\left(\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}\right)}{\left(\frac{S_X^2}{n_X}\right)^2(n_X-1)+\left(\frac{S_Y^2}{n_Y}\right)^2(n_Y-1)}\ \ \&\ S \equiv \sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}$$

- Si en este último caso se considera que la muestra es grande, entonces la distribución asintótica que sigue el estadístico es normal y se puede usar la siguiente cantidad pivotal

$$P\left(z_{\alpha/2} < \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{(S_X^2/n_X) + (S_Y^2/n_Y)}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\bar{X} - \bar{Y} + z_{\alpha/2}S < \mu_X - \mu_Y < \bar{X} - \bar{Y} + z_{1-\alpha/2}S\right) = 1 - \alpha$$

$$where\ S \equiv \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$$

- Otro caso importante proviene de que $X_1, X_2, \dots, X_n$ y $Y_1, Y_2, \dots, Y_n$ sean muestras aleatorias de distribuciones $Binom(n_X, p_X)$ y $Binom(n_Y, p_Y)$. Debido a que las medias seguirían una $N\left(p, \frac{p(1-p)}{n}\right)$ para $n \to \infty$ (gracias al teorema del límite central), es posible utilizar una cantidad pivotal que se distribuya de esta manera para crear intervalos de confianza

One proportion	$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Two proportions	$(\hat{p}_A - \hat{p}_B) \pm z_{\alpha/2}\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$ where $\hat{p}_0 = \frac{n_A\hat{p}_A + n_B\hat{p}_B}{n_A + n_B}$

  - Si se quiere crear un intervalo para una proporción desconocida, se puede utilizar un procedimiento similar al caso de varianza conocida para la normal y utilizar un estadístico de estandarización como cantidad pivotal

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

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

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

  - Si se quiere crear un intervalo para la diferencia de proporciones desconocidas, se puede utilizar un procedimiento similar al caso de varianzas iguales y utilizar un estadístico de estandarización como cantidad pivotal

$$P\left(z_{\alpha/2} < \frac{\hat{p}_X - \hat{p}_Y - (p_X - p_Y)}{\sqrt{\left(\frac{\hat{p}_X(1-\hat{p}_X)}{n_X}\right) + \left(\frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}\right)}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(z_{\alpha/2} < \frac{\hat{p}_X - \hat{p}_Y - (p_X - p_Y)}{\sqrt{p(1-p)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

$$where\ \ p \equiv \frac{n_X\hat{p}_X + n_Y\hat{p}_Y}{n_X + n_Y}$$

$$\Rightarrow P\left(\hat{p}_X - \hat{p}_Y + z_{\alpha/2}S < \mu_X - \mu_Y < \hat{p}_X - \hat{p}_Y + z_{1-\alpha/2}S\right) = 1 - \alpha$$

$$where\ \ S \equiv \sqrt{p(1-p)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}$$

- Ahora que se han desarrollado métodos para la obtención de estimadores interválicos, es necesario tener métodos para evaluarlos. En la estimación de conjuntos, dos cantidades son los principales indicadores de la calidad: el tamaño y las probabilidades de cobertura

    - Antes de poder optimizar cualquier conjunto de confianza con respecto a su tamaño y a las probabilidades de cobertura, es necesario decidir como se miden estas cantidades

        - Las probabilidades de cobertura de un conjunto serán, excepto en casos especiales, función del parámetro, por lo que no hay un solo valor que considerar, sino un número infinito de valores

        - Para la mayor parte del desarrollo, sin embargo, se mide el rendimiento de la probabilidad de cobertura a través del nivel de confianza, que es el ínfimo de estas probabilidades. Aunque esta es una forma de resumir las probabilidades de cobertura, no es la única (se podrían calcular las probabilidades medias, por ejemplo)

        - Cuando se habla del tamaño de un conjunto de confianza, normalmente se refiere a su longitud si es un intervalo. Si el conjunto no es un intervalo, sino uno multidimensional, entonces se usa el volumen

    - Para poder seleccionar el intervalo de confianza con la menor longitud para un nivel de confianza definido, es posible utilizar el siguiente teorema, basado en la condición de unimodalidad:

- Una función de densidad de probabilidad $f(y)$ es unimodal si existe una y^* tal que $f(y)$ no es decreciente para $y \leq y^*$ y $f(y)$ no es creciente para $y \geq y^*$

- Siendo $f(y)$ una función de densidad unimodal, si el intervalo $[a, b]$ satisface las condiciones $\int_a^b f(y)\, dy = 1 - \alpha$, $f(a) = f(b) > 0$ y $a \leq y^* \leq b$ donde y^* es la moda de $f(y)$, entonces $[a, b]$ es el intervalo más corto entre todos los intervalos que satisface la primera condición

- El teorema anterior, no obstante, no suele funcionar directamente al lidiar con distribuciones que no sean de la familia de localización, por lo que se tiene que ser más cuidadoso

  - En particular, al trabajar con distribuciones de la familia de escala, el teorema no es directamente aplicable, dado que la longitud del intervalo no es proporcional a la distancia $b - a$. Como el teorema anterior utiliza b como una función de a, se puede utilizar un problema de minimización restringido, en donde el factor al que y es proporcional se minimice cumpliendo la restricción del nivel de confianza:

$$Y \sim Gamma(1, \beta) \Rightarrow \min_a \frac{1}{a} - \frac{1}{b(a)} \;\; s.t. \;\; \int_a^b f_Y(y)\, dy = 1 - \alpha$$

- Otro aspecto importante para la evaluación de los intervalos y los conjuntos de confianza es el tamaño de la muestra. El tamaño muestral depende de que tan precisa se quiere la estimación, del nivel de confianza, del valor real de la desviación estándar y del tamaño poblacional

  - Mientras más precisión y nivel de confianza se desee para el conjunto, mayor deberá ser el tamaño muestral

  - Para poder obtener el tamaño muestral deseado, es necesario seguir el siguiente procedimiento: primero, se escoge un nivel de confianza determinado; segundo, se decide el margen de error de la estimación (la longitud del intervalo); después, se utiliza una estimación de la desviación estándar o una proporción de $p = 50\%$ (en caso de usar proporciones) y se aísla n en la ecuación o inecuación

- Siendo (L, U) un IC de nivel de confianza $100(1 - \alpha)\%$ (o un IC de muestra grande) para θ, A_α es la longitud asintótica del IC si se cumple la siguiente condición:

$$n^\delta (U_n - L_n) \xrightarrow{P} A_\alpha \quad for \;\; \forall \theta \in \Theta$$

- Para una δ y un nivel de confianza $1-\alpha$ dado, un IC con una A_α más pequeña es preferido que un IC con una más grande

- Si $A_{1,\alpha}$ y $A_{2,\alpha}$ son dos longitudes asintóticas con la misma δ, entonces la *ratio* entre ambas es una medida de eficiencia asintótica relativa para intervalos de confianza

$$\left(\frac{A_{2,\alpha}}{A_{1,\alpha}}\right)^{\frac{1}{\delta}}$$

La teoría asintótica

- La teoría asintótica se usa para aproximar la distribución de un estimador cuando el tamaño de la muestra n es grande, y es extremadamente útil si la distribución exacta de la muestra del estimador es complicada o desconocida. Primero es necesario estudiar los modos de convergencia y la consistencia de los estimadores

  - Siendo $\{Z_n\}_{n=1}^{\infty}$ para $n \in \mathbb{N}$ una secuencia de variables aleatorias con función de distribución de probabilidad F_n y X una variable aleatoria con función de distribución de probabilidad F, Z_n converge en distribución o converge en ley a X si $F_n(t) \to F(t)$ cuando $n \to \infty$ para cada punto de continuidad t de F

$$\lim_{n\to\infty} F_n(t) = F(t) \;\; at \; \forall t \in F \;\Rightarrow\; Z_n \xrightarrow{D} X$$

    - La distribución de X se llama distribución en el límite o distribución asintótica de Z_n. Esta distribución no depende del tamaño de la muestra n

    - La convergencia en distribución expresa que si la función de distribución $F_n(t)$ de X_n se acerca a la función de distribución $F(t)$ de X cuando $n \to \infty$ dado que t es un punto de continuidad de F. Por lo tanto, para toda $\varepsilon > 0$ existe N_t (que depende del valor de t) tal que si $n > N_t$ entonces $|F_n(t) - F(t)| < \varepsilon$

$$\forall \varepsilon > 0, \exists N_t \; such \; that \; if \; n > N_t, then \; |F_n(t) - F(t)| < \varepsilon$$

    - La convergencia en distribución es útil porque si la distribución de una variable X_n es desconocida o complicada y la distribución de X es fácil de usa, entonces para un tamaño grande n se puede aproximar la probabilidad de que X_n esté en un intervalo con la probabilidad de que X esté en ese intervalo y se pueden crear contrastes de hipótesis con esto

$$X_n \xrightarrow{D} X \;\Rightarrow\; P(a < X_n \leq b) \xrightarrow{D} P(a < X \leq b)$$

$$\Rightarrow\; F_n(b) - F_n(a) \xrightarrow{D} F(b) - F(a)$$

- La convergencia en distribución, por lo tanto, no implica que las variables $X_n \equiv X_n(\omega)$ converge a la variable aleatoria $X \equiv X(\omega)$ para toda ω

- Una secuencia de variables aleatorias X_n converge en distribución a una constante $\tau(\theta)$ si X_n converge en distribución a X en donde $P\big(X = \tau(\theta)\big) = 1$. La distribución de la variable aleatoria X es degenerada en $\tau(\theta)$ o es un punto de masa en $\tau(\theta)$

$$X_n \xrightarrow{D} \tau(\theta) \quad if \quad X_n \xrightarrow{D} X \;\; where \;\; P\big(X = \tau(\theta)\big) = 1$$

- Una secuencia de variables aleatorias X_n converge en probabilidad a una variable aleatoria X denotado por $X_n \xrightarrow{P} X$, si para toda $\varepsilon > 0$ se cumple lo siguiente:

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1 \quad or \quad \lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

  - La convergencia en probabilidad expresa que, en el límite, la variable aleatoria X_n que depende del tamaño de la muestra n se va acercando cada vez más a X cuanto mayor es el tamaño

  - Por lo tanto, que una secuencia X_n converja en probabilidad a X quiere decir $X_n - X \xrightarrow{P} 0$ cuando $n \to \infty$

  - Una secuencia de variables aleatorias X_n converge en probabilidad a una constante $\tau(\theta)$, denotado por $X_n \xrightarrow{P} \tau(\theta)$, si para toda $\varepsilon > 0$ se cumple lo siguiente:

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \varepsilon) = 1 \quad or \quad \lim_{n\to\infty} P(|X_n - \tau(\theta)| \geq \varepsilon) = 0$$

- Una secuencia de estimadores T_n de $\tau(\theta)$ es consistente para $\tau(\theta)$ si $T_n \xrightarrow{P} \tau(\theta)$ para toda $\theta \in \Theta$. Por lo tanto, se tienen que cumplir las siguientes condiciones:

$$\lim_{n\to\infty} P(|T_n - \tau(\theta)| < \varepsilon) = 1 \quad or \quad \lim_{n\to\infty} P(|T_n - \tau(\theta)| \geq \varepsilon) = 0$$

  - Si T_n es consiste para $\tau(\theta)$, entonces T_n es un estimador consistente para $\tau(\theta)$

- La probabilidad $P \equiv P_\theta$ es la distribución de probabilidad "real" o la probabilidad subyacente que depende de θ

- La consistencia es una propiedad débil que normalmente se satisface para buenos estimadores, y esta quiere decir que la probabilidad de que el valor de T_n caiga en la cercanía de $\tau(\theta)$ tiende a 1 sin importar el valor de θ

  - Para un número real $r > 0$, Y_n converge en media r-ésima a una variable aleatoria Y, denotado por $Y_n \xrightarrow{r} Y$, si $E(|Y_n - Y|^r) \to 0$ cuando $n \to \infty$

$$Y_n \xrightarrow{r} Y \quad if \lim_{n\to\infty} E(|Y_n - Y|^r) = 0$$

  - En particular, si $r = 2$, entonces Y_n converge en media cuadrática a Y, denotado por $Y_n \xrightarrow{2} Y$ si $E[(Y_n - Y)^2] \to 0$ cuando $n \to \infty$

$$Y_n \xrightarrow{2} Y \quad if \lim_{n\to\infty} E[(Y_n - Y)^2] = 0$$

- La desigualdad de Chebyshev generalizada permite estudiar la convergencia en probabilidad de manera más sencilla y, además, permite derivar desigualdades complementarias, tales como la desigualdad de Chebyshev y la de Markov

  - La desigualdad de Chebyshev generalizada expresa que, siendo $u: \mathbb{R} \to [0, \infty)$ una función no negativa, si $E[u(Y)]$ existe, entonces para cualquier $c < 0$ se cumple la siguiente desigualdad

$$P(u(Y) \geq c) \leq \frac{E[u(Y)]}{c}$$

  - Si $\mu = E(Y)$ existe, tomando $u(y) \equiv |y - \mu|^r$ y $\tilde{c} \equiv c^r$ se obtiene la desigualdad de Markov. Esta expresa que, para cualquier $r > 0$ y $c > 0$, se obtiene la siguiente desigualdad:

$$P(|Y - \mu|^r \geq c^r) \leq \frac{E[|Y - \mu|^r]}{c^r}$$

  - Si $r = 2$ y $Var(Y)$ es constante y existe, entonces se puede obtener la desigualdad de Chebyshev. Esta expresa que, para $r = 2$, se obtiene la siguiente desigualdad:

$$P((Y - \mu)^2 \geq c^2) = P(|Y - \mu| \geq c) \leq \frac{E[(Y - \mu)^2]}{c^2} = \frac{Var(Y)}{c^2}$$

- Como estas son consecuencia de la desigualdad de Chebyshev generalizada, solo hace falta demostrar esta para demostrar las otras dos. La demostración se basa en la definición de esperanza y de función de distribución

$$E[u(Y)] = \int_{\mathbb{R}} u(y)f(y)\,dy =$$

$$= \int_{\{y:u(y)\geq c\}} u(y)f(y)\,dy + \int_{\{y:u(y)<c\}} u(y)f(y)\,dy \geq$$

$$\geq \int_{\{y:u(y)\geq c\}} u(y)f(y)\,dy$$

$$u(y)f(y) \geq 0 \;\Rightarrow\; E[u(Y)] \geq c\int_{\{y:u(y)\geq c\}} f(y)\,dy = cP(u(Y)\geq c)$$

- A partir de las desigualdades, las siguientes proposiciones dan condiciones suficientes para que T_n sea un estimador consistente para $\tau(\theta)$:

  - Si $MSE_{\tau(\theta)}(T_n) \to 0$ cuando $n \to \infty$ para toda $\theta \in \Theta$, entonces T_n es un estimador consistente de $\tau(\theta)$. Esta proposición se puede demostrar a través de demostrar que el estimador tiende en probabilidad a $\tau(\theta)$

$$P(|T_n - \tau(\theta)| \geq c) \leq \frac{E\left[\left(T_n - \tau(\theta)\right)^2\right]}{c^2}$$

$$\Rightarrow\; T_n \xrightarrow{P} \tau(\theta) \;\; if \;\; \lim_{n\to\infty} P(|T_n - \tau(\theta)| \geq c) = 0$$

$$\Rightarrow\; T_n \xrightarrow{P} \tau(\theta) \;\; if \;\; \frac{1}{c^2}\lim_{n\to\infty} E\left[\left(T_n - \tau(\theta)\right)^2\right] = 0$$

$$\Rightarrow\; T_n \xrightarrow{P} \tau(\theta) \;\; if \;\; \lim_{n\to\infty} E\left[\left(T_n - \tau(\theta)\right)^2\right] = 0$$

$$\Rightarrow\; T_n \xrightarrow{P} \tau(\theta) \;\; if \;\; \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0$$

  - Si $Var_\theta(T_n) \to 0$ y $E_\theta(T_n) \to \tau(\theta)$ cuando $n \to \infty$ para toda $\theta \in \Theta$, entonces T_n es un estimador consistente de $\tau(\theta)$. Esta proposición se puede demostrar a través de la definición del MSE y de su tendencia en probabilidad a cero

$$MSE_{\tau(\theta)}(T_n) = Var_\theta(T_n) + \left[Bias_{\tau(\theta)}(T_n)\right]^2$$

$$\Rightarrow T_n \xrightarrow{P} \tau(\theta) \quad if \quad \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0 \quad (last\ result)$$

$$\Rightarrow T_n \xrightarrow{P} \tau(\theta) \quad if \quad \lim_{n\to\infty} Var_\theta(T_n) = 0 \ \& \ \lim_{n\to\infty} Bias_{\tau(\theta)}(T_n) = 0$$

- Las siguientes dos proposiciones muestran como los estimadores que convergen a una velocidad $\sqrt{n}$ son consistentes y que, junto al método delta permite demostrar que $g(T_n)$ es un estimador consistente de $g(\theta)$:

  - Siendo $X_\theta \sim N\big(0, v(\theta)\big)$ y $0 < \delta \leq 1$, si $n^\delta\big(T_n - \tau(\theta)\big) \xrightarrow{D} X_\theta$ para toda $\theta \in \Theta$, entonces $T_n \xrightarrow{P} \tau(\theta)$

  - En consecuencia, si $\sqrt{n}\big(T_n - \tau(\theta)\big) \xrightarrow{D} N\big(0, v(\theta)\big)$ para toda $\theta \in \Theta$, entonces $T_n \xrightarrow{P} \tau(\theta)$ y T_n es un estimador consistente para $\tau(\theta)$

- Siendo T_n una secuencia de estimadores consistentes de un parámetro $\tau(\theta)$ y siendo $a_1, a_2, a_3, \ldots$ y $b_1, b_2, b_3, \ldots$ secuencias de constantes que satisfacen que $a_n \to 1$ y $b_n \to 0$ cuando $n \to \infty$, entonces la secuencia $U_n = a_n T_n + b_n$ es una secuencia consistente de estimadores de θ

  - Esto se puede demostrar a través de tomar el límite de las secuencias a la vez:

$$\lim_{n\to\infty} U_n = \lim_{n\to\infty} a_n T_n + b_n = \lim_{n\to\infty} a_n \lim_{n\to\infty} T_n + \lim_{n\to\infty} b_n =$$

$$= \lim_{n\to\infty} T_n = \tau(\theta)$$

- Finalmente, la noción de convergencia casi en todas partes o *almost everywhere* permite definir la ley fuerte y la ley débil de números grandes, y la demostración de estos teoremas utiliza las desigualdades anteriores

  - Una secuencia de variables aleatorias X_n converge en casi todas partes o con probabilidad 1 a X, denotado por $X_n \xrightarrow{ae} X$, si se cumple la siguiente condición:

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

- La ley fuerte de grandes números expresa que $\bar{Y}_n \xrightarrow{ae} \mu$ y la ley débil de grandes números expresa que $\bar{Y}_n \xrightarrow{P} \mu$

- La demostración de la ley débil de números grandes se puede construir a partir de la desigualdad de Chebyshev y asumiendo que la varianza de Y_i sea constante (de modo que $Var(Y_i) = \sigma^2$)

$$P(|\bar{Y}_n - \mu| \geq c) \leq \frac{Var(\bar{Y}_n)}{c^2} \;\Rightarrow\; P(|\bar{Y}_n - \mu| \geq c) \leq \frac{\sigma^2}{nc^2}$$

$$\Rightarrow \bar{Y}_n \xrightarrow{P} \mu \quad if \quad \lim_{n\to\infty} P(|\bar{Y}_n - \mu| \geq c) = 0$$

$$\Rightarrow\; \bar{Y}_n \xrightarrow{P} \mu \quad if \quad \lim_{n\to\infty} \frac{\sigma^2}{nc^2} = 0 \;\Rightarrow\; \bar{Y}_n \xrightarrow{P} \mu$$

- A partir de las maneras de converger de las variables aleatorias y de la propiedad de consistencia, es posible establece resultados muy importantes: el teorema central del límite y el método delta

  - Siendo $Y_1, Y_2, \ldots, Y_n$ observaciones i.i.d. con $E(Y) = \mu$ y $Var(Y) = \sigma^2$ y siendo $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ la media muestral, entonces la normalización de la media muestral sigue una distribución normal estándar cuando $n \to \infty$

$$\sqrt{n}(\bar{Y}_n - \mu) = \sqrt{n}\left(\sum_{i=1}^{n} Y_i - n\mu\right) \xrightarrow{D} N(0, \sigma^2)$$

$$\sqrt{n}\left(\frac{\bar{Y}_n - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{\sum_{i=1}^{n} Y_i - n\mu}{n\sigma}\right) \xrightarrow{D} N(0,1)$$

    - Suponiendo que $Y_n \xrightarrow{D} F(\boldsymbol{\theta})$, entonces $Y_n \sim AF(\boldsymbol{\theta})$ o $Y_n \approx F(\boldsymbol{\theta})$ significa que se aproxima la función de distribución por la función de distribución F

    - La distribución aproximada de $\bar{Y}_n$ sí depende de n (a diferencia de la distribución asintótica) y esta es la siguiente:

$$\bar{Y}_n \approx N(\mu, \sigma^2/n)$$

  - Las dos aplicaciones más importantes del teorema central del límite son dar la distribución asintótica de $\sqrt{n}(\bar{Y}_n - \mu)$ y la distribución asintótica de $\sqrt{n}(Y_n/n - \mu_X)$ para una variable aleatoria $Y_n \equiv \sum_{i=1}^{n} X_i$ tal que X_i son i.i.d. con $E(X) = \mu_X$ y $Var(X) = \sigma_X^2$

$$\sqrt{n}\left(\frac{Y_n/n - \mu_X}{\sigma_X}\right) = \sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - n\mu_X}{n\sigma_X}\right) \xrightarrow{D} N(0,1)$$

- Si se quiere encontrar la distribución asintótica de una variable Y_n sola, se pueden utilizar las propiedades de la distribución de la variable Y_n y obtener así la esperanza y la varianza de esta, las cuales serán la media y la varianza de la distribución asintótica normal

$$Y_n \xrightarrow{D} N\big(E(Y_n), Var(Y_n)\big)$$

- Muchas variables aleatorias vistas anteriormente se pueden aproximar de esta manera. Además, el teorema expresa que $\bar{Y}_n \approx N(\mu, \sigma^2/n)$ y que $Y_n/n \approx N(\mu_X, \sigma_X^2/n)$

- El método delta permite encontrar la distribución asintótica de una variable T_n. Si g no depende de n, $g'(\theta) \neq 0$ y $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)$, entonces se cumple la siguiente propiedad:

$$\sqrt{n}\big(g(T_n) - g(\theta)\big) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2)$$

- Si $\sqrt{n}(T_n - k) \xrightarrow{D} N(0, \sigma^2)$, entonces se evalúa la derivada en k, por lo que la derivada a utilizar en el método delta es $g'(k)$, aunque k sea una función simple

- La demostración de este teorema se basa en la expansión de Taylor de $g(T_n)$ con respecto a $T_n = \theta$ y en el teorema de Slutsky

$$g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + resid.$$

$$if\ \ residual \to 0\ \ when\ \ T_n \to \theta\ \ as\ \ n \to \infty, then:$$

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)\ \ \Rightarrow\ \ T_n \xrightarrow{P} \theta\ \ \Rightarrow\ resid. \xrightarrow{P} 0$$

$$\Rightarrow \sqrt{n}[g(T_n) - g(\theta)] = g'(\theta)\sqrt{n}(T_n - \theta)$$

$$\Rightarrow\ g'(\theta)\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2)$$

$$as\ \ E\big[\sqrt{n}g'(\theta)(T_n - \theta)\big] = 0$$

$$\&\ \ Var\big[\sqrt{n}g'(\theta)(T_n - \theta)\big] = [g'(\theta)]^2 Var\big[\sqrt{n}(T_n - \theta)\big] = [g'(\theta)]^2\sigma^2$$

$$\Rightarrow \sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2)$$

- El método delta permite encontrar la distribución asintótica de una variable T_n. Si g no depende de n, $g'(\theta) \neq 0$ y $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)$, entonces se cumple la siguiente propiedad:

- A partir del teorema del límite central, la teoría asintótica ofrece métodos para evaluar la eficiencia de los estimadores

  - Siendo $Y_1, Y_2, \dots, Y_n$ variables aleatorias i.i.d. y $T_n \equiv T_n(Y_1, Y_2, \dots, Y_n)$ un estimador del parámetro μ_T tal que $\sqrt{n}(T_n - \mu_T) \xrightarrow{D} N(0, \sigma_A^2)$, entonces la varianza asintótica de $\sqrt{n}(T_n - \mu_T)$ es σ_A^2 y la varianza asintótica de T_n es σ_A^2/n, denotada por $AV(T_n)$

    - Si S_n^2 es un estimador consistente de σ_A^2, entonces el error estándar asintótico de T_n es $S_n/\sqrt{n}$, denotada por $SE(T_n)$

    - Si $Y_1, Y_2, \dots, Y_n$ son variables aleatorias i.i.d. con función de distribución F, entonces $\sigma_A^2 \equiv \sigma_A^2(F)$ depende de F

    - En este caso σ_A^2 es una función del estimador T_n y de la función de distribución subyacente F de Y_i

  - Siendo $T_{1,n}$ y $T_{2,n}$ dos estimadores de un parámetro θ tal que se cumple $n^{\delta}(T_{1,n} - \theta) \xrightarrow{D} N(0, \sigma_1^2)$ y $n^{\delta}(T_{2,n} - \theta) \xrightarrow{D} N(0, \sigma_2^2)$, entonces la eficiencia relativa asintótica (ARE) de $T_{1,n}$ con respecto a $T_{2,n}$ es la siguiente *ratio*:

$$ARE(T_{1,n}, T_{2,n}) = \frac{\sigma_2^2(F)}{\sigma_1^2(F)}$$

    - Los estimadores deben tener la misma tasa de convergencia n^{δ}, dado que si un estimador tiene una $\delta_i > \delta_j$ entonces se juzgaría como un estimador mejor

    - Los dos estimadores deben de estimar el mismo parámetro θ, lo cual no siempre pasa a no ser que la distribución sea simétrica respecto a μ

    - El estimador con menor varianza se considerará un mejor estimador en términos relativos asintóticos

  - Asumiendo que $\tau'(\theta) \neq 0$, el estimador T_n para $\tau(\theta)$ es asintóticamente eficiente si se cumple la siguiente condición:

$$\sqrt{n}\big(T_n - \tau(\theta)\big) \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{I_n(\theta)}\right) \sim N\left(0, FCRLB_n\big(\tau(\theta)\big)\right)$$

- En particular, el estimador el estimador T_n para θ es asintóticamente eficiente si se cumple la siguiente condición:

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N\left(0, \frac{1}{I_n(\theta)}\right) = N\big(0, FCRLB_n(\theta)\big)$$

- Si $T_{2,n}$ es un estimador asintóticamente eficiente de θ, $I_1(\theta)$ y $v(\theta)$ son funciones continuas y $T_{1,n}$ es un estimador tal que se cumple que $\sqrt{n}\big(T_{1,n} - \theta\big) \xrightarrow{D} N\big(0, v(\theta)\big)$, entonces bajo condiciones de regularidad, $v(\theta) \geq 1/I_n(\theta)$ y eso hace que el estimador $T_{2,n}$ sea mejor que cualquier otro estimador $T_{1,n}$

$$ARE\big(T_{1,n}, T_{2,n}\big) = \frac{1/I_n(\theta)}{v(\theta)} = \frac{1}{I_n(\theta)v(\theta)} \leq 1$$

- Siendo $Y_1, Y_2, \dots, Y_n$ son observaciones independientes e idénticamente distribuidas de una función de densidad $f(\boldsymbol{y}|\theta)$, $\hat{\theta}_n$ el MLE o UMVUE para θ y $\tau(\theta)$ una función continua de θ tal que $\tau'(\theta) \neq 0$, bajo ciertas condiciones de regularidad, el estimador es asintóticamente eficiente

$$\sqrt{n}\left(\tau\big(\hat{\theta}_n\big) - \tau(\theta)\right) \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{nI_1(\theta)}\right) = N(0, FCRLB_n[\tau(\theta)])$$

- Uno de los teoremas más importantes en la teoría asintótica es el teorema de Slutsky, dado que permite obtener propiedades sobre la tendencia en distribución y en probabilidad de las variables y permite derivar otros resultados

  - Suponiendo que $Y_n \xrightarrow{D} Y$ y que $W_n \xrightarrow{D} w$ para alguna constante w, entonces el teorema de Slutsky expresa que se cumplen las siguientes propiedades:

$$\begin{aligned} &Y_n + W_n \xrightarrow{D} Y + w, \\ &Y_n W_n \xrightarrow{D} wY, \text{ and} \\ &Y_n / W_n \xrightarrow{D} Y/w \text{ if } w \neq 0. \end{aligned}$$

  - Para una secuencia de variables aleatorias X_n, una variable aleatoria X y un parámetro θ, el teorema de Slutsky permite demostrar que se cumplen las siguientes propiedades:

$$\text{If } X_n \xrightarrow{P} X \text{ then } X_n \xrightarrow{D} X.$$
$$\text{If } X_n \xrightarrow{\text{ae}} X \text{ then } X_n \xrightarrow{P} X \text{ and } X_n \xrightarrow{D} X.$$
$$\text{If } X_n \xrightarrow{r} X \text{ then } X_n \xrightarrow{P} X \text{ and } X_n \xrightarrow{D} X.$$
$$X_n \xrightarrow{P} \tau(\theta) \text{ iff } X_n \xrightarrow{D} \tau(\theta).$$
$$\text{If } X_n \xrightarrow{P} \theta \text{ and } \tau \text{ is continuous at } \theta, \text{ then } \tau(X_n) \xrightarrow{P} \tau(\theta).$$
$$\text{If } X_n \xrightarrow{D} \theta \text{ and } \tau \text{ is continuous at } \theta, \text{ then } \tau(X_n) \xrightarrow{D} \tau(\theta).$$

- Estas propiedades muestran que, si para toda $\theta \in \Theta$ se cumple que $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$ o que $T_n \xrightarrow{ae} \tau(\theta)$, entonces $T_n \xrightarrow{P} \tau(\theta)$ y eso hace que el estimador T_n sea un estimador consistente para $\tau(\theta)$

- Asumiendo que la función g no depende de n, se puede demostrar el teorema generalizado de mapeado continuo y el teorema de mapeado continuo

  - Si $X_n \xrightarrow{D} X$ y la función g es tal que $P\big(X \in C(g)\big) = 1$, donde $C(g)$ es el conjunto de puntos en donde g es continua, entonces $g(X_n) \xrightarrow{D} g(X)$

  - Si $X_n \xrightarrow{D} X$ y la función g es continua, entonces $g(X_n) \xrightarrow{D} g(X)$

  - Estos teoremas expresan que la convergencia en distribución se preserva por las funciones continuas, y que hasta algunas discontinuidades están permitidas mientras la probabilidad asignada a los puntos de $C(g)$ sea 1 (la probabilidad de que se de una discontinuidad es nula). Un ejemplo claro de lo último es la función $1/X_n$, la cual tiende en distribución a $1/X$, dado que $P(X = 0) = 0$ y $x = 0$ es el único punto de discontinuidad

- Aunque anteriormente se ha mencionado que el estimador MLE y el estimador UMVUE es consistente y asintóticamente eficiente, la eficiencia se define sobre estimadores asintóticamente normales, lo cual a su vez implica consistencia

  - Suponiendo que $\sqrt{n}(T_n - \mu)/\sigma$ tiende en distribución a una normal estándar, se puede aplicar el teorema de Slutsky para ver que T_n tiende en distribución a μ y que, como la convergencia en distribución a un punto es equivalente a la convergencia en probabilidad, entonces T_n es consistente

$$\frac{\sqrt{n}(T_n - \mu)}{\sigma} \xrightarrow{D} Z \sim N(0,1) \;\Rightarrow\; T_n - \mu = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{n}(T_n - \mu)}{\sigma}$$

$$\Rightarrow \lim_{n\to\infty} T_n - \mu = \lim_{n\to\infty} \frac{\sigma}{\sqrt{n}} \frac{\sqrt{n}(T_n - \mu)}{\sigma} = \lim_{n\to\infty} \frac{\sigma Z}{\sqrt{n}} = 0$$

$$\Rightarrow T_n \xrightarrow{D} \mu \Leftrightarrow T_n \xrightarrow{P} \mu \Rightarrow T_n \ is\ consistent$$

- Por lo tanto, como se puede demostrar que los estimadores MLE y UMVUE son asintóticamente eficientes (debido al teorema visto anteriormente), entonces también son consistentes

El modelo de regresión lineal simple

- En el análisis de la varianza se analizaba como un factor influenciaba las medias de la variable respuesta, por lo que ahora se pasa al análisis de la regresión simple, en donde se intenta entender la dependencia funcional de una variable con otra
    - En particular, en la regresión lineal simple se tiene una relación lineal entre la variable aleatoria Y_i y una variable observable x_i

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

        - Las cantidades α y β se denominan intercepto y pendiente, se asume que son parámetros fijos y desconocidos y que ε_i es necesariamente una variable aleatoria
        - Lo más común es asumir que $E(\varepsilon_i|x_i) = 0$ (ya que el exceso se podría tener en cuenta a través de escalar α), de modo que se obtiene la siguiente igualdad:

$$E(Y_i|x_i) = \alpha + \beta x_i$$

        - En general, la función que da $E(Y_i|x_i)$ como función de x_i se denomina la función de regresión poblacional, y la expresión anterior es la función de regresión poblacional para el modelo de regresión simple
    - Un motivo principal de utilizar una regresión es predecir Y_i a partir del conocimiento de x_i usando una relación como la propuesta
        - Normalmente se dice que Y_i depende de x_i, siendo Y_i la variable dependiente y x_i la variable independiente (aunque no se refiera a que las x_i sean estadísticamente independiente, ya que no hace falta que sean variables aleatorias)
        - En el contexto de predicción, no obstante, Y_i se conoce como la variable respuesta y x_i como el predictor

        - Como se puede ver, no importa si x_i es un valor fijo o una realización de un valor para la variable aleatoria x_i, ya que $E(Y_i|x_i)$ tendrá la misma interpretación

    - La palabra regresión se utiliza en estadística para referirse a una relación entre variables. Cuando se dice que esta es lineal, quiere decir que $E(Y_i|x_i)$ es una función lineal de x_i y que la especificación es lineal en los parámetros

        - Por lo tanto, da igual si las variables son o no son lineales, solo es necesario que la expresión sea lineal en sus parámetros

        - Cuando se utiliza $E(Y_i|x_i) = \alpha + \beta x_i$, se hace la suposición implícita que la relación entre ambas variables es lineal, aunque no haya teoría o evidencia que apoye esto. Por lo tanto, no se espera que la ecuación se cumpla con exactitud, pero sí que sea una aproximación razonable

$$E(Y_i|x_i) \approx \alpha + \beta x_i$$

    - Cuando se realiza un análisis de regresión lineal (de la relación entre ambas variables) se realizan dos pasos: uno enfocado en los datos y otro enfocado en la estadística

        - El primer paso consiste en resumir los datos observados a través de diversos estadísticos (medias, varianzas, etc.) y no se hace ninguna suposición sobre lo parámetros

        - El segundo paso consiste en hacer inferencias sobre la relación en la población (de la función de regresión poblacional). Para hacer esto, es necesario hacer suposiciones sobre la población y sobre los parámetros poblacionales (que corresponderían a las cantidades del intercepto y la pendiente)

- En un problema de regresión lineal simple, normalmente se observan datos consistiendo en n pares $(x_1, y_1), \dots, (x_n, y_n)$, y se pueden considerar varios modelos para estos datos

    - Cada modelo conlleva diferentes suposiciones sobre las variables, y en cada modelo se intenta analizar una relación lineal entre x e y

        - Los n datos no caerán exactamente en una línea recta, pero lo interesante es resumir la información muestral ajustando una línea a los puntos de datos observados

- Se podrá ver que varios enfoques conllevan estimar la misma recta

- Basándose en los datos $(x_1, y_1), \dots, (x_n, y_n)$ se definen unas relevantes que componen los estimadores más comunes para la recta del modelo:

  - Las medias muestrales de las variables se definen de la siguiente manera:

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

  - Las sumas de cuadrados de las variables se definen de la siguiente manera:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x}) \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})$$

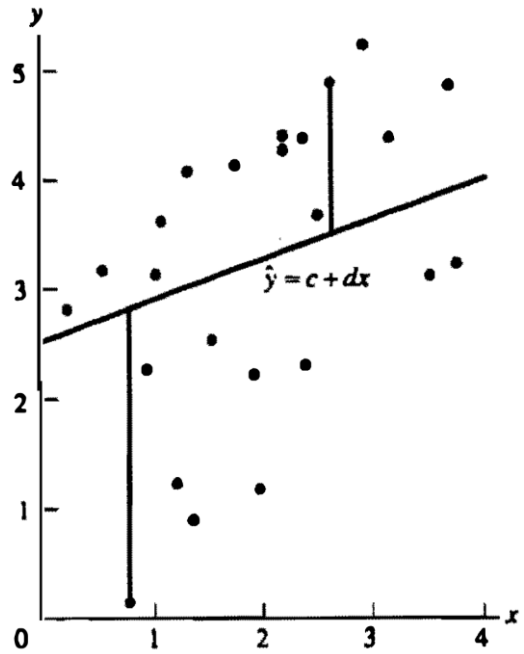
  - La suma de productos cruzados de las variables se define de la siguiente manera:

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

  - Los estimadores más comunes de α y de β en $E(y|x) = \alpha + \beta x$ (lo cual se justifica bajo varios modelos que se comentarán) se denotan por a y b, respectivamente, y sus expresiones son las siguientes:

$$a = \bar{y} - b\bar{x} \qquad b = \frac{S_{xy}}{S_{xx}}$$

- La primera derivación de los estimadores no implica hacer suposiciones sobre las observaciones. Considerando los datos $(x_1, y_1), \dots, (x_n, y_n)$ como n pares en un gráfico, el objetivo es dibujar una recta que esté lo más cerca posible a todos los puntos en la nube de puntos



x	y	x	y	x	y	x	y
3.74	3.22	0.20	2.81	1.22	1.23	1.76	4.12
3.66	4.87	2.50	3.71	1.00	3.13	0.51	3.16
0.78	0.12	3.50	3.11	1.29	4.05	2.17	4.40
2.40	2.31	1.35	0.90	0.95	2.28	1.99	1.18
2.18	4.25	2.36	4.39	1.05	3.60	1.53	2.54
1.93	2.24	3.13	4.36	2.92	5.39	2.60	4.89
$\bar{x} = 1.95$	$\bar{y} = 3.18$	$S_{xx} = 22.82$		$S_{yy} = 43.62$		$S_{xy} = 15.48$	

- Para cualquier línea $y = c + dx$, la suma cuadrada de residuos (SSR) es una medida de la distancia vertical desde cada punto de datos a la línea $c + dx$ que después suma los cuadrados de esta distancia. Esta se define de la siguiente manera:

$$SRR = \sum_{i=1}^{n} \hat{\varepsilon}^2 = \sum_{i=1}^{n} \big(y_i - (c + dx)\big)^2$$

- Los estimadores de mínimos cuadrados de α y β están definidos como aquellos valores a y b tal que la línea $a + bx$ minimiza el SSR, por lo que satisfacen el siguiente problema de minimización:

$$\min_{c,d} \sum_{i=1}^{n} \big(y_i - (c + dx_i)\big)^2 = \sum_{i=1}^{n} \big(y_i - (a + bx_i)\big)^2$$

- Resolviendo el problema de minimización, se obtiene que los estimadores que minimizan esta función (que es convexa) son los vistos anteriormente:

$$\frac{\partial}{\partial c} = -2\sum_{i=1}^{n}(y_i - c - dx_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - nc - d\sum_{i=1}^{n} x_i = 0 \;\Rightarrow\; a = \bar{y} - d\bar{x}$$

$$\frac{\partial}{\partial d} = -2\sum_{i=1}^{n}(y_i - c - dx_i)x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \bar{y} + d\sum_{i=1}^{n} x_i \bar{x} - d\sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow\; b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- Ambas condiciones de primer orden se pueden reescribir en términos de los residuos, de modo que la condición necesaria es que los residuos $(y_i - \hat{y})$ son ortogonales y la suma de estos residuos es nula (es ortogonal al regresor constante $x_0 = 1$ parar toda $i = 1,2, \dots, n$)

$$\frac{\partial}{\partial c} = -2\sum_{i=1}^{n}(y_i - c - dx_i) = 0 \;\Rightarrow\; \sum_{i=1}^{n}(y_i - \hat{y}) = 0$$

$$\frac{\partial}{\partial d} = -2\sum_{i=1}^{n}(y_i - c - dx_i)x_i = 0 \;\Rightarrow\; \sum_{i=1}^{n}(y_i - \hat{y})x_i = 0$$

- El SSR es solo una de las muchas maneras razonables de medir la distancia de la línea $c + dx$ a los puntos

  - Si se quieren utilizar, por ejemplo, distancias horizontales, entonces es lo mismo que hacer el mismo proceso, pero con loss ejes invertidos, de modo que intercambiando los roles de las variables se obtendría la línea $\hat{x} = a' + b'y$ y los siguientes estimadores:

$$a' = \bar{x} - d\bar{y} \qquad b' = \frac{S_{xy}}{S_{yy}}$$

  - Expresando $\hat{x} = a' + b'y$ en términos de y, se obtiene $\hat{y} = -(a'/b') + (1/b')x$

- Si ambas líneas fueran las mismas, entonces $b/(1/b') = 1$. No obstante, $b/(1/b') \leq 1$, dado que se puede aplicar la desigualdad de Cauchy-Schwarz en esta razón de la siguiente manera

$$\frac{b}{1/b'} = bb' = \frac{S_{xy}^2}{S_{xx}S_{yy}} \;\Rightarrow\; S_{xy}^2 \leq S_{xx}S_{yy} \;\Rightarrow\; b/(1/b') \leq 1$$

- Si x es una variable predictora, y es la variable respuesta y se intenta predecir y a partir de x, entonces la distancia vertical medida en el SSR es razonable, midiendo la distancia entre y_i y $\hat{y}_i = c + dx_i$. No obstante, si no se hace esta distinción entre x e y, entonces es raro que la distancia horizontal (otro criterio razonable) de una línea diferente

- El método de mínimos cuadrados solo se puede considerar como un método para ajustar una línea a un conjunto de datos, pero no como un método para inferencia estadística. Entonces, se entiende que a y b son soluciones a este problema más que estimadores

- Los estimadores para a y b también son óptimos en la clase de estimadores no sesgados bajo un modelo estadístico general. Asumiendo que $x_1, \dots, x_n$ son valores fijos conocidos y que los valores observados $y_1, \dots, y_n$ provienen de variables aleatorias no correlacionadas $Y_1, \dots, Y_n$

  - La relación lineal asumida entre x e y es la siguiente ecuación, en donde también se asume que la varianza es constante e igual para todas las Y:

$$E(Y_i) = \alpha + \beta x_i \quad for \quad i = 1,2, \dots, n \quad \& \quad Var(Y_i) = \sigma^2$$

  - Este modelo se puede expresar equivalentemente incluyendo un término de error ε_i en la expresión para Y_i, donde $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ son variables aleatorias no correlacionadas con esperanza nula y varianza constante e igual para todas las variables

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad for \quad i = 1,2, \dots, n$$

$$E(\varepsilon_i) = 0 \quad \& \quad Var(\varepsilon_i) = \sigma^2 \quad for \quad i = 1,2, \dots, n$$

  - Como Y_i depende solo de ε_i y las ε_i no están correlacionadas, las Y_i están no correlacionadas. Además, $E(\varepsilon_i)$ y $Var(\varepsilon_i)$ cumplen las mismas suposiciones que el modelo anterior

- Un estimador lineal es un estimador de la forma $\sum_{i=1}^{n} d_i Y_i$ donde $d_1, d_2, \dots, d_n$ son cantidades fijas, y un estimador no sesgado para un parámetro β es un estimador tal que $E(\sum_{i=1}^{n} d_i Y_i) = \beta$ independientemente del valor real de los parámetros α y β, implicando lo siguiente:

$$\beta = E\left(\sum_{i=1}^{n} d_i Y_i\right) = \sum_{i=1}^{n} d_i E(Y_i) = \sum_{i=1}^{n} d_i(\alpha + \beta x_i) =$$

$$= \alpha\left(\sum_{i=1}^{n} d_i\right) + \beta\left(\sum_{i=1}^{n} d_i x_i\right)$$

- Esta igualdad solo es verdad para todos los valores α y β si y solo si $\sum_{i=1}^{n} d_i = 0$ y $\sum_{i=1}^{n} d_i x_i = 1$. Por lo tanto, $d_1, d_2, \dots, d_n$ deben satisfacer estas condiciones para un estimador β

- El método para obtener estimadores de mínimos cuadrados no suponía propiedades estadísticas, mientras que las suposiciones estadísticas para los estimadores lineales no sesgados solo eran para los primeros dos momentos, de modo que no se pueden crear contrastes ni intervalos de confianza los estimadores exactos. Por lo tanto, se pueden presentar modelos estadísticos más concretos que especifican completamente la estructura probabilística de los datos para hacer inferencias

    - El modelo normal condicional es el modelo de regresión lineal más simple y más sencillo de analizar. Este modelo tiene la siguiente serie de suposiciones e implicaciones:

        - Los valores $x_1, x_2, \dots, x_n$ son fijos y conocidos, mientras que $y_1, y_2, \dots, y_n$ son observaciones de las variables aleatorias $Y_1, Y_2, \dots, Y_n$

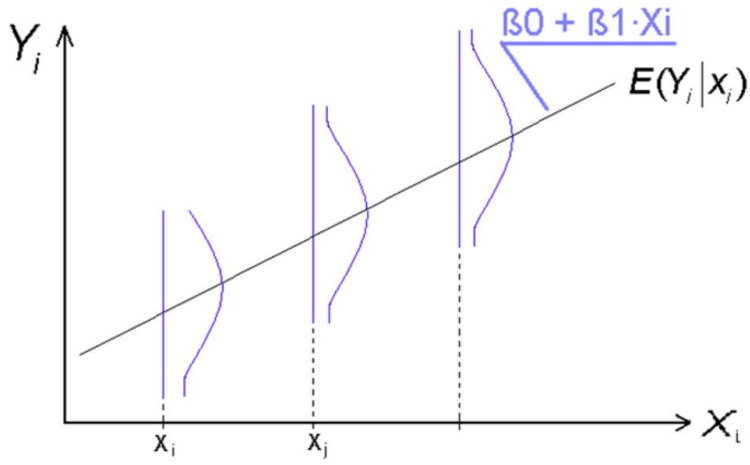
        - Las variables $Y_1, Y_2, \dots, Y_n$ son independientes y, además, se asume que las Y_i se distribuyen normalmente con media $\alpha + \beta x_i$ y varianza constante σ^2 (homocedasticidad):

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2) \quad for \;\; i = 1,2, \dots, n$$

        - Por lo tanto, la línea de regresión poblacional es una función de x, de modo que $E(Y_i|x_i) = \alpha + \beta x_i$ y todas las Y_i tienen la misma varianza σ^2. El modelo normal condicional puede expresarse con un término de error, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ son independientes e idénticamente distribuidos siguiendo $N(0, \sigma^2)$

$$Y_i = \alpha + \beta x_i + \varepsilon_i \;\; for \;\; i = 1,2, \dots, n$$

$$where \;\; \varepsilon_i \sim N(0, \sigma^2) \;\; and \;\; Cov(\varepsilon_i, \varepsilon_j) = 0 \;\; for \;\; i \neq j$$



- Este es un caso especial del modelo estadístico considerado anteriormente, manteniendo media y varianza igual, pero fortaleciendo la no correlación con independencia y definiendo una distribución de probabilidad completamente. La función de distribución conjunta de $Y_1, Y_2, \dots, Y_n$ se especifica de la siguiente manera:

$$f(\boldsymbol{y}|\alpha, \beta, \sigma^2) = \prod_{i=1}^{n} f(y_i|\alpha, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\frac{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right]$$

- En todos los modelos anteriores se asumía que $x_1, x_2, \dots, x_n$ eran valores fijos conocidos, pero hay ocasiones estas son observaciones de variables aleatorias $X_1, X_2, \dots, X_n$, por lo que es necesario considerar modelos en las que la variable predictora y la variable respuesta sean aleatorias. El modelo más utilizado en este caso es el modelo normal bivariante

  - En el modelo bivariante, los datos $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ son los valores observados de las variables aleatorias $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$. Los vectores aleatorios son independientes y la distribución conjunta de (Y_i, X_i) es una normal bivariante

$$(Y_i, X_i) \sim N_2(\mu_Y, \mu_X, \sigma_X, \sigma_Y, \rho)$$

  - La función de densidad conjunta de todos los datos es el producto de las funciones de densidad bivariantes de cada par

- En un modelo de regresión lineal simple se sigue interpretando x como la variable predictora e y como la variable respuesta, por lo que esto lleva a basar la inferencia en la distribución condicional Y dado $X = x$. Para un modelo normal bivariante, la distribución condicional es normal y ahora la función de regresión poblacional es una esperanza condicionada real

$$E(Y|X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \alpha + \beta x$$

$$where \quad \alpha = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X \quad \& \quad \beta = \rho\frac{\sigma_Y}{\sigma_X}$$

- El modelo normal bivariante implica que la función de regresión poblacional es lineal en x, por lo que no es necesario asumir esto como en otros modelos. Además, igual que en el modelo condicional, la varianza condicional de la variable respuesta Y no depende de x

$$Var(Y|X = x) = \sigma_Y^2(1 - \rho^2)$$

- El análisis de regresión en el modelo normal bivariante se lleva a cabo usando la distribución condicional, más que con la distribución incondicional. Por lo tanto, se está en el mismo caso y no se necesita utilizar el hecho de normalidad bivariante excepto que para definir la distribución condicional

- Asumiendo que se cumple el modelo de regresión condicional normal definido anteriormente, es posible desarrollar procedimientos de inferencia

  - Para poder desarrollar estos procedimientos es necesario obtener los estimadores máximo verosímiles tanto de los parámetros como de la varianza del término de error, los cuales dependen de la siguiente función de verosimilitud

$$\log L\,(\alpha, \beta, \sigma^2|\boldsymbol{x}, \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}{2\sigma^2}$$

    - Para cualquier valor fijo de σ^2, se puede ver como $\log L$ se maximiza como una función de α y β por esos valores $\hat{\alpha}$ y $\hat{\beta}$ que minimizan la suma cuadrada de residuos. Por lo tanto, los estimadores de mínimos cuadrados ordinarios son los estimadores máximo verosímiles de α y β para cualquier σ^2 fija

$$\hat{\beta} = \hat{\beta}_{MLE} = \hat{\beta}_{OLS} = \frac{S_{xy}}{S_{xx}} \qquad \hat{\alpha} = \hat{\alpha}_{MLE} = \hat{\alpha}_{OLS} = \bar{y} - \hat{\beta}\bar{x}$$

- Sustituyendo estos valores en la función de verosimilitud logarítmica, se puede obtener el estimador máximo verosímil para la varianza, el cual es la suma cuadrada de residuos dividido entre el número de observaciones

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}{2\sigma^2}$$

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

- Aunque $\hat{\alpha}$ y $\hat{\beta}$ son estimadores lineales no sesgados de α y β, pero el estimador $\hat{\sigma}^2$ es un estimador sesgado de σ^2. Esto se puede demostrar a través de la esperanza de $\hat{\sigma}^2$

  - Un lema que será útil para la obtención de la esperanza del estimador es el siguiente: siendo $Y_1, Y_2, \dots, Y_n$ variables aleatorias no correlacionadas con varianza $Var(Y_i) = \sigma^2$ para toda $i = 1,2, \dots, n$ y siendo $c_1, c_2, \dots, c_n$ y $d_1, d_2, \dots, d_n$ dos conjuntos de constantes, entonces se da la siguiente igualdad:

$$Cov\left(\sum_{i=1}^{n} c_i Y_i\,, \sum_{i=1}^{n} d_i Y_i\right) = \sigma^2 \sum_{i=1}^{n} c_i d_i$$

  - Debido a que el error se define como $\varepsilon_i = Y_i - \alpha - \beta x_i$, los residuos se pueden definir como $\hat{\varepsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$, de modo que es posible expresar $\hat{\sigma}^2$ en términos de $\hat{\varepsilon}_i$ y así obtener la esperanza de $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$$

$$\begin{cases} E(\hat{\varepsilon}_i) = E(Y_i) - E(\hat{\alpha}) - E\left(\hat{\beta}\right)x_i = \alpha - \beta x_i - \alpha + \beta x_i = 0. \\ Var(\hat{\varepsilon}_i) = \left[\frac{n-2}{n} + \frac{1}{S_{xx}}\left(\frac{1}{n}\sum_{j=1}^{n} x_j^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i\bar{x}\right)\right]\sigma^2 \end{cases}$$

$$\Rightarrow E(\hat{\sigma}^2) = \frac{1}{n}E\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right) = \frac{1}{n}\sum_{i=1}^{n}E(\hat{\varepsilon}_i^2) = \frac{1}{n}\sum_{i=1}^{n}Var(\hat{\varepsilon}_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n-2}{n} + \frac{1}{nS_{xx}}\left(\sum_{j=1}^{n} x_j^2 + \sum_{i=1}^{n} x_i^2 - 2S_{xx} - \frac{2}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)\right]\sigma^2$$

$$\Rightarrow\ S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \Rightarrow\ E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2$$

- Como el estimador está sesgado, entonces se utiliza un estimador que ajusta por los grados de libertad

$$S_R^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- Parar desarrollar los procedimientos de estimación y de contraste basados en los estimadores es necesario saber su distribución muestral, las cuales se resumen en el siguiente teorema: bajo el modelo de regresión condicional normal, las distribuciones muestrales de los estimadores $\hat{\alpha}$, $\hat{\beta}$ y $\hat{\sigma}^2$ son las siguientes:

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n}x_i^2\right) \qquad \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \qquad \frac{(n-2)S_R^2}{\sigma^2} \sim \chi_{n-2}^2$$

  - Además, existe correlación entre $\hat{\alpha}$ y $\hat{\beta}$, pero $(\hat{\alpha}, \hat{\beta})$ y S_R^2 son independientes

$$Cov(\hat{\alpha}, \hat{\beta}) = \frac{-\sigma^2\bar{x}}{S_{xx}}$$

  - Los estimadores $\hat{\alpha}$ y $\hat{\beta}$ son funciones lineales de las variables aleatoria normales independientes $Y_1, Y_2, \ldots, Y_n$. Por lo tanto, como una combinación lineal de variables aleatorias normales es una variable normal, entonces cada estimador sigue una distribución normal

  - La esperanza y la varianza de cada una de las distribuciones, junto a su covarianza, se puede obtener de la siguiente manera:

$$E(\hat{\beta}) = \beta \qquad Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$E(\hat{\alpha}) = E\left(\sum_{i=1}^{n}c_iY_i\right) = \sum_{i=1}^{n}c_iE(Y_i) = \alpha\sum_{i=1}^{n}c_i - \beta\sum_{i=1}^{n}c_ix_i = \alpha$$

$$Var(\hat{\alpha}) = \sum_{i=1}^{n}c_i^2Var(Y_i) = \sigma^2\sum_{i=1}^{n}c_i^2 = \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n}x_i^2$$

$$Cov(\hat{\alpha}, \hat{\beta}) = Cov\left(\sum_{i=1}^{n}c_iY_i, \sum_{i=1}^{n}d_iY_i\right) = \sigma^2\sum_{i=1}^{n}c_id_i =$$

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

- La suma cuadrada de residuos contiene información lineal sobre el valor de un ajuste polinómico de un mayor orden (sobre el ajuste lineal). Como en este modelo se asume que la regresión poblacional es lineal, la variación en el ajuste de mayor orden solo es variación aleatoria

    - Alternativamente, se puede utilizar el teorema de Cochran para establecer que $\sum_i \hat{\varepsilon}_i^2 \, / \sigma^2 \sim \chi_{n-2}^2$

- Las inferencias con respecto a los parámetros α y β normalmente se basan en las siguientes distribuciones t-Student, las cuales se derivan inmediatamente de la normalidad de sus estimadores y de la distribución muestral de σ^2:

$$\frac{\hat{\alpha} - \alpha_0}{S_R \sqrt{\sum_{i=1}^{n} x_i^2 \, / n S_{xx}}} \sim t_{n-2} \qquad \frac{\hat{\beta} - \beta_0}{S_R / \sqrt{S_{xx}}} \sim t_{n-2}$$

    - La distribución conjunta de ambos estadísticos forma una distribución t-Student bivariante. Esta distribución se deriva de manera análoga al caso univariante, utilizando el hecho de que la distribución conjunta de $\hat{\alpha}$ y $\hat{\beta}$ es una normal bivariante y que se S está en ambos estadísticos univariantes

    - Aunque se podría utilizar la distribución t-Student bivariante para contrastar estimaciones a la vez, se suele contrastar de manera separada

    - Debido a que el estadístico t para β sigue una distribución t-Student con $n - 2$ grados de libertad, se puede invertir este contraste para poder obtener un intervalo de confianza para β. Esto también se puede hacer para α

$$\hat{\beta} - t_{n-2,\frac{\alpha}{2}} \frac{S_R}{\sqrt{S_{xx}}} \leq \beta \leq \hat{\beta} + t_{n-2,\frac{\alpha}{2}} \frac{S_R}{\sqrt{S_{xx}}}$$

$$\hat{\alpha} - t_{n-2,\frac{\alpha}{2}} S_R \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n S_{xx}}} \leq \alpha \leq \hat{\alpha} + t_{n-2,\frac{\alpha}{2}} S_R \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n S_{xx}}}$$

- El interés suele residir en contrastar β más que α, dado que el parámetro α es el valor esperado $E(Y_i | x_i = 0)$, dependiendo del problema, esta cantidad puede ser de interés o no

- En particular, puede ser que $x_i = 0$ no sea un valor razonable para la variable predictora

- No obstante, β es la cantidad en la que $E(Y_i|x_i)$ cambia cuando x_i cambia en una unidad, de modo que se relaciona con el rango entero de valores de x_i y contiene la información sobre la relación lineal que existe entre Y_i y x_i. Un valor particularmente interesante es $\beta = 0$, ya que eso haría que $Y_i \sim N(\alpha, \sigma^2)$, y que, por tanto, no dependa de x

- Para contrastar la hipótesis nula $H_0: \beta = 0$ contra $H_1: \beta \neq 0$, se puede usar el estadístico t o, equivalentemente, elevar al cuadrado este para obtener un estadístico con distribución $F_{1,n-2,\alpha}$ (ya que el estadístico t cuadrado sigue una distribución F)

$$\left|\frac{\beta}{S_R/\sqrt{S_{xx}}}\right| > t_{n-2.\alpha/2} \qquad \frac{\hat{\beta}^2}{S_R^2/S_{xx}} > F_{1,n-2,\alpha}$$

- Si se expande el estadístico t cuadrado, se puede ver como hay una expresión equivalente que resulta ser el estadístico F de la tabla ANOVA

$$\frac{\hat{\beta}^2}{S_R^2/S_{xx}} = \frac{S_{xy}^2/S_{xx}}{RSS/(n-2)} = \frac{Regression\ sum\ of\ squares}{Residual\ sum\ of\ squares/df}$$

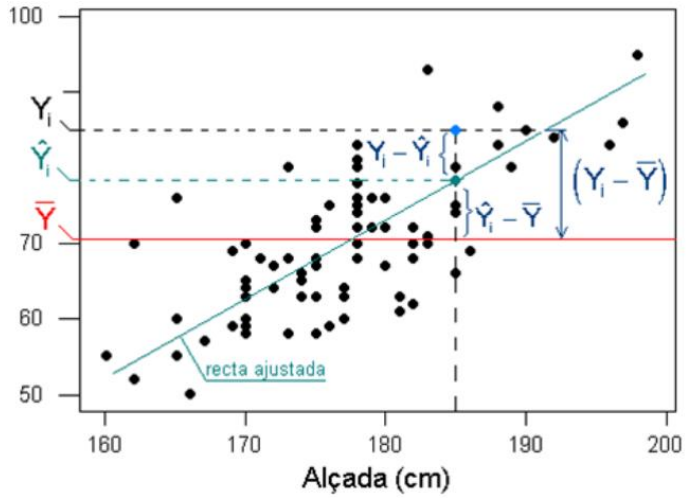
- Por lo tanto, el estadístico del análisis de la varianza para una regresión lineal simple permite contrastar si β es cero o si es diferente de cero

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Regression (slope)	1	Reg. SS = S_{xy}^2/S_{xx}	MS(Reg) = Reg. SS	$F = \frac{\text{MS(Reg)}}{\text{MS(Resid)}}$
Residual	$n-2$	RSS = $\sum \hat{\epsilon}_i^2$	MS(Resid) = RSS/$(n-2)$	
Total	$n-1$	SST = $\sum (y_i - \bar{y})^2$		

- La partición de la suma de cuadrados total para el ANOVA tiene un análogo para el caso de regresión. En este caso, la SSR mide la desviación de la línea ajustada a los valores observados, mientras que la suma de cuadrados de la regresión es análoga a la suma de cuadrados de tratamiento de la ANOVA y mide la desviación de los valores predichos de la media total

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$TSS = Reg.SS + SSR$$



- La suma de cuadrados de la regresión se puede expresar en términos de S_{xy} y de S_{xx}, conectando así con el contraste t visto anteriormente

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}} = \sqrt{Var(\hat{\beta})}$$

- La suma de cuadrados total no depende de los valores de x (solo de y) y no cambia a no ser que se aplique una transformación sobre y

- Un estadístico que permite cuantificar que tan bien la línea ajustada describe los datos (la bondad del ajuste) es el coeficiente de determinación R^2, definido como la razón entre la suma de cuadrados de la regresión y la suma de cuadrados total

$$R^2 = \frac{Regression\ sum\ of\ squares}{Total\ sum\ of\ squares} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

- El coeficiente de determinación mide la proporción de la variación en $y_1, y_2, \dots, y_n$ (medida por S_{yy}) que se explica por la línea predicha (medida por la suma de cuadrados de la regresión)

- Debido a la desigualdad de Cauchy-Schwarz se puede ver como $R^2 \in [0,1]$. Si $y_1, y_2, \dots, y_n$ caen exactamente en la línea ajustada, entonces $y_i = \hat{y}_i$ para toda i y $R^2 = 1$, pero si $y_1, y_2, \dots, y_n$ están lejos de la línea, la SSR será grande y $R^2 \approx 0$

- En este caso, R^2 también se puede derivar como el cuadrado del coeficiente de correlación muestral de los n pares $(y_1, x_1), \dots, (y_n, x_n)$ o de los n pares $(y_1, \hat{y}_1), \dots, (y_n, \hat{y}_n)$

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad \rho_{xy} = \frac{Cov(x, y)}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

$$\Rightarrow R^2 = \rho_{xy}^2$$

- Existe una población valores de Y asociado con un valor específico de la variable predictora $x = x_0$. Una vez estimados los parámetros del modelo, el experimentador puede fijar $x = x_0$ y obtener una nueva observación Y_0, de modo que puede haber interés en estimar la media de la población $\hat{Y}(x_0)$ de la que esta observación se ha extraído o en predecir cuál será la observación Y_0

  - Se asume que $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ satisfacen el modelo condicional normal y que se obtienen los estimadores $\hat{\alpha}$, $\hat{\beta}$ y S_R^2 a partir de las n observaciones

    - Siendo x_0 un valor especificado para la variable predictora, se considera estimar la media de la población Y asociada a x_0, es decir, $E(Y|x_0) = \alpha + \beta x_0$

    - Para poder estimarla, un estimador obvio es $\hat{\alpha} + \hat{\beta} x_0$, el cual no tiene sesgo y tiene la siguiente varianza:

$$E\left(\hat{Y}(x_0)\middle|x_0\right) = E\left(\hat{\alpha} + \hat{\beta}x_0\middle|x_0\right) = \alpha + \beta x_0$$

$$Var\left(\hat{Y}(x_0)\middle|x_0\right) = Var(\hat{\alpha}|x_0) + Var\left(\hat{\beta}\middle|x_0\right)x_0 + 2x_0 Cov\left(\hat{\alpha}, \hat{\beta}\right)$$

$$= \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n} x_i^2 + \frac{\sigma^2 x_0^2}{S_{xx}} - \frac{2\sigma^2 x_0 \bar{x}}{S_{xx}} =$$

$$= \frac{\sigma^2}{S_{xx}}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 + \bar{x}^2 - 2x_0\bar{x} + x_0^2\right) =$$

$$= \frac{\sigma^2}{S_{xx}}\left(\frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] + (x_0 - \bar{x})^2\right) =$$

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

- Como $\hat{\alpha}$ y $\hat{\beta}$ son una función lineal de $Y_1, Y_2, \ldots, Y_n$, $\hat{\alpha} + \hat{\beta}x_0$ también es una función lineal de estas variables. Por lo tanto, también sigue una distribución normal

$$\hat{\alpha} + \hat{\beta}x_0 \sim N\left(\alpha + \beta x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

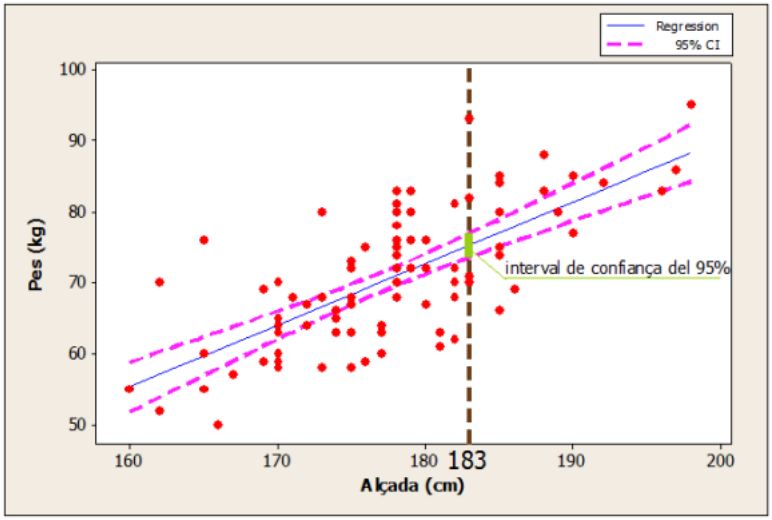
  - Debido a que $(\hat{\alpha}, \hat{\beta})$ y S^2 son independientes, entonces S^2 también es independiente de $\hat{\alpha} + \hat{\beta}x_0$ y, por tanto, se puede construir un estadístico t que sigue una distribución t_{n-2}

$$\frac{\hat{\alpha} + \hat{\beta}x_0 - \alpha - \beta x_0}{S_R\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

  - Este estadístico se puede utilizar como una cantidad pivotal y así obtener un intervalo de confianza para $\alpha + \beta x_0$

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\frac{\alpha}{2}} S_R \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

  - La longitud del intervalo para $\alpha + \beta x_0$ dependerá de los valores de $x_1, x_2, \ldots, x_n$ a través del valor $(x_0 - \bar{x})^2/S_{xx}$. Está claro que el ancho será menor cuanto más cerca esté x_0 de $\bar{x}$ (lógicamente, dado que se intenta estimar para la esperanza), por lo que, en un diseño de experimento, lo que se querrá es escoger unos valores $x_1, x_2, \ldots, x_n$ tales que $x_0 = \bar{x}$ o cerca



- Un tipo de inferencia que no se ha discutido hasta ahora es la predicción de una variable aleatoria no observada Y, un tipo de inferencia que es de interés en un marco de regresión

- Un intervalo de predicción $100(1-\alpha)\%$ para una variable no observada Y basado en los datos observados $\boldsymbol{X}$ es un intervalo aleatorio $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ que cumple la siguiente propiedad para cualquier valor del parámetro θ

$$P_\theta[L(\boldsymbol{X}) \leq Y \leq U(\boldsymbol{X})] \geq 1-\alpha$$

- La definición del intervalo de predicción y del intervalo de confianza son parecidas, pero la diferencia está en que el intervalo de predicción es un intervalo en una variable aleatoria, mientras que el de confianza lo es para un parámetro

○ Intuitivamente, como una variable aleatoria es más variable que un parámetro (que es constante), se espera que el intervalo de predicción sea más ancho que el de confianza para un mismo nivel

- Asumiendo que la nueva observación Y_0 que se toma en $x = x_0$ tiene una distribución $N(\alpha + \beta x_0, \sigma^2)$, independiente de los datos anteriores $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$, los estimadores $\hat{\alpha}$, $\hat{\beta}$ y S^2 son independientes de Y_0 y $Y_0 - \hat{\alpha} - \hat{\beta}x_0$ tiene la siguiente distribución:

$$E\left(Y_0 - \hat{\alpha} - \hat{\beta}x_0\right) = \alpha + \beta x_0 - \alpha - \beta x_0 = 0$$

$$Var\left(Y_0 - \hat{\alpha} - \hat{\beta}x_0\right) = Var(Y_0) + Var\left(\hat{\alpha} + \hat{\beta}x_0\right) =$$

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

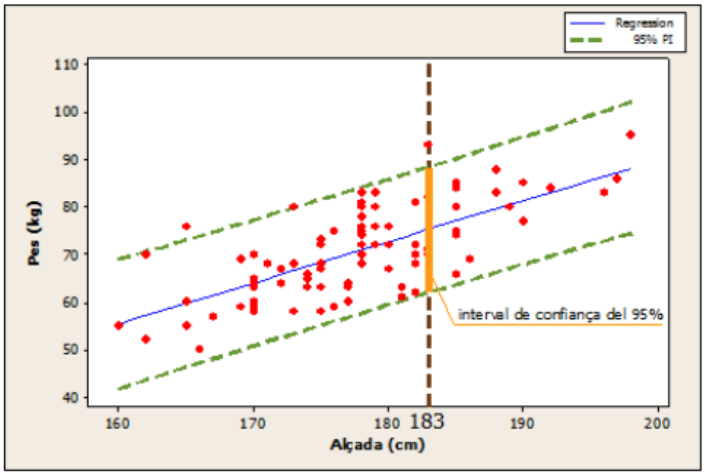
$$\Rightarrow N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

- Usando la condición de independencia entre $Y_0 - \hat{\alpha} - \hat{\beta}x_0$ y S^2, es posible construir un estadístico t que siga una distribución t-Student con $t-2$ grados de libertad y que puede invertirse para obtener un intervalo de predicción

$$\frac{Y_0 - \hat{\alpha} - \hat{\beta}x_0}{S_R\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

$$\Rightarrow \hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\frac{\alpha}{2}} S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- Debido a que los límites del intervalo dependen solo de los datos observados, la expresión define el intervalo de predicción para la nueva observación Y_0



  - Es importante remarcar que en todo momento se hace la suposición implícita de que x_0 está dentro del rango de valores de x en la muestra usada para ajustar el modelo

    - En este contexto, la predicción es una interpolación de valores para aquellos dentro de la muestra

    - De otro modo, la predicción es una extrapolación que se puede justificar si se puede asumir que el modelo teórico validado y ajustado en el rango de la muestra también es válido en el área donde la predicción se hace

- Aunque se ha visto la predicción para un solo valor x_0, hay circunstancias en las que es de interés predecir varios valores de x, el cual es un problema de inferencia simultánea y se basa en controlar el nivel de confianza del intervalo general para la inferencia simultánea

  - Anteriormente se ha visto cuál es el intervalo de predicción que corresponde a un valor $x = x_0$, pero no se ha visto una metodología para inferenciar sobre la media poblacional de Y para varios valores x_0. Es decir, se quieren diferentes intervalos para $E(Y|x_{0i})$ para $i = 1,2, \dots, m$

    - Se sabe que si se construyen m intervalos como los anteriores, cada uno a un nivel de confianza $1 - \alpha$, la inferencia no será a un nivel $1 - \alpha$

  - Debido a que la suposición sobre la línea de regresión poblacional implica que la ecuación $E(Y|x) = \alpha + \beta x$ se mantiene para toda x, por

lo que se podría hacer inferencias para toda x. El siguiente teorema de Scheffé permite obtener un intervalo que se mantiene para toda x:

- Bajo el modelo de regresión normal condicional, la probabilidad es al menos $1-\alpha$ de que se cumpla la siguiente igualdad, simultáneamente para toda x:

$$P\left(\frac{\left[\left(\hat{\alpha}+\hat{\beta}x\right)-\left(\alpha+\beta x\right)\right]^2}{S_R^2\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right]} \leq M_\alpha^2\right) = 1-\alpha \;\; for\; \forall x$$

$$where\;\; M_\alpha = \sqrt{2F_{2,n-2,\alpha}}$$

- Esta cantidad pivotal se puede invertir para obtener una desigualdad estricta para $\alpha+\beta x$, las cuales conformarán las bandas de confianza o bandas de Scheffé

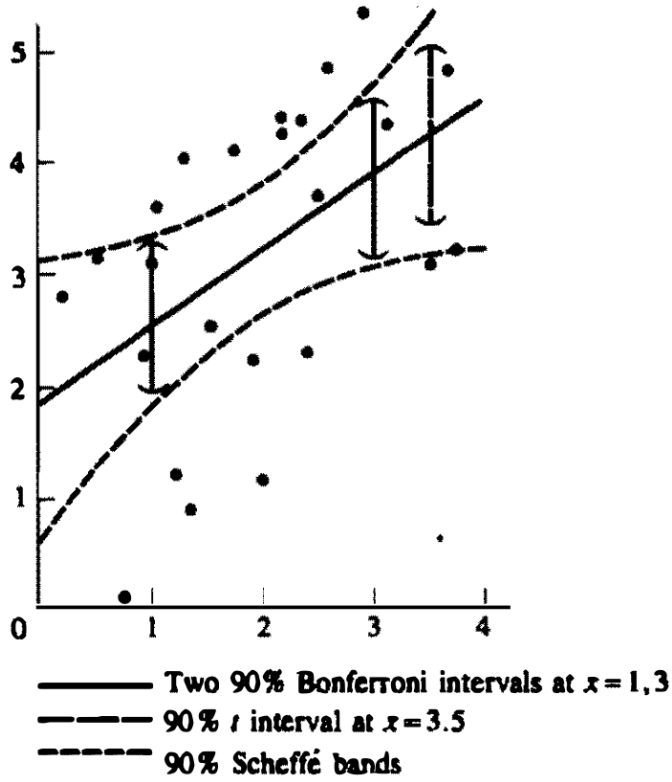
$$\hat{\alpha}+\hat{\beta}x - M_\alpha S_R\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}} < \alpha+\beta <$$

$$< \hat{\alpha}+\hat{\beta}x + M_\alpha S_R\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$$

- Debido a que esta última desigualdad se mantiene para toda x, entonces en verdad se tiene una banda de confianza para toda la línea de regresión poblacional

  - Igual que un intervalo de confianza cubre un solo parámetro con un solo valor, una banda de confianza cubre toda una línea con una banda

  - Por lo tanto, esta banda cubriría toda la línea de regresión de manera horizontal, mientras que el intervalo para Y_0 (el que está más a la derecha en el gráfico) sería un intervalo vertical en un punto x_0 (los otros dos intervalos en el gráfico)



- Como se puede observar, las bandas de confianza de Scheffé delimitan la longitud vertical del intervalo para $\hat{Y}(x_0)$ en el cual se ha aplicado el método de Bonferroni, pero no determinan la longitud vertical del intervalo de predicción para Y_0

El modelo de regresión lineal múltiple

- El modelo de regresión lineal simple se puede extender a más variables explicativas utilizando más regresores y coeficientes, de modo que ahora la variable respuesta se explica por un conjunto de variables
    - Desde el punto de vista teórico, pasar del modelo de regresión lineal simple al múltiple es trivial, ya que la única diferencia es cambiar el hecho de que la variable respuesta se explica a través de un plano (y no una recta), pero la teoría es idéntica
        - No obstante, la práctica es más complicada porque hay muchos modelos posibles (no siempre es fácil escoger el mejor), es más difícil identificar observaciones atípicas y hay dependencia entre variables explicativas, lo que dificulta la interpretación del modelo
    - El modelo teórico $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ tiene asociado un modelo ajustado $y_i = b_0 + b_1 x_i + \hat{\varepsilon}_i = \hat{y}_i + \hat{\varepsilon}_i$ para $i = 1,2,\dots,n$, por lo que se pueden generalizar ambos modelos a través de matrices

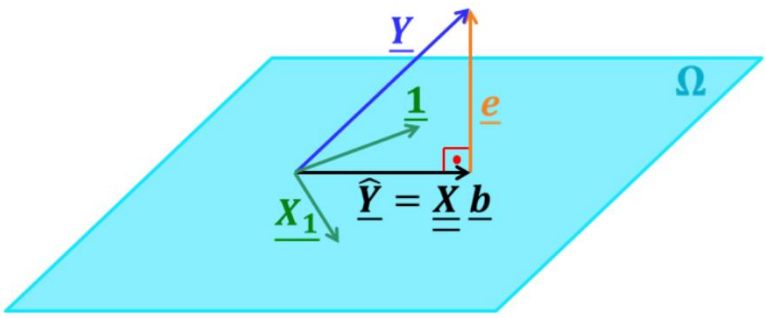
$$y_i = \beta_0 + \beta x_i + \varepsilon_i \;\Rightarrow\; \boldsymbol{y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}$$

$$y_i = b_0 + b_1 x_i + \hat{\varepsilon}_i = \hat{y}_i + \hat{\varepsilon}_i \;\Rightarrow\; \boldsymbol{y} = \boldsymbol{bX} + \hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$$

- En este caso, la matriz $\boldsymbol{X}$ es una matriz con una columna de unos (representando un regresor constante para β_0) y con las observaciones de cada x que se incluya en el modelo en las otras columnas (cada variable representará una columna)

- Se puede ver como $\hat{\boldsymbol{y}} = \boldsymbol{bX}$ es una combinación lineal de $\boldsymbol{X}$, de modo que $\hat{\boldsymbol{y}}$ pertenece al subespacio Ω generado por los vectores columna (las variables)

- Debido a que $\boldsymbol{b}$ es el vector de coeficientes de mínimos cuadrados que minimiza el SSR, entonces minimiza $\|\hat{\boldsymbol{\varepsilon}}\|^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ y eso hace que $\hat{\boldsymbol{\varepsilon}}$ sea ortogonal al vector de valores esperados $\hat{\boldsymbol{y}}$ y a los vectores columna en $\boldsymbol{X}$



- Las ecuaciones normales obtenidas para el caso de la regresión simple muestra como el vector de residuos es ortogonal a los vectores columna

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \;\Rightarrow\; \hat{\boldsymbol{\varepsilon}}'\mathbf{1} = \mathbf{0} \qquad \sum_{i=1}^{n} \hat{\varepsilon}_i x_i = 0 \;\Rightarrow\; \hat{\boldsymbol{\varepsilon}}'\boldsymbol{x_1'} = \mathbf{0}$$

- Una vez visto como funcionaría el modelo de regresión simple en el marco de la regresión múltiple, ahora es posible aumentar el número de variables para poder trabajar con el marco generalizado

  - La relación entre la variable respuesta y las variables explicativas se pueden modelar por $y_i = f\left(x_{1i}, x_{2i}, \dots, x_{(K-1)i}; \beta\right) + \varepsilon_i(z_{1i}, z_{2i}, \dots)$, donde las variables z son variables que explican y_i pero que no están en el modelo

    - La función $f\left(x_{1i}, x_{2i}, \dots, x_{(K-1)i}; \beta\right)$ es normalmente una función complicada y ε_i tiene una distribución que depende de las variables z

- No obstante, se puede aproximar la relación a través de un modelo lineal, en donde ε_i cumple las mismas propiedades que en el modelo lineal simple. Por lo tanto, el modelo teórico en esta situación es el siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{(K-1)i} + \varepsilon_i \;\Rightarrow\; \boldsymbol{y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

  - En este caso, $\boldsymbol{X}$ es una matriz $n \times K$ que incluye un vector de n unos y $K-1$ vectores columnas que representan las k variables. Del mismo modo, $\boldsymbol{\beta}$ es el vector de coeficientes para cada una de las $K-1$ variables y para el término constante

  - Además, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n\times n})$, de modo que se cumple $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $Cov(\varepsilon_i, \varepsilon_j) \neq 0$ para toda $i \neq j$ y ε_i tiene una distribución normal univariante

  - El vector respuesta $\boldsymbol{y}$ se modela por un modelo lineal múltiple si el rango de la matriz $\boldsymbol{X}$ es completo, de modo que $n \geq K$ y las columnas son linealmente independientes entre ellas, lo cual permitirá invertir la matriz

  - Si el rango de la matriz no es completo, entonces las variables no son linealmente independientes entre ellas (hay multicolinearidad perfecta) y no se puede invertir la matriz $\boldsymbol{X}$

- El modelo ajustado o muestral se puede escribir de manera equivalente, aunque, igual que en el modelo de regresión lineal simple, la distinción es que no se sabe el valor de los coeficientes y sus valores dependen de la muestra y del criterio de ajuste

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{(k-1)i} + \varepsilon_i$$

$$\Rightarrow\; \boldsymbol{y} = \boldsymbol{X}'\boldsymbol{b} + \hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$$

  - Por lo tanto, el modelo teórico sigue siendo un modelo único y desconocido

  - Las causas que complican la construcción y el uso de modelos lineales múltiples son que hay varios posibles modelos, que hay que tener en cuenta la dependencia entre las variables explicativas al interpretar y que no es fácil detectar observaciones atípicas

  - Para calcular el modelo ajustado, se utiliza el mismo criterio de mínimos cuadrados que se había visto anteriormente, solo que

se utilizan vectores y matrices. Esto permite obtener la siguiente ecuación:

$$\boldsymbol{X'y - X'Xb = 0} \;\Rightarrow\; \boldsymbol{X'y = X'Xb} \;\Rightarrow\; \boldsymbol{b = (X'X)^{-1}X'y}$$

- A partir de esta ecuación es posible entender por qué se necesita rango completo y que $\boldsymbol{b}$ es una combinación lineal de $\boldsymbol{y}$

- Igual que antes, también es posible realizar una tabla ANOVA para el modelo de regresión lineal múltiple que permita analizar la variabilidad de la variable dependiente y dividir esta variabilidad

Cause	Degrees Freedom	Sum of squares	Mean Squares
Explained by regression	$\nu_E = p - 1$	$SQ_E = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	$\frac{SQ_E}{p-1}$
Residual	$\nu_R = n - p$	$SQ_R = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$S_R^2 = \frac{SQ_R}{n-p}$
Total corrected	$\nu_T = n - 1$	$SQ_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	

  - Igual que antes, la suma de cuadrados total no depende de las x, solo de y, y no variará a no ser que se transformen las y. Sin embargo, la SSR y la suma de cuadrados de la regresión si dependen de las x

  - Se tienen K grados de libertad en la SSR correspondiente al modelo de regresión múltiple porque en verdad es como si hubieran K ecuaciones normales. En cambio, para la suma de cuadrados de la regresión habrían $K - 1$ grados de libertad (solo hay $K - 1$ variables explicativas a parte del regresor constante)

- A partir de esta división de la variabilidad, es posible obtener la medida de bondad del ajuste R^2 y una versión que tiene en cuenta la cantidad de regresores

  - En este caso, R^2 se define de la misma manera que antes, solo que ahora se interpreta como la correlación entre $\boldsymbol{y}$ e $\hat{\boldsymbol{y}}$, representando igual el porcentaje de variabilidad de y explicada por el modelo. Además, sigue cumpliéndose que $R^2 \in [0,1]$, pero se puede ver como si $K = n$ (cuanto más variables se añadan, más alto es el coeficiente), entonces se obtiene que $R^2 = 1$ y se estaría interpolando la muestra con residuos iguales a $\mathbf{0}$

$$R^2 = \frac{Regression\ sum\ of\ squares}{Total\ sum\ of\ squares}$$

- Por lo tanto, R^2 solo es útil para comparar modelos con el mismo número de variables explicativas, pero no modelos con diferente número. Esto hace que sea necesario tener medidas de bondad del ajuste que solo incrementen cuando las variables añadidas mejoren el modelo (que penalicen modelos grandes)

- Para ello, se puede definir el coeficiente de determinación ajustado $\bar{R}^2$, el cual se define de la siguiente manera:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-K} = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-K)}{S_{yy}^2/(n-1)}$$

$$\Rightarrow SSR = \frac{S_{yy}^2}{n-1}(1 - \bar{R}^2)$$

  - El grado de multicolinearidad es el grado de dependencia lineal entre las variables explicativas dentro del modelo. Tener multicolinearidad en el modelo es inevitable, pero tener un alto nivel de multicolinearidad puede ser complicado

    - Esto se debe a que hace la interpretación del modelo ajustado muy difícil e incrementa la varianza de cada coeficiente estimado y la varianza de la predicción

    - Además, si la dependencia entre variables explicativas es muy fuerte, el cálculo de la inversa de $\boldsymbol{X}'\boldsymbol{X}$ puede ser complicado, y normalmente hará que se tengan que eliminar algunas variables del modelo

    - Es por eso que a veces, en un experimento controlado, se escogen unos valores de $\boldsymbol{X}$ tales que $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ se convierte en una matriz diagonal (no hay dependencia entre variables)

    - El rango completo elimina la multicolinearidad perfecta (que una variable explicativa sea función lineal de otra), pero no la multicolinearidad común

- Del mismo modo que antes, es posible derivar las distribuciones muestrales de los estimadores para poder realizar inferencias y obtener intervalos de predicción

  - Debido a que $\boldsymbol{b}$ es una combinación lineal de $\boldsymbol{y}$ y se asume el modelo condicional normal (extendido para el caso de múltiples regresores), entonces $\boldsymbol{b}$ también sigue una distribución normal

$$\boldsymbol{b} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

- Igual que antes, se puede demostrar que la esperanza y la varianza (condicional) de esta distribución son $\boldsymbol{\beta}$ y $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, respectivamente

$$E(\boldsymbol{b}|\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\beta}$$

$$Cov(\boldsymbol{b}|\boldsymbol{X}) = E[(\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})'|\boldsymbol{X}] =$$

$$= E\left[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta})\left((\boldsymbol{\beta}'\boldsymbol{X}'+\boldsymbol{\varepsilon}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}-\boldsymbol{\beta}'\right)|\boldsymbol{X}\right]$$

$$= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}]$$

$$= \sigma^2 E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- Por lo tanto, las distribuciones marginales (la distribución para cada coeficiente) es normal y tiene como esperanza su coeficiente poblacional y la varianza del estimador b_i

$$b_i \sim N\left(\beta_i, \sqrt{\sigma^2_{b_i}}\right)$$

- Como no se sabe σ^2, esta se estima con S_R^2, la cual tiene la expresión vista anteriormente:

$$S_R^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-K} = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-K}$$

- Como se vio anteriormente en el modelo lineal simple, es posible construir un intervalo de confianza para cada estimador. En el caso de la regresión lineal múltiple, el intervalo de confianza individual para cada estimador es el siguiente:

$$b_i - t_{n-K,\frac{\alpha}{2}}S^2_{b_i} \leq \beta_i \leq b_i + t_{n-K,\frac{\alpha}{2}}S^2_{b_i}$$

  - En este caso, la distribución t-Student tiene $n-K$ grados de libertad

  - Si $n > 20$ y el intervalo tiene un nivel de confianza del 95%, entonces se puede sustituir si $t_{n-K,\alpha}$ por 2

- Los contrastes para un solo coeficiente son los mismos que se han visto anteriormente en el caso del modelo de regresión lineal simple, solo se tiene que tener en cuenta que la distribución es una t-Student con $n-K$ grados de libertad (y sus valores críticos para cada nivel de significación)

$$\begin{cases} H_0: \beta_k = \beta_k^0 \\ H_1: \beta_k \neq \beta_k^0 \end{cases} \Rightarrow \quad t_k = \frac{b_k - \beta_k^0}{\sqrt{s^2 S^{kk}}}$$

$$where \;\; S^{kk} = k^{th} \; diagonal \; elem. \; of \; (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- Asumiendo que β_k es igual a β_k^0, el estadístico sigue una distribución t-Student con $n - K$ grados de libertad

- Si se quiere contrastar si todos los coeficientes son nulos (la esperanza condicional de y_i es constante y no depende de las x) o no (hay al menos una variable x que explica y_i), es posible utilizar un estadístico F

  - Bajo la hipótesis nula, el siguiente estadístico sigue una distribución F con $K - 1$ y $n - K$ grados de libertad (el mismo estadístico que en la tabla ANOVA)

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{n - K}{K - 1} \sim F_{K-1,n-K}$$

- La hipótesis lineal general de un conjunto de J restricciones de un modelo de regresión lineal $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ se puede escribir de la siguiente manera:

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1K}\beta_K &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2K}\beta_K &= q_2 \\ \cdots & \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{JK}\beta_K &= q_J. \end{aligned}$$

  - El caso general se puede escribir en notación matricial como $\boldsymbol{R\beta} = \boldsymbol{q}$, donde cada fila de $\boldsymbol{R}$ son los coeficientes de una de las restricciones. Típicamente, $\boldsymbol{R}$ solo tendrá unas pocas filas y muchos ceros en cada fila

A set of the coefficients sum to one, $\beta_2 + \beta_3 + \beta_4 = 1$,

$$\mathbf{R} = [0 \quad 1 \quad 1 \quad 1 \quad 0 \quad \cdots]; \mathbf{q} = 1.$$

A subset of the coefficients are all zero, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$,

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} = [\mathbf{I} \mid \mathbf{0}]; \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Several linear restrictions, $\beta_2 + \beta_3 = 1$, $\beta_4 + \beta_6 = 0$, and $\beta_5 + \beta_6 = 0$,

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}; \quad \mathbf{q} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

All the coefficients in the model except the constant term are zero,

$$\mathbf{R} = [\mathbf{0} | \mathbf{I}_{K-1}]; \quad \mathbf{q} = \mathbf{0}.$$

- La matriz $\boldsymbol{R}$ tiene K columnas (para coincidir con las filas de $\boldsymbol{\beta}$), J filas para un total de J restricciones, y tiene rango completo de filas, de modo que J tiene que ser menor o igual a K

- No obstante, $K \neq J$, ya que, de no ser así, la matriz sería cuadrada e invertible y $\boldsymbol{\beta} = \boldsymbol{R^{-1}q}$ (no habría problema de estimación o inferencia porque se ha dado un valor para cada parámetro), haciendo que $J < K$. La restricción $\boldsymbol{R\beta} = \boldsymbol{q}$ impone J restricciones en los K parámetros, haciendo que hayan $K - J$ parámetros libres

- Para poder extender estos métodos a restricciones no lineales, se plantea una hipótesis no lineal general, que involucra un conjunto $\boldsymbol{c(\beta)}$ de J restricciones no lineales posibles de $\boldsymbol{\beta}$

$$\boldsymbol{c(\beta) = q}$$

  - La contraparte de los requerimientos vistos para el caso especial en donde $\boldsymbol{c(\beta)}$ es no lineal es que $J < K$ y que la matriz jacobiana de estas restricciones (formada por el gradiente de $\boldsymbol{c(\beta)}$ con respecto a cada uno de los parámetros en $\boldsymbol{\beta}$) tenga rango de fila completo

  - En el caso lineal, la jacobiana equivale a $\boldsymbol{R}$ y la independencia funcional que se requiere sería equivalente a la independencia lineal

- Considerando que se quiere contrastar un conjunto de J restricciones lineales en la hipótesis nula contras las de la hipótesis alternativa, se puede utiliza el criterio de Wald

$$\begin{cases} H_0: \boldsymbol{R\beta - q = 0} \\ H_1: \boldsymbol{R\beta - q \neq 0} \end{cases}$$

- Dado el estimador de mínimos cuadrados $\boldsymbol{b}$, el interés se centra en el vector de discrepancia $\boldsymbol{Rb} - \boldsymbol{q} = \boldsymbol{m}$. Es improbable que $\boldsymbol{m} = \boldsymbol{0}$, pero es razonable preguntarse si la desviación se atribuye a la variabilidad muestral o si es significativa

- Como $\boldsymbol{b}$ sigue una distribución normal y $\boldsymbol{m}$ es una función lineal de $\boldsymbol{b}$, $\boldsymbol{m}$ también se distribuye normalmente. Por lo tanto, si la hipótesis nula es verdad, entonces se obtienen los siguientes resultados:

$$E(\boldsymbol{m}|\boldsymbol{X}) = \boldsymbol{R}E(\boldsymbol{b}|\boldsymbol{X}) - \boldsymbol{q} = \boldsymbol{R\beta} - \boldsymbol{q} = \boldsymbol{0}$$

$$Var(\boldsymbol{m}|\boldsymbol{X}) = \boldsymbol{R}Var(\boldsymbol{b}|\boldsymbol{X})\boldsymbol{R}' = \boldsymbol{R}[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}'$$

$$\Rightarrow \boldsymbol{Rb} \sim N(\boldsymbol{q}, \boldsymbol{R}[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}')$$

- Condicionando en $\boldsymbol{X}$, se puede comprobar que el estadístico W sigue una distribución χ_J^2 si la hipótesis nula es cierta, y así ver como valores grandes del estadístico indican que la hipótesis nula es rechazable (la distancia entre $\boldsymbol{b}$ y $\boldsymbol{\beta}$ es muy grande)

$$W = \boldsymbol{m}'Var(\boldsymbol{m}|\boldsymbol{X})^{-1}\boldsymbol{m} =$$

$$= (\boldsymbol{Rb} - \boldsymbol{q})'[\boldsymbol{R}[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}']^{-1}(\boldsymbol{Rb} - \boldsymbol{q}) \sim \chi_J^2$$

- Este estadístico, no obstante, no se puede usar porque depende de σ^2, por lo que se puede obtener un estadístico F haciendo esa sustitución y dividendo por los grados de libertad

$$F = \frac{W}{J}\frac{\sigma^2}{s^2} = \frac{(\boldsymbol{Rb} - \boldsymbol{q})'[\boldsymbol{R}[s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}']^{-1}(\boldsymbol{Rb} - \boldsymbol{q})}{J} \sim F_{J,n-K}$$

- Para comprobar una sola restricción lineal de la forma $\boldsymbol{r}'\boldsymbol{\beta} = q$, el estadístico F será el siguiente:

$$H_0: r_1\beta_1 + r_2\beta_2 + \cdots + r_K\beta_K = \mathbf{r}'\boldsymbol{\beta} = q$$

$$F[1, n-K] = \frac{\left(\sum_j r_j b_j - q\right)^2}{\sum_j \sum_k r_j r_k Est.Var.\left(b_j, b_k\right)}$$

- Para contrastar la hipótesis de que el coeficiente j es igual a un valor particular, entonces $\boldsymbol{R}$ tiene una sola fila con un 1 en la posición j, $\boldsymbol{R}[s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{R}'$ es el elemento diagonal j de la matriz de varianzas y covarianzas estimada y $\boldsymbol{Rb} - \boldsymbol{q}$ es $\left(b_j - q\right)$. El estadístico, por tanto, será el siguiente:

$$F[1, n-K] = \frac{\left(b_j - q\right)^2}{Est.Var.\left(b_j\right)}$$

- Considerando un enfoque alternativo, se puede plantear una estimación muestral $\boldsymbol{r'b} = \hat{q}$ para $\boldsymbol{r'\beta} = q$. Si $\hat{q}$ difiere significativamente de q, se puede concluir que los datos muestrales no son consistentes con la hipótesis nula, de modo que se puede usar el estadístico t

$$t = \frac{\hat{q} - q}{\sqrt{Est.Var(\hat{q}|\boldsymbol{X})}}$$

$$where\ \ Est.Var(\hat{q}|\boldsymbol{X}) = \boldsymbol{r'}[s^2(\boldsymbol{X'X})^{-1}]\boldsymbol{r}$$

- Existe una relación útil entre el estadístico t y el F, ya que es posible escribir el cuadrado del estadístico t en términos del estadístico F cuando hay una sola restricción

$$t^2 = \frac{(\hat{q} - q)^2}{Est.Var(\hat{q} - q|\boldsymbol{X})} =$$

$$= \frac{(\boldsymbol{r'b} - q)'[\boldsymbol{r}[s^2(\boldsymbol{X'X})^{-1}]\boldsymbol{r'}]^{-1}(\boldsymbol{r'b} - q)}{1} \sim F_{1,n-K}$$

- Finalmente, también se puede utilizar el modelo de regresión múltiple para predecir valores puntuales o construir intervalos de confianza

  - Para poder obtener una predicción puntual, solo es necesario sustituir $\boldsymbol{X}$ por $\boldsymbol{x_0}$ en el modelo ajustado, de modo que se dará una predicción $\boldsymbol{y_0}$ para los valores concretos $\boldsymbol{x_0}$

  - El intervalo de confianza para $E(\boldsymbol{y}|\boldsymbol{X})$ de nivel de confianza $1 - \alpha$ se construye de manera similar, solo que la distribución t-Student tendrá $n - K$ grados de libertad

$$\boldsymbol{x_0'b} \pm t_{n-K,\frac{\alpha}{2}} S_R \sqrt{\boldsymbol{x_0'}(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}}$$

  - El intervalo de confianza para $\boldsymbol{y}(\boldsymbol{x_0})$ de nivel de confianza $1 - \alpha$ se construye de manera similar, solo que la distribución t-Student tendrá $n - K$ grados de libertad

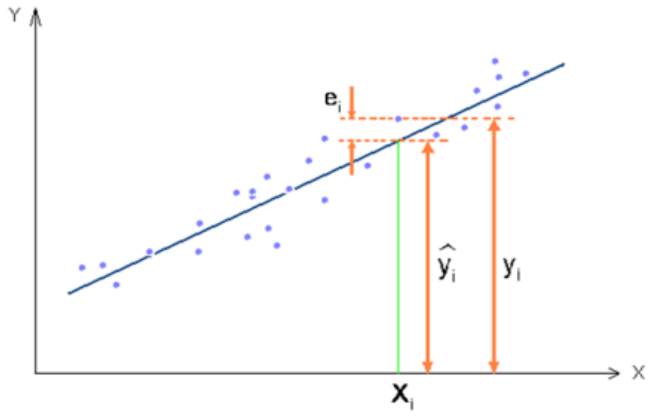
$$\boldsymbol{x_0'b} \pm t_{n-K,\frac{\alpha}{2}} S_R \sqrt{1 + \boldsymbol{x_0'}(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}}$$

El análisis de los residuos: normalidad, homocedasticidad y no linealidad

- Antes de poder realizar inferencias sobre los parámetros de la regresión, es necesario estar seguro que las suposiciones del modelo lineal se cumplen, por lo que se tiene que hacer un análisis de los residuos

    - Para poder inferenciar se han hecho suposiciones sobre las perturbaciones o errores, definidos como $\varepsilon_i = y_i - (\alpha + \beta x_i)$. Estos son parte del modelo teórico y no son observables

        - Por lo tanto, se necesita utilizar estimadores de estos. Para ello, se utilizan los residuos del modelo de regresión ajustado, definidos como $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$



        - Los residuos son aquella parte de la respuesta que no se explica por el modelo, aunque en el contexto de predicción se pueden interpretar como el error de predicción

        - Los residuos tienen información sobre la relación entre x e y contenida en los datos que el modelo ajustado no tiene. Si el modelo es correcto, toda la información de la relación entre x e y debería estar capturada, y, por tanto, los residuos no deberían contener ninguna información sobre la relación y no deberían mostrar ningún patrón significativo

        - Si se encuentra cualquier patrón al analizar los residuos, entonces las suposiciones del modelo no se cumplen, pero se puede utilizar esta información obtenida para mejorar el modelo

    - El análisis de residuos se realiza para verificar las hipótesis del modelo, para sugerir mejoras al modelo o proponer modelos alternativos y para detectar observaciones atípicas (poco explicadas por el modelo y con mucha influencia en la regresión)

- Las cuatro hipótesis que implican modelos como el condicional normal son que $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $\varepsilon_i \sim N(0, \sigma^2)$ para toda i y que $Cov(\varepsilon_i, \varepsilon_j) = 0$ para toda $i \neq j$

- Se utilizan los residuos observados para poder contrastar las hipótesis hechas para los errores (los cuales no son observables)

- Debido a que ε y $\hat{\varepsilon}$ son diferentes, uno no puede esperar que tengan exactamente la misma distribución, aunque el modelo sea válido

ε_i	e_i
Unknown	Known
Linearity	
$E(\varepsilon_i) = 0$	$E(e_i) = 0$
Constant variance	
$V(\varepsilon_i) = \sigma^2$	$V(e_i) = \sigma^2 \cdot (1 - h_{ii})$
Normality	
$N(0, \sigma)$	$N(0, \sigma_{e_i})$
Independence	
Yes	No

- En el caso de la varianza de los residuos $\hat{\varepsilon}$, esta dependerá de las observaciones de x a través de un término llamado h_{ii}, el cual mide la distancia entre x_i y $\overline{x}$

$$Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}) \quad where \quad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

- El término h_{ii} cumple dos propiedades que permiten derivar una expresión para la desviación estándar de los residuos y para el estimador de esta desviación

$$\sum_{i=1}^{n} h_{ii} = \sum_{i=1}^{n} \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} = \frac{n}{n} + \frac{S_{xx}}{S_{xx}} = 2 \quad \Rightarrow \quad \overline{h} = \frac{2}{n}$$

$$\sigma_{\hat{\varepsilon}_i} = \sigma\sqrt{1 - h_{ii}} \quad \Rightarrow \quad \hat{\sigma}_{\hat{\varepsilon}_i} = S_R\sqrt{1 - h_{ii}}$$

- Normalmente se supone que h_{ii} es muy pequeño, de modo que se puede hacer la aproximación $\sigma^2_{\hat{\varepsilon}_i} \approx \sigma^2$. Por lo tanto, la varianza residual estimada, $\hat{\sigma}^2_{\hat{\varepsilon}_i}$, es aproximadamente constante y similar a S_R^2

- Si el modelo es correcto, entonces la distribución de los residuos es muy similar a la distribución de los errores y tiene sentido verificar las hipótesis con los residuos en este caso. Para ello, se utilizan dos tipos de residuos: los residuos comunes y los residuos estandarizados

    - Los residuos comunes $\hat{\varepsilon}_i$ son residuos expresados en las mismas unidades que la respuesta. Cuanto mayor sea el residuo, más pobre es la explicación de la observación por el modelo, pero con este tipo de residuos no hay manera de saber si un residuo es muy grande o muy pequeño

    $$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

    - Los residuos estandarizados $\hat{\varepsilon}_i^*$ son residuos sin escala, de modo que permite hacer comparaciones de tamaño. Los residuos estandarizados siguen una distribución t-Student con $n-2$ grados de libertad

    $$\hat{\varepsilon}_i^* = SRES = \frac{\hat{\varepsilon}_i - \bar{\varepsilon}}{\hat{\sigma}_{\hat{\varepsilon}_i}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{\hat{\varepsilon}_i}} \qquad \hat{\varepsilon}_i^* \sim t_{n-2}$$

    - Un valor absoluto $|\hat{\varepsilon}_i^*| > 2$ indica que la observación es atípica, dado que se explica peor que la mayoría de observaciones en el modelo. Esto puede ser debido a la observación, al modelo o a ambas, y la mayoría de veces esto se soluciona añadiendo variables o transformándolas

- Los residuos tienen que graficarse para poder descubrir si hay información escondida que puede usarse para mejorar el modelo para Y, aunque diferentes problemas requieren diferentes enfoques gráficos

    - Hay tres tipos de gráficos que son muy usados: los de residuos contra valores ajustados, los *Q-Q plots* y los de residuos contra variables explicativas fuera del modelo

        - Un gráfico de residuos contra valores ajustados es un gráfico cuyos ejes son los valores $\hat{y}$ obtenidos por el modelo ajustado y los valores de loss residuos (ya sean los comunes o los estandarizados)

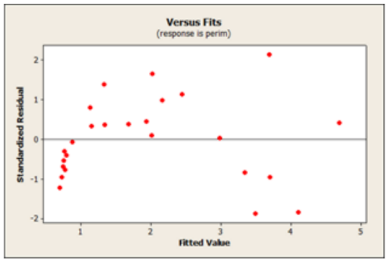
        - Un *Q-Q plot* es un gráfico que permite diagnosticar diferencias entre la distribución de probabilidad de una muestra y la de una distribución cualquiera, de modo que se comparan los cuartiles

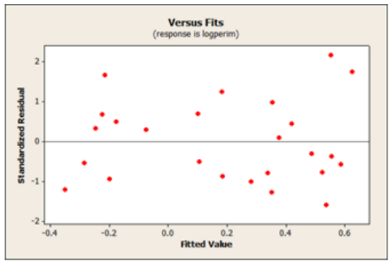
        - Un gráfico de residuos contra variables explicativas fuera del modelo es simplemente un gráfico con los valores de los

residuos en un eje y con valores de una variable explicativa fuera del modelo

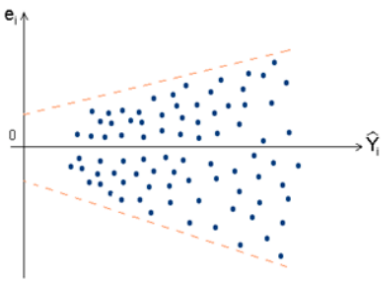
- A través del gráfico de residuos contra valores ajustados es posible detectar violaciones de la primera y de la segunda hipótesis, planteadas anteriormente, y detectar observaciones atípicas

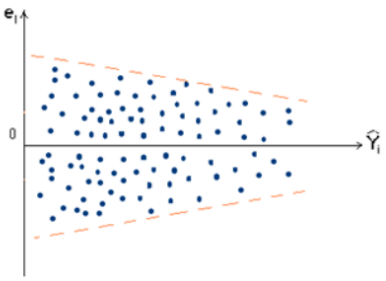
  - Es posible detectar no linealidad (violación de la primera hipótesis) través de un gráfico de residuos contra valores ajustados, ya que tendría no tendría que haber ningún patrón para un buen ajuste. Para poder solucionar este problema, se puede ajustar un modelo diferente que permita hacer un mejor ajuste





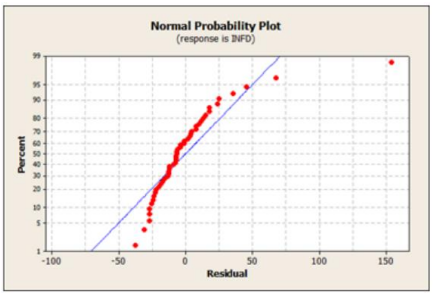
  - También es posible detectar heterocedasticidad (violación de la homocedasticidad, la segunda hipótesis) a partir del mismo tipo de gráfico, ya que muestra la amplitud de la dispersión. Para poder solucionar este problema, se puede transformar la variable respuesta o utilizar un criterio de ajuste de mínimos cuadrados ponderados

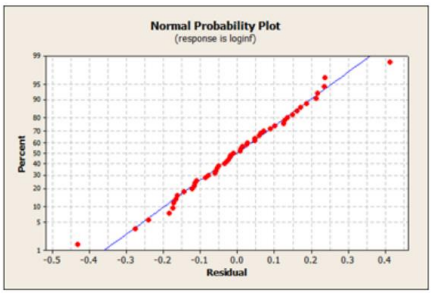




  - Obviamente, también es posible detectar la violación de ambas hipótesis a la vez (de la primera y la segunda) y de valores atípicos (a través de utilizar residuos estandarizados)

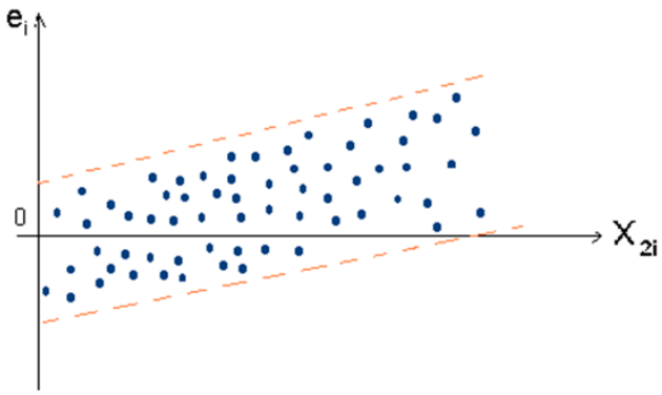
- Para poder verificar la hipótesis de normalidad de las perturbaicones o errores, normalmente se utiliza un *Q-Q plot* para la distribución normal





- Si los residuos para el ajuste concuerdan con los cuartiles que se tendrían que tener si siguieran una distribución normal, entonces los residuos se distribuyen aproximadamente de manera normal. En cambio, una diferencia relativamente grande haría que se concluya lo contrario (violación de la tercera hipótesis)

- Finalmente, a través de un gráfico de los residuos contra una variable predictora diferente de x fuera del modelo se puede comprobar si los errores están correlacionados



- Si se detecta un patrón, entonces la variable fuera del modelo tiene correlación con los residuos y, por tanto, ayudan a predecir y. Esto, además, quiere decir que los residuos tienen una correlación entre sí, dado que se ven afectados por la misma variable subyacente (violación de la cuarta hipótesis)

- Si no se detecta ningún patrón claro, entonces la variable no ayuda a explicar y y eso hace que tampoco haya correlación entre los errores

- El objetivo principal del análisis gráfico de los residuos es poder juzgar si se cumplen las hipótesis hechas sobre los errores en el modelo y poder ver si el modelo ajustado es correcto. No obstante, el análisis cuantitativo permite analizar las observaciones atípicas y medir su grado de anormalidad, su distancia en el espacio de las x y el grado de influencia en el modelo

- Para poder medir el grado de anormalidad se utilizan los residuos estandarizados c, y concretamente, se estudian las observaciones con $|\hat{\varepsilon}_i| > 2$, dado que son aquellas mal explicadas por el modelo y permiten obtener información útil

    - Si el modelo es correcto, se espera que el 5% de las observaciones tengan un valor $|\hat{\varepsilon}_i| > 2$

    - Como ya se ha mencionado, esto se puede deber a la observación, al modelo, o a ambas. Además, a veces esto indica un error en el valor de la observación, si el error puede ser corregido, entonces se corrige, pero si no, entonces se puede eliminar de la muestra

    - Viendo el valor absoluto de los residuos estandarizados con $|\hat{\varepsilon}_i| > 2$ es posible obtener una medida del grado de anormalidad de esta observación (dado que no hay diferencias de escala)

- Para poder medir la distancia en el espacio de las x se puede utilizar la medida h_{ii}, la cual mide la distancia entre x_i y el centro de gravedad de todas las x en la muestra usada para ajustar el modelo (una media muestral de todos los valores de todas estas variables x)

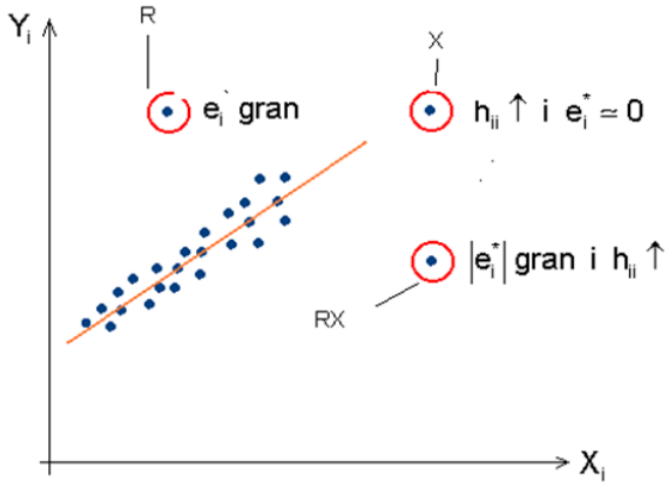
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \;\Rightarrow\; \sum_{i=1}^{n} h_{ii} = k = 2 \;\Rightarrow\; \bar{h} = \frac{k}{n} = \frac{2}{n}$$

$$where\ \ k = n^{\underline{o}}\ of\ \ expl.var.$$

    - Esta medida se puede extender para el caso multivariante, en cuyo caso se añade

    - El criterio que se utiliza normalmente para clasificar las observaciones como atípicas con h_{ii} es de prestar atención a observaciones cuya h_{ii} es mayor a $3k/n$

    - Es necesario combinar el primer criterio de los residuos estandarizados con este, dado que hay veces que la medida no permite discernir que tan atípica es en términos de distancia de la línea de regresión y del espacio de x (desde el origen)



- Una medida para poder medir el grado de influencia de una observación en el modelo es la distancia de Cook, la cual se calcula de la siguiente manera:

$$DC_i = \frac{(\boldsymbol{b} - \boldsymbol{b}^*)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{b}^*)}{kS_R^2} = \hat{\varepsilon}_i^{*2}\frac{h_{ii}}{1 - h_{ii}}\frac{1}{k} \sim F_{k,n-k}$$

  - Se considera que una observación (x_i, y_i) tiene mucha influencia cuando el vector de coeficientes estimados $\boldsymbol{b}$ para el modelo ajustado usando todas las n observaciones es muy diferente al vector $\boldsymbol{b}^*$ que se obtendría si no se tuviera en cuenta esta. En este caso, $\boldsymbol{X}$ es una matriz $n \times k$ con los valores de las variables x

  - Debido a que el estadístico sigue una distribución $F_{k,n-k}$, un criterio normalmente usado es el de ver si el valor supera a su punto mediano $F_{k,n-k,0.5}$. Como este está muy cerca de 1, se suele simplificar el criterio y solo se mira si $DC_i > 1$

  - No obstante, esta distancia no siempre permite detectar correctamente observaciones con mucha influencia

El análisis de la varianza

- El método ANOVA es un método para estimar las medias de varias poblaciones, las cuales normalmente se asumen que están normalmente distribuidas, pero su motivación es una cuestión de diseño estadístico

  - El ANOVA clásico tenía el contraste de hipótesis como su objetivo principal (contrastando la llamada hipótesis nula del ANOVA), pero los experimentadores se han dado cuenta que contrastar con un contraste de hipótesis no hace que la inferencia experimental sea buena

- Por lo tanto, aunque se derivara el contraste de hipótesis para la hipótesis nula del ANOVA, esto está lejos de la parte más importante del análisis de la varianza

- La parte importante se refiere a la estimación, tanto puntual como interválica, y a la inferencia basada en contrastes

- En el ANOVA de una vía o *oneway ANOVA* (también llamada clasificación de una vía) asume que los datos Y_{ij} se observan acorde al siguiente modelo, donde θ_i son parámetros desconocidos y ε_{ij} son errores aleatorios:

$$Y_{ij} = \theta_i + \varepsilon_{ij} \quad for \quad i = 1,2,\dots,k \quad \& \quad j = 1,2,\dots,n_i$$

  - Esquemáticamente, los datos para un ANOVA de una vía se verían de la siguiente manera, en donde no se asume que hay el mismo número de observaciones para cada grupo de tratamiento:

Treatments				
1	2	3	...	k
y_{11}	y_{21}	y_{31}	⋯	y_{k1}
y_{12}	y_{22}	y_{32}	⋯	y_{k2}
⋮	⋮	⋮	⋯	y_{k3}
		y_{3n_3}		⋮
y_{1n_1}				
	y_{2n_2}			y_{kn_k}

  - Sin pérdida de generalidad, se puede asumir que $E(\varepsilon_{ij}) = 0$ (como en los modelos de regresión, dado que si no se podría absorber la media restante de esta en θ_i) y por tanto ver que θ_i son las medias de las Y_{ij}. Estas normalmente se denominan medias de tratamiento, dado que el subíndice i corresponde a los diferentes tratamientos o niveles de un tratamiento particular

$$E(Y_{ij}) = \theta_i \quad for \quad j = 1,2,\dots,n_i$$

  - Alternativamente, se puede considerar un modelo sobreparametrizado, el cual se basa en definir el término θ_i como $\mu + \tau_i$ y en donde $E(\varepsilon_{ij}) = 0$

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad for \quad i = 1,2,\dots,k \quad \& \quad j = 1,2,\dots,n_i$$

$$\Rightarrow E(Y_{ij}) = \mu + \tau_i \quad for \quad j = 1,2,\dots,n_i$$

- En la formulación alternativa, μ se interpreta como el nivel medio común de todos los tratamientos, mientras que los parámetros τ_i denotan el efecto único debido al tratamiento i (la desviación de la media causada por el tratamiento)

- No obstante, los dos elementos de la formulación alternativa no se pueden estimar separadamente por problemas de identificación

  - Un parámetro θ para una familia de distribuciones $\{f(\boldsymbol{y}|\theta) : \theta \in \Theta\}$ es identificable si los valores distintos de θ corresponden a distintas funciones de densidad de probabilidad o de masa de probabilidad. Eso quiere decir que si $\theta \neq \theta'$, entonces $f(\boldsymbol{y}|\theta)$ no es la misma función de $\boldsymbol{y}$ que $f(\boldsymbol{y}|\theta')$

  - La identificación es una propiedad del modelo, no de un estimador o procedimiento de estimación. Sin embargo, si el modelo no es identificable, entonces hay una dificultad para hacer una inferencia, aunque los problemas de identificación normalmente pueden ser resueltos redefiniendo el modelo

  - En la parametrización del modelo alternativo, hay $k + 1$ parámetros, pero solo hay k medias, por lo que, sin ninguna restricción en los parámetros, más de un conjunto de valores para $(\mu, \tau_1, \tau_2, \dots, \tau_k)$ haría que se obtuviera la misma distribución. En consecuencia, es común añadir la restricción de que $\sum_{i=1}^{k} \tau_i = 0$, lo cual efectivamente reduce el número de parámetros a k y hace que el modelo sea identificable

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad where \quad \sum_{i=1}^{k} \tau_i = 0$$

$$for \quad i = 1,2,\dots,k \quad \& \quad j = 1,2,\dots,n_i$$

  - Esta restricción hace que las τ_i tengan una interpretación como desviaciones de la media común. No obstante, si los grupos no están balanceados, normalmente se pondera esta suma por el número de observaciones de cada grupo n_i, quedando de la siguiente manera:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad where \quad \sum_{i=1}^{k} n_i \tau_i = 0$$

$$for \quad i = 1,2,\dots,k \quad \& \quad j = 1,2,\dots,n_i$$

- La terminología común que se suele utilizar en este modelo es la siguiente:

  - Las observaciones de llaman réplicas. Cuando el número de réplicas es igual para cada grupo o tratamiento, se dice que se está en un diseño experimental balanceado

  - La fuente de variación controlada θ_i (o τ_i en la formulación alternativa) se denomina factor. Los diferentes valores que puede tomar este factor se denominan niveles

- Como se puede ver en la formulación alternativa, el modelo del ANOVA es como un modelo para variables predictoras categóricas (una regresión de Y_{ij} sobre τ_i)

  - En este caso, μ sería el intercepto, mientras que τ_i sería una variable categórica que toma un valor diferente dependiendo del grupo o categoría

  - Los parámetros de estos modelos se pueden obtener mediante mínimos cuadrados o mediante estimación por máxima verosimilitud

- Bajo el modelo alternativo, una suposición mínima que es necesaria antes de cualquier estimación es que $E(\varepsilon_{ij}) = 0$ y que $Var(\varepsilon_{ij}) < \infty$ para toda i y j. No obstante, es necesario hacer suposiciones de las distribuciones para poder estimar intervalos de confianza, de modo que se presentan las suposiciones clásicas del ANOVA

  - Las variables aleatorias Y_{ij} son observadas acorde al siguiente modelo, el cual cumple las suposiciones explicadas:

$$Y_{ij} = \theta_i + \varepsilon_{ij} \quad for \;\; i = 1,2,\dots,k \;\; \& \;\; j = 1,2,\dots,n_i$$

    - Se asume que $E(\varepsilon_{ij}) = 0$ y $Var(\varepsilon_{ij}) < \infty$ para toda i y j y, además, $Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$ para toda i, i', j, j'

    - Se asume que los errores se distribuyen normalmente y son independientes, de modo que $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ y $Y_{ij} \sim N(\theta_i, \sigma_i^2)$

    - Se asume que hay homocesdasticidad, de modo que $\sigma_i^2 = \sigma^2$ y eso hace que $\varepsilon_{ij} \sim N(0, \sigma^2)$ y $Y_{ij} \sim N(\theta_i, \sigma^2)$

  - Sin la segunda suposición, solo se podría hacer una estimación puntual y buscar posibles estimadores que minimicen la varianza dentro de una

clase, pero no se podría hacer un intervalo de confianza ni un contraste de hipótesis

- Si se asume alguna distribución diferente de la normal, los intervalos y los contrastes pueden ser difíciles de obtener (pero posibles). Con tamaños muestrales razonables y con poblaciones no muy asimétricas, el teorema del límite central apoya la suposición de normalidad

- La suposición de igualdad de la varianza también es importante, dado que está vinculada con la suposición de normalidad también

- En general, si se sospecha que los datos violan las suposiciones del modelo ANOVA, entonces un primer intento para solucionar el problema es hacer una transformación no lineal

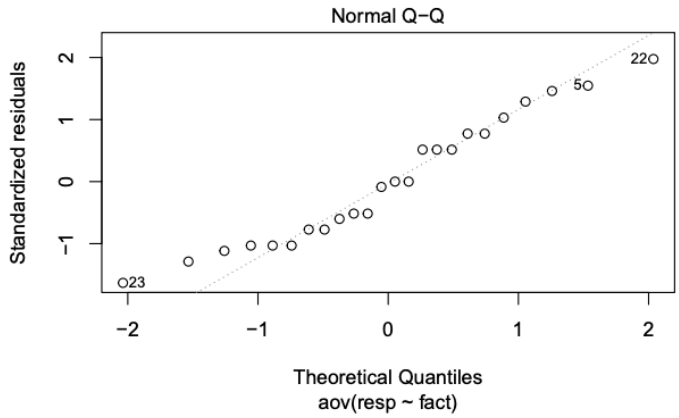
  - Una investigación de 1984 de Box muestra que la robustez del ANOVA a la suposición de normalidad depende de que tan iguales son las varianzas. Dados tamaños entre grupos iguales, la varianza más grande no debería ser tres veces superior a la varianza más pequeña

  - Además, también es importante el número de observaciones en las poblaciones de las cuales se obtiene la muestra. Los diseños balanceados pueden ayudar a mitigar el efecto de la heterocedasticidad

- No obstante, se pueden hacer algunas comprobaciones antes de implementar el ANOVA para ver si es adecuado utilizar este análisis

  - Para poder comprobar la suposición de normalidad de los errores, se pueden utilizar los siguientes métodos:

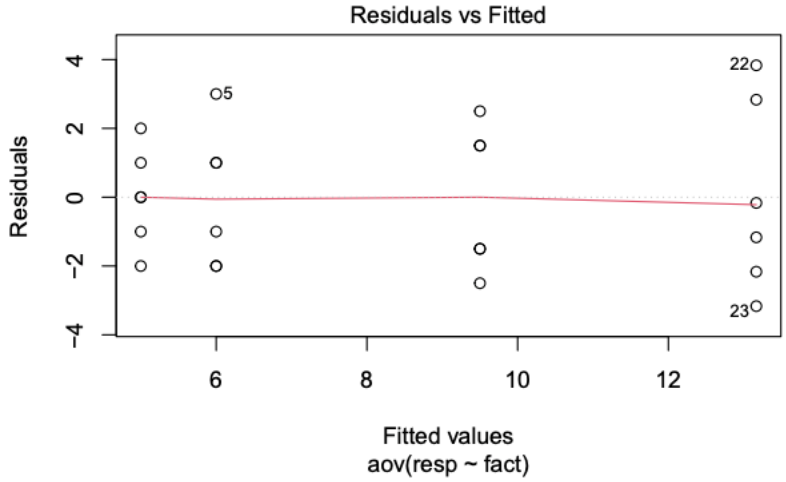
    - Se pueden utilizar *Q-Q plots* para poder comparar los residuos estandarizados con los cuantiles de la distribución normal. Si estos coinciden aproximadamente, entonces los errores se distribuyen aproximadamente de manera normal



- También es posible utilizar contrastes de normalidad, tales como el de Wilk-Shapiro, contrastes de bondad del ajuste o diagramas de caja con tal de comprobar la suposición de normalidad

- Para poder comprobar la suposición de homocedasticidad de los errores, se pueden utilizar los siguientes métodos:

  - Es posible utilizar diagramas en donde se muestran los residuos y los valores ajustados por el modelo



  - También es posible utilizar contrastes para la homogeneidad de la varianza (cuya hipótesis nula es la igualdad de varianzas entre grupos), tales como los contrastes de Bartlett, de Hartley, de Cochran o de Levene

- Los términos de errores de las observaciones deberían ser independientes entre si, de modo que se debería mirar el diseño y la recolección de datos, dado que no se puede compensar después

  - Una correlación positiva entre errores dentro de los grupos resulta en una subestimación de la varianza verdadera, de modo

que la probabilidad de rechazar incrementa (el porcentaje de errores de tipo I aumenta)

- Una correlación negativa entre errores dentro de los grupos resulta en una sobreestimación de la varianza verdadera, de modo que la probabilidad de rechazar disminuye (el porcentaje de errores de tipo II aumenta)

- Una vez visto el modelo ANOVA, es posible desarrollar las hipótesis de este análisis, los contrastes de hipótesis pertinentes y los intervalos de confianza para poder inferenciar a través de métodos como el de unión-intersección

  - El contraste de hipótesis clásico del contraste ANOVA es un contraste de la hipótesis nula $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$, la cual normalmente no se suele cumplir

    - El experimento se realiza, sin embargo, para encontrar que tratamientos son mejores (cuales tienen una media más alta), por lo que el interés real del ANOVA no es el contraste de hipótesis, sino la estimación

    - La hipótesis alternativa a la ANOVA es que existe alguna media que no es igual, expresada como $H_1 : \theta_i \neq \theta_j$ para $i \neq j$

    - Si H_0 se rechaza, solo se puede concluir que hay alguna diferencia en las θ_i, pero no se puede inferenciar sobre dónde está esta diferencia

  - Un problema con las hipótesis del ANOVA es que la interpretación de estas hipótesis no es fácil. De este modo, sería más interesante realizar una descripción estadística de las θ_i, lo cual se puede hacer separando las hipótesis ANOVA en hipótesis más pequeñas (más fáciles de describir)

    - Ya se han visto maneras de separar las hipótesis en hipótesis menos complicadas y pequeñas: el método de unión-intersección y el método de intersección-unión

    - En el caso del ANOVA, el método de unión-intersección funciona mejor, dado que la hipótesis nula de la ANOVA es la intersección de hipótesis univariantes más fáciles, expresadas en términos de contrastes

  - También es posible derivar el estadístico F del ANOVA a través de obtener el contraste de razón de verosimilitud, el cual es equivalente

- Bajo la hipótesis nula, el espacio paramétrico Θ_0 consiste en un solo parámetro θ que es igual para todos los tratamientos o grupos, por lo que se obtiene la siguiente función de verosimilitud:

$$L_0 = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\frac{-\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu)^2}{2\sigma^2}}$$

- A través de esta función, se pueden obtener los estimadores de máxima verosimilitud bajo la hipótesis nula y usarlos en la función, de modo que se obtienen los siguientes resultados:

$$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{k} n_i} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2}{\sum_{i=1}^{k} n_i}$$

$$\Rightarrow L_0 = \left(\frac{n}{2\pi\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}\right)^2}\right)^{n/2} e^{-n/2}$$

- Bajo la hipótesis alternativa, el espacio paramétrico Θ consiste de todos los parámetros θ_i , los cuales difieren según tratamientos o grupos, por lo que se obtiene la siguiente función de verosimilitud:

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\frac{-\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2}{2\sigma^2}}$$

- A través de esta función, se pueden obtener los estimadores de máxima verosimilitud bajo la hipótesis nula y usarlos en la función, de modo que se obtienen los siguientes resultados:

$$\hat{\mu}_i = \bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2}{N}$$

$$\Rightarrow L = \left(\frac{n}{2\pi\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2}\right)^{n/2} e^{-n/2}$$

- Finalmente, se obtiene la siguiente razón, la cual será el contraste Λ para el contraste de hipótesis. Esta razón será la razón entre la suma de cuadrados dentro de los grupos y la suma cuadrada total, de las cuales se hablará posteriormente

$$\Lambda = \left(\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}\right)^2} \right)^{\frac{n}{2}} = \left(\frac{SSW}{SST}\right)^{\frac{n}{2}}$$

$$\Rightarrow\ R = \{(y_1, y_2, \dots, y_n) | \Lambda \leq k_\alpha\}$$

- El ANOVA proporciona una manera útil de pensar sobre la manera en la que diferentes tratamientos afectan una variable observada: la idea de asignar variación a diferentes fuentes

    - Para cualquier número y_{ij} para $i = 1,2, \dots, k$ y $j = 1,2, \dots, n_i$, se cumple la siguiente igualdad:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2 = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{\bar{y}})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$$

$$where\ \ \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad \& \quad \bar{y}_i = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i \bar{y}_i$$

        - Estas sumas se denominan suma de cuadrados, y se entienden como una medida de variación en los datos asignable a diferentes fuentes

    - Los términos en el modelo ANOVA de una vía corresponden uno a uno con los términos de la descomposición anterior, en donde la varianza asignada a los tratamientos es el primer término y la varianza asignada al error es el segundo término

        - El primer término es la variación entre grupos o tratamientos, la cual mide la variación que depende de los tratamientos (variación al comparar los grupos), llamada variación entre grupos. El segundo término es la varianza que proviene del error dentro de cada grupo (variación entre las observaciones de cada grupo), llamada variación dentro de los grupos

        - El término a la izquierda de la ecuación es la varianza total observada, también llamada suma total de cuadrados, la cual mide la variación sin tener en cuenta la categorización por tratamientos o grupos

        - El hecho de que las fuentes de variación cumplan la identidad muestra que la variación en los datos, medida en sumas de cuadrados, es aditiva de la misma manera que lo es el modelo ANOVA

- Una razón por la que es más fácil lidiar con las sumas de cuadrados es que, bajo normalidad, las sumas de cuadrados siguen una distribución chi cuadradas, de modo que variables aleatorias de este tipo se pueden sumar para obtener una nueva variable con distribución chi cuadrada

    - Bajo las suposiciones ANOVA, en particular si $Y_{ij} \sim N(0, \sigma^2)$, se puede comprobar lo siguiente:

$$\frac{1}{\sigma^2}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 \sim \chi^2_{n_i-1} \quad for\ i = 1,2,\dots,k$$

$$\Rightarrow\ \chi^2_{n_i-1}\ are\ independent\ for\ i = 1,2,\dots,k$$

$$\Rightarrow \sum_{i=1}^{k}\chi^2_{n_i-1} = \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 \sim \chi^2_{N-k}$$

    - Además, si $\theta_i = \theta_j$ para toda i y j, entonces se cumple lo siguiente:

$$\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 \sim \chi^2_{k-1} \qquad \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{\bar{Y}}\right)^2 \sim \chi^2_{N-1}$$

- Entonces, bajo $H_0: \theta_1 = \theta_2 = \cdots = \theta_k$, la partición de la suma de cuadrados es una partición de variables chi cuadradas

    - Cuando estas se escalan, la suma total de cuadrados es una variable que se distribuye χ^2_{N-1}, y la suma de sumas de cuadrados se distribuyen χ^2_{k-1} y χ^2_{N-k} respectivamente

    - Esta partición solo es verdad cuando los términos de la parte derecha de la descomposición de la suma de cuadrados total son independientes entre si, lo cual proviene de la suposición de normalidad del ANOVA

    - La partición de las χ^2 se mantiene para un contexto un poco más general, y una caracterización de esto viene dada por el teorema de Cochrane

- Solo con estos resultados es posible demostrar que, dividiendo las variables que siguen χ^2_{k-1} y χ^2_{N-k} por sus grados de libertad, y obteniendo la *ratio* entre ellas, es posible construir un estadístico que siga una distribución $F_{k-1,N-k}$ como se ha visto anteriormente

$$\frac{\frac{1}{\sigma^2}\frac{\sum_{i=1}^{k} n_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2}{k-1}}{\frac{1}{\sigma^2}\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2}{N-k}} = \frac{\frac{\sum_{i=1}^{k} n_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2}{k-1}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2}{N-k}} = \frac{MSB}{MSW} \sim F_{k-1,N-k}$$

- Como se puede ver, este estadístico se puede resumir como la razón entre el MSB y el MSW, los cuales serán posteriormente definidos

- En este caso, el contraste que se realiza dependiendo de si el valor del estadístico es mayor o igual a un valor crítico $F_{k-1,N-k,\alpha}$

- Lo más común es resumir los resultados del contraste de hipótesis ANOVA en una tabla, la cual se denomina tabla del ANOVA

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Between treatment groups	$k-1$	SSB = $\sum n_i(\bar{y}_i - \bar{\bar{y}})^2$	MSB = SSB/$(k-1)$	$F = \frac{\text{MSB}}{\text{MSW}}$
Within treatment groups	$N-k$	SSW = $\sum\sum(y_{ij} - \bar{y}_{i\cdot})^2$	MSW = SSW/$(N-k)$	
Total	$N-1$	SST = $\sum\sum(y_{ij} - \bar{\bar{y}})^2$		

- A partir de la identidad anterior y de la suma de variables aleatorias chi cuadradas, se puede ver como se pueden sumar los grados de libertad y las sumas de cuadrados

$$SSB + SSW = SST$$

- Para el estadístico F del ANOVA se utiliza el MSB y el MSW porque son estimadores no sesgados de la varianza:

$$E(SSB) = \frac{\sum_{i=1}^{k} n_i\tau_i}{k-1} + \sigma^2 \;\Rightarrow\; E(MSB) = \sigma^2$$

$$E(SSB) = (N-k)\sigma^2 \;\Rightarrow\; E(MSW) = \sigma^2$$

- No obstante, la columna de los cuadrados medios no suma para obtener los cuadrados medios totales, ya que estos serían $SST/(N-1)$

- Una vez visto como estimar y contrastar para un solo contraste en el ANOVA, normalmente se quiere hacer más de una inferencia, pero se sabe que la inferencia simultánea de múltiples contrastes un nivel de confianza α no necesariamente está a un nivel α de confianza, de modo que es necesario utilizar otros métodos

    - Rechazar la hipótesis nula de que todas las medias de los grupos o tratamientos son iguales solo indica que hay al menos un grupo poblacional para el cual la media difiere, pero no indica qué grupos difieren de otros

        - Se pueden hacer comparaciones a pares no planificadas *ex post*, de modo que se comparan los posibles pares de medias de tratamientos para revelar las diferencias entre grupos. Esto se hace cuando no es posible justificar comparaciones específicas sobre otras comparaciones antes del análisis y la recolección de datos

        - Se pueden hacer comparaciones planeadas, las cuales son comparaciones específicas que normalmente están planeadas durante el diseño del experimento

    - Existen varios procedimientos para controlar el *family-wise type I error rate* al hacer comparaciones a pares no planificadas, los cuales minimizan la probabilidad de cometer errores de tipo I

        - Estos procedimientos reducen la potencia de cada comparación a pares individual (se incrementa el error de tipo II), y la reducción de potencia está directamente relacionada con el número de grupos comparándose

        - Cuanto mayor es el número de grupos que se comparan, menor es la potencia del contraste y mayor es la probabilidad de cometer un error de tipo II

        - Es posible diferenciar los contrastes entre contrastes conservadores (en donde la probabilidad de error de tipo I nunca es mayor a la del nivel nominal establecido) y liberales (en donde puede tener un valor mayor)

    - Los procedimientos más comunes para comparaciones a pares no planificadas son los siguientes:

        - El contraste de Scheffé, el cual es muy conservador y no es tan eficiente para comparar todos los pares de medias. Este no está restringido a comparaciones a pares, y se analizará con mayor detalle

- El contraste HSD de Tukey, el cual compara cada media de grupo con cualquier otra media de los otros grupos de par en par y controla el *family-wise type I error rate* a no más que el nivel nominal

- El contraste de diferencia protegida menos significativa de Fisher (o contraste PLSD), el cual se basa en contrastes t a pares usando la MSW para el error estándar

- El contraste de Duncan, el cual es un contraste basado en rangos y diseñado para comparar la media de cada grupo con la de los otros grupos (a pares)

$$\begin{cases} H_0\colon\ \mu_i = \mu_j \quad for\ \ all\ \ i \neq j \\ H_1\colon\ \mu_i \neq \mu_j \quad for\ \ any\ \ (i,j) \end{cases}$$

$$R_p = r_\alpha(p,l)\sqrt{\frac{CM_{error}}{n}}$$

$$D_\alpha(k-1,l) = Dunnett's\ \ comparison\ parameter$$

$$k-1 = treatments\ minus\ control\ group$$

$$l = degrees\ of\ freedom$$

- El contraste de Dunnett, el cual es un contraste t modificado y diseñado específicamente para comparar la media de cada grupo con un único grupo de control (todos los grupos tienen el mismo número de casos). Bajo este escenario, hay menos comparaciones que al hacer comparaciones a pares de todas las medias, por lo que tiene más potencia que otros procedimientos

$$\begin{cases} H_0\colon\ \mu_i = \mu_{cont} \\ H_1\colon\ \mu_i \neq \mu_{cont} \end{cases}$$

$$reject\colon\ \ |\bar{Y}_i - \bar{Y}_{cont}| > D_\alpha(k-1,l)\sqrt{CM_{error}\left(\frac{1}{n_i}+\frac{1}{n_{cont}}\right)}$$

$$accept\colon\ |\bar{Y}_i - \bar{Y}_{cont}| \leq D_\alpha(k-1,l)\sqrt{CM_{error}\left(\frac{1}{n_i}+\frac{1}{n_{cont}}\right)}$$

$$D_\alpha(k-1,l) = Dunnett's\ \ comparison\ parameter$$

$$k - 1 = treatments\ minus\ control\ group$$

$$l = degrees\ of\ freedom$$

- También es posible aplicar ajustes de los *p-values* para múltiples contrastes de diferencia de medias

- Para poder escoger qué tipo de procedimiento utilizar, es necesario considerar el objetivo de la investigación experimental

  - Si el propósito es decidir cuál de un grupo de tratamientos es más probable que tenga un efecto, entonces es mejor utilizar un contraste más liberal como el PLSD, ya que es mejor no perder un posible efecto

  - Si el objetivo, en cambio, es ser lo más acertado posible al decidir si un tratamiento particular tiene un efecto, un procedimiento más conservativo como el de Scheffé sería apropiado

  - Sin embargo, ninguno de estos métodos es un sustituto efectivo para un experimento diseñado específicamente para hacer comparaciones planeadas entre las medias de tratamientos

- Debido a que el contraste de Tukey no es muy liberal ni muy conservador, este se puede utilizar más a menudo

```
HSD Test for resp

Mean Square Error:  4.516667

fact,  means

      resp      std r Min Max
1  6.00000 2.000000 6   4   9
2  5.00000 1.414214 6   3   7
3  9.50000 2.073644 6   7  12
4 13.16667 2.786874 6  10  17

Alpha: 0.05 ; DF Error: 20
Critical Value of Studentized Range: 3.958293

Minimun Significant Difference: 3.434325

Treatments with the same letter are not significantly different.

      resp groups
4 13.16667      a
3  9.50000      b
1  6.00000      c
2  5.00000      c
```

  - Los resultados del contraste serán una clasificación de los tratamientos en la muestra dependiendo de si sus medias se

pueden diferenciar o no. De este modo, cada tratamiento se asigna a una categoría, pudiendo haber más de un tratamiento en cada categoría si es que sus medias no se pueden diferenciar significativamente

- Estas categorías no son excluyentes, de modo que un tratamiento puede pertenecer a dos categorías diferentes. En este caso, eso quiere decir que el tratamiento tiene una media que puede no ser significativamente diferente a las medias de los tratamientos en cada una de las categorías, pero no quiere decir que las medias de las diferentes categorías no sean significativamente diferentes (lo son, por eso se crean diferentes categorías)

- Este análisis solo tiene sentido si se rechaza la hipótesis nula de que todas las medias son iguales. Si no se rechaza, entonces hacer este análisis no tendría sentido porque no habría medias diferentes

- Otra manera de hacer un análisis de varianza es hacer un análisis de varianza en donde los factores o efectos son aleatorios, dado que no es interesante comparar los grupos en sí, sino que se utiliza una muestra genérica para poder inferenciar sobre la población entera

    - En el ANOVA de una vía clásico se tiene un solo factor de tratamiento con varios niveles o grupos y réplicas en cada nivel. En cambio, en el ANOVA de una vía de efectos aleatorios, los niveles o grupos comparados se escogen de manera aleatoria

        - El modelo matemático utilizado en este ANOVA es similar al visto anteriormente con formulación alternativa, pero en este caso se incluye un factor aleatorio A_i

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}$$

        - Las suposiciones básicas que se realizan en este modelo es que tanto A_i como ε_{ij} son variables aleatorias que siguen una distribución normal con varianza diferente y que son independientes entre ellas

$$A_i \sim N(0, \sigma_A^2) \qquad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \qquad Cov\left(A_i, \varepsilon_{ij}\right) = 0$$

$$\Rightarrow Var\left(Y_{ij}\right) = \sigma_\varepsilon^2 + \sigma_A^2 \Rightarrow Y_{ij} \sim N(0, \sigma_\varepsilon^2 + \sigma_A^2)$$

        - Las varianzas σ_A^2 y σ_ε^2 se llaman componentes de la varianza, y el modelo se suele llamar modelo de efectos aleatorios o de componentes de la varianza

- En este caso, contrastar la hipótesis de que las medias de los tratamientos o grupos es la misma no tiene sentido, de modo que se plantean hipótesis alternativas

  - La hipótesis nula en este tipo de ANOVA es que la fuente de variación del modelo se determina únicamente por el error ε, de modo que H_0: $\sigma_A^2 = 0$. La hipótesis alternativa, por lo tanto, es que $\sigma_A^2 > 0$ (dado que la varianza siempre tiene que ser positiva)

  - Si no se puede rechazar que $\sigma_A^2 = 0$, entonces los tratamientos o grupos serían todos idénticos, mientras que si $\sigma_A^2 > 0$, entonces existiría variabilidad (diferencias) entre los grupos

  - Debido a la naturaleza de este tipo de contraste, cualquier tipo de metodología para poder estimar la magnitud de los efectos o las diferencias entre los grupos o tratamientos (en caso de que haya diferencias entre grupos) no tendría sentido, dado que no se está interesado concretamente en los tratamientos o grupos de la muestra, sino en la población general

- La descomposición anterior de la suma de cuadrados total en la suma de cuadrados entre grupos y dentro de grupos sigue siendo válida

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{\bar{y}}\right)^2 = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{\bar{y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2$$

  - Debido a esto, es posible plantear una tabla ANOVA equivalente a la anteriormente vista

Source of variation	Degrees of freedom	Sum of squares	Mean square	F statistic
Between treatment groups	$k-1$	SSB = $\sum n_i(\bar{y}_i-\bar{\bar{y}})^2$	MSB = SSB/$(k-1)$	$F = \frac{\text{MSB}}{\text{MSW}}$
Within treatment groups	$N-k$	SSW = $\sum\sum(y_{ij}-\bar{y}_{i\cdot})^2$	MSW = SSW/$(N-k)$	
Total	$N-1$	SST = $\sum\sum(y_{ij}-\bar{\bar{y}})^2$		

  - La esperanza de cada uno de los términos en la descomposición, no obstante, difieren, y se obtienen las siguientes equivalencias:

$$E(MSB) = \frac{\left[\left(\sum_{i=1}^{k} n_i\right)^2 - \sum_{i=1}^{k} n_i^2\right]\sigma_A^2}{\sum_{i=1}^{k} n_i\,(k-1)} + \sigma_\varepsilon^2$$

$$E(MSW) = \sigma_\varepsilon^2$$

- Como uno no se enfoca en las magnitudes de los efectos ni en las diferencias entre tratamientos, uno se interesa en la cantidad de variabilidad entre los tratamientos o grupos (los efectos aleatorios) comparada con la variabilidad dentro de cada grupo (el error)

  - La varianza entre dentro de cada grupo o tratamiento se estima a través del MSW, dado que es un estimador no sesgado de σ_ε^2. No obstante, para estimar σ_A^2 se utiliza el siguiente estadístico:

$$S_A^2 = \frac{MSB - MSW}{\frac{\sum_{i=1}^{k} n_i - \sum_{i=1}^{k} n_i^2 / \sum_{i=1}^{k} n_i}{k-1}}$$

$$if \;\; n_i = n \;\; for \;\; i = 1,2,\dots,k \;\;\; then \;\;\; S_A^2 = \frac{MSB - MSW}{n}$$

  - A través de estos estimadores es posible obtener la proporción de varianza total explicada por la variación entre grupos (de los efectos aleatorios), lo cual se suele llamar correlación intraclase. A su vez, esto permite explicar la variación explicada por el error

$$\rho_1 = \frac{S_A^2}{S_A^2 + MSW} = \%\; explained\; by\; A_i$$

$$\Rightarrow \; 1 - \rho_1 = \%\; explained\; by\; \varepsilon$$

La examinación y transformación de los datos

- Una de las cosas más importantes en la inferencia estadística es el concepto de muestra, y en particular, los conceptos de muestra aleatoria y de muestra aleatoria simple, que se suelen utilizar a la hora de hacer inferencias

  - Los tres términos más importantes para la inferencia estadística son la población, la muestra y el muestreo

    - La población se define como un conjunto de elementos con una característica que se desea entender, la muestra se define como un subconjunto de la población, y el muestreo es el procedimiento por el cual se obtiene esta muestra

  - Para que esta sea útil a la hora de extraer conclusiones de la población, esta tiene que ser representativa

- Una condición que asegura que la muestra es representativa es que los individuos de la población se hayan escogido aleatoriamente. De este modo, cada individuo tiene la misma probabilidad de ser escogido en cualquier fase del proceso de muestreo y cada subconjunto de k individuos tiene la misma probabilidad de ser escogido que cualquier otro subconjunto de k individuos

- No obstante, una muestra puede ser representativa de la población, aunque esta no haya sido escogida de manera totalmente aleatoria

- El muestreo aleatorio simple se basa en que todos los subconjuntos de la población tengan la misma probabilidad (por lo que se asegura representatividad). Además, cualquier par de elementos tiene la misma probabilidad de ser escogido que otro par (por lo que se asegura independencia)

- El muestreo sistemático se basa en organizar el estudio de la población acorde a algún tipo de esquema de orden y, después, seleccionar los elementos en intervalos regulares a través de la lista ordenada. Este tipo de muestreo involucra un comienzo aleatorio y sigue con la selección de cualquier elemento k a partir de ahí

- El muestreo estratificado se basa en que, cuando la población tiene diferentes estratos o categorías, el muestreo se puede organizar por categorías o estratos. Se obtiene una muestra de cada uno de los estratos como si fueran subpoblaciones independientes, en donde cada individuo se selecciona de manera aleatoria, y el tamaño de cada estrato se puede determinar con diferentes métodos (proporcionalidad, optimalidad, etc.)

- El muestreo de clústeres se basa en escoger individuos en grupos o clústeres, y normalmente se sacan muestras por área geográfica o periodos de tiempo

- El muestreo de cuotas se basa en hacer una partición de la población (en subgrupos, como con un muestreo estratificado), pero en donde el criterio para seleccionar los sujetos para cada subgrupo se basa en una proporción (no es aleatorio)

- El muestreo de panel se basa en escoger un grupo de individuos a través del muestreo aleatorio simple y tomar una muestra (las

observaciones) en diferentes momentos dentro de un periodo de tiempo