MÉTODOS ESTADÍSTICOS

Análisis de la varianza

Problema: determinar si la media de una variable aleatoria Y es la misma en varias poblaciones distintas.
La variable respuesta Y debe tener distribución normal y su varianza debe ser la misma en todas las poblaciones.

Análisis de la varianza de un factor

Las poblaciones están definidas por varios niveles de un solo factor.
Modelo: n_i datos de la población $i \in I$, $Y_{ij} : j = 1, \ldots, n_i$.
Tabla ANOVA: de ella nos interesan principalmente los valores del estadístico F, su p-valor y también la media cuadrática de los residuos, que utilizamos como estimación de la varianza común: S_R^2.
Si el p-valor es inferior al nivel de significación α acordado entonces se rechaza la hipótesis nula de igualdad de todas las medias.
En este caso por lo general querremos contrastar igualdad de medias dos a dos utilizando contrastes t (que pueden hacerse por medio de intervalos de confianza para la diferencia de medias). Como se hacen varios contrastes se utiliza el método de Bonferroni para evitar decisiones erróneas debidas al azar.

Análisis de la varianza de más de un factor

Las poblaciones están definidas por todas las combinaciones posibles de los niveles de varios factores.

Dos factores sin réplicas. Los datos recogen un solo experimento por cada combinación de niveles de los factores.
Tabla ANOVA: Nos interesan principalmente los dos estadísticos F y sus correspondientes p-valores y también la media cuadrática de los residuos (varianza residual, S_R^2) que se utilizará como estimación de la varianza común.

Dos factores con réplicas. Interacción. Los datos recogen más de un experimento por cada combinación de nivles de los factores. Es posible entonces estudiar una posible interación de los niveles de uno de los factores en las medias de la variable respuesta según los niveles del otro factor y viceversa.
La tabla ANOVA incluye un estadístico F que se utiliza para contrastar la existencia de interacción.

Tres factores con igual número de clases para cada factor. Por medio de un cuadrado latino se reduce al mínimo el número de experimentos necesario para que en cada uno de los niveles de uno de los factores se realice un experimento con cada uno de los niveles de los otros dos factores. Por ejemplo, si el número de nivles de cada factor es de 4, en vez de realizarse 64 experimentos se realizan solamente 16.

Regresión lineal simple

Hipótesis: entre dos variables cuantitativas X, Y existe una relación del tipo siguiente. Para cada valor de x de X la variable Y tiene distribución normal de media $\beta_0 + \beta_1 x$ y de desviación típica σ independiente del valor de X.
Los datos que nos permiten estimar los parámetros (β_0, β_1, σ) pueden ser de dos tipos distintos: los valores de X están fijados o los valores de la X son aletorios (X, Y es un vector aleatorio con distribución normal bivariante). En cualquiera de los casos, los datos recogidos son independientes.

Regresión múltiple

Los valores de una variable cuantitativa Y se explican como función de K variables $X_1, X_2, \ldots, X_k$ de la manera siguiente: para cada selección de valores $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$ los valores de Y tienen distribución normal de media $\beta_0 + \beta_1 x_1 + \beta_2 + x_2 + \beta_k x_k$ y varianza σ^2 independiente de los valores de las X.

Los valores de los parámetros β_i, σ se estiman a partir de una muestra $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ de datos independientes, $i = 1, 2, \ldots, n$.