Análisis Multivariado

1. Exploración multivariada
   - Estadísticos básicos y gráficos
   - Gráficas de dispersión
   - Correlaciones
   - ANOVA
2. Modelos lineales
   - Regresión lineal simple
   - Regresión lineal múltiple
   - Regresión logística
3. Análisis multivariado
   - Análisis de discriminante
   - Análisis de componentes principales
   - Análisis de clústers

Estadísticos básicos y gráficas Teoría, definiciones, ejemplos y ejercicios

Estadísticos básicos y gráficos - definición

El análisis exploratorio de datos es el paso previo a la aplicación de cualquier método de análisis inferencial.

Una primera aproximación consiste en reducir toda la información muestral a un reducido grupo de valores numéricos, los parámetros estadísticos, cuyo conocimiento puede dar pie a la detección de algún patrón de comportamiento; algunos parámetros estadísticos son:

- Media
- Mediana
- Varianza
- Coeficiente de variación
- Coeficiente de simetría
- Coeficiente de curtosis

Además, es casi preceptivo en la fase exploratoria la confección de gráficos que permitan comprender mejor algunos comportamientos de los datos.

Estadísticos básicos y gráficos - tipos de gráficos

Algunas gráficas utilizadas en la exploración de datos son:

- Box plot
- Histogramas
- Multi-vari chart
- Run chart
- Pareto chart
- Scatter plot
- Parallel coordinates
- Odds ratio

Pero...¿cuándo utilizar cuál?

Estadísticos básicos y gráficos - ejemplos y ejercicios

Ejemplo sobre estadísticas y gráficos

- Análisis de operaciones por tipo de aduana
- Análisis de operaciones por tipo de aduana y aduana
- Box plot, histogramas y estadísticos
- Multi-vari, run chart, pareto y el resto

Ejercicios sobre estadísticas y gráficos

Replicar lo antes visto sobre las siguientes tablas y registrar resultados:

- ABT_AGEN_TARGET
- ABT_APOD_TARGET
- ABT_DEPENDIENTES_TARGET
- ABT_MANDATARIOS_TARGET
- DAT_GRAL_AGE_AGREG
- DAT_GRAL_MAN_AGREG

Correlaciones
Teoría, definiciones, ejemplos y ejercicios

Correlación - definición

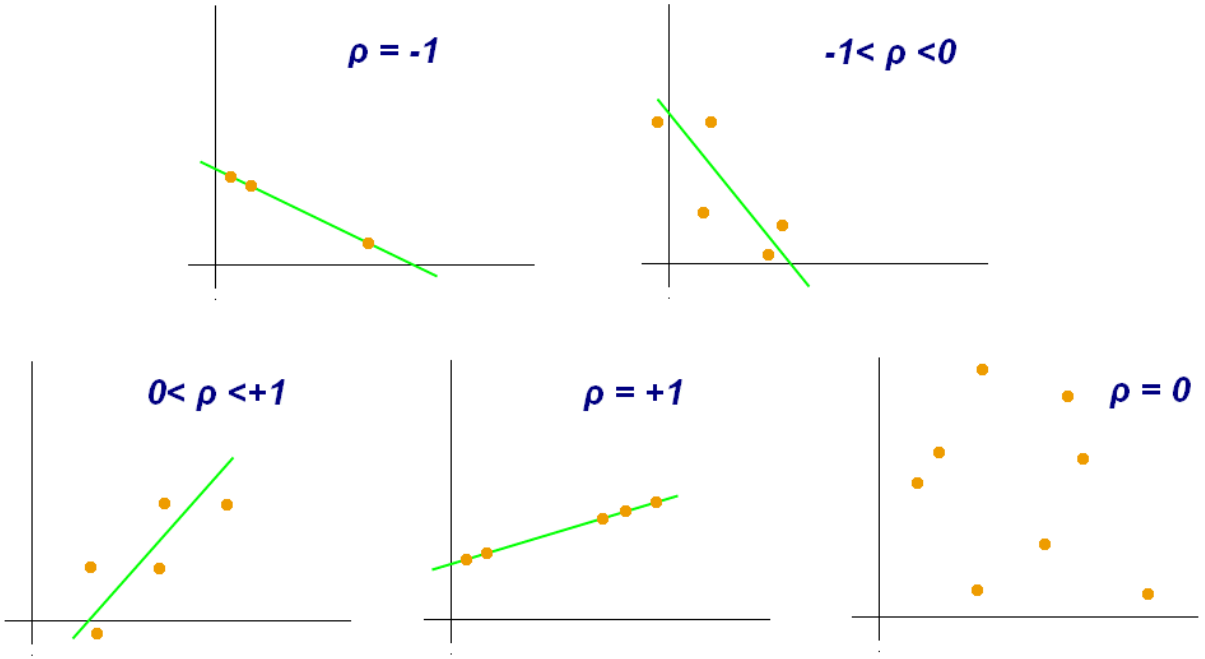
La correlación es una medida de la fuerza de relación entre variables aleatorias. La correlación poblacional entre dos variables X y Y se define como:

$$\rho(x, y) = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}},$$

ρ es conocida como el Coeficiente de correlación. Es un número que resume el sentido y la cercanía de las relaciones lineales entre dos variables.

Correlación - interpretación

El coeficiente de correlación puede tomar valores entre $[-1, 1]$. El signo ($+$ o $-$) de la correlación define la dirección de la relación. Cuando la correlación es positiva ($\rho > 0$), significa que a medida que el valor de una variable aumenta, también lo hace la otra.



Correlación - ejemplos y ejercicios

Ejemplos sobre correlación

- Correlación tres variables
- Correlación más variables
- Correlación y semaforización

Ejercicios sobre correlación

Replicar lo antes visto sobre las siguientes tablas y registrar resultados:

- ABT_AGEN_TARGET
- ABT_APOD_TARGET
- ABT_DEPENDIENTES_TARGET
- ABT_MANDATARIOS_TARGET
- DAT_GRAL_AGE_AGREG
- DAT_GRAL_MAN_AGREG

ANOVA
Teoría, definiciones, ejemplos y ejercicios

ANOVA - definición

El análisis de la varianza (ANOVA) prueba la hipótesis de que las medias de dos o más poblaciones son iguales. El ANOVA evalua la importancia de uno o más factores mediante la comparación de la respuesta media en los diferentes niveles de los factores. La hipótesis nula establece que todas las medias poblaciones son iguales, mientras que la hipótesis alternativa establece que al menos una es diferente.

Two-Sample t test - supuestos

- datos independientes
- datos distribuidos normalmente para cada grupo
- varianzas iguales para cada grupo

ANOVA - tipos

Type of Predictors / Type of Response	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

ANOVA - modelo

One-Way ANOVA

- H_0 medias iguales
- H_1 medias no iguales
- Partición de la variabilidad

Variabilidad total = variabilidad entre grupos + variabilidad en grupos. El modelo ANOVA es:

$$y_{ik} = \mu + x_{ik} + \epsilon_{ik},$$

One-Way ANOVA - supuestos

- observaciones independientes
- errores distribuidos normal
- varianza de los errores iguales para todos los grupos

ANOVA - ejemplos y ejercicios

Ejemplos sobre correlación

- Two-sample ANOVA
- One-way ANOVA

Ejercicios sobre correlación

Replicar lo antes visto sobre las siguientes tablas y registrar resultados:

- ABT_AGEN_TARGET
- ABT_APOD_TARGET
- ABT_DEPENDIENTES_TARGET
- ABT_MANDATARIOS_TARGET
- DAT_GRAL_AGE_AGREG
- DAT_GRAL_MAN_AGREG

Regresión lineal simple
Teoría, definiciones, ejemplos y ejercicios

El modelo lineal

El modelo lineal más simple implica sólo una variable independiente, plantea que la verdadera media de la variable dependiente cambia a una tasa constante cuando el valor de la variable independiente aumenta o disminuye. Así la relación funcional entre $\mathbb{E}[Y_i]$ y X_i es:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i,$$

donde β_0 es la intersección, el valor de $\mathbb{E}[Y_i]$ cuando $X_i = 0$ y β_1 es la pendiente de la recta, la tasa de cambio en $\mathbb{E}[Y_i]$ por una unidad de cambio en X_i.

Modelo lineal y supuestos

- Y_i son obs. aleatorias de poblaciones de v.a. con media de cada población $\mathbb{E}[Y_i]$.
- La desviación de Y_i de su media se tiene en cuenta mediate la suma de un error aleatorio ϵ_i dando el modelo estadístico

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

- X_i son v.a.i. y son medidas sin error.
- Los ϵ_i son planteados como

$$\epsilon_i \sim N.I.D.\left(0, \sigma^2\right).$$

Estimación mínimos cuadrados

El procedimiento de estimación de mínimos cuadrados utiliza el criterio de que la solución debe dar la suma más pequeña posible de las desviaciones al cuadrado de la Y_i observada de las estimaciones de sus verdaderas medias proporcionadas por la solución. Sea $\hat{\beta}_0$ y $\hat{\beta}_1$ las estimaciones numéricas de los parámetros β_0 y β_1, respectivamente, y sea

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

la media estimada de Y para cada X_i, $i = 1, 2, \ldots, n$. El principio de mínimos cuadrados escoge $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizen la suma de los cuadrados de los residuales, SS_{res}:

Estimación mínimos cuadrados

$$SS_{res} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} e_i^2.$$

Los estimadores de β_0 y β_1 son obtenidos usando cálculo para encontrar los valores que minimizen SS_{res}. Las derivadas de SS_{res} con respecto a los parámetros e igualando las a ceros, proporcionan dos ecuaciones conocidas como **ecuaciones normales**:

$$\begin{aligned} n\left(\hat{\beta}_0\right) + \left(\sum_{i=1}^{n} X_i\right)\hat{\beta}_1 &= \sum_{i=1}^{n} Y_i \\ \left(\sum_{i=1}^{n} X_i\right)\hat{\beta}_0 + \left(\sum_{i=1}^{n} X_i^2\right)\hat{\beta}_1 &= \sum_{i=1}^{n} Y_i X_i. \end{aligned}$$

Estimación mínimos cuadrados

Resolviendo las ecuaciones normales simultaneamente para $\hat{\beta}_0$ y $\hat{\beta}_1$ dada las estimaciones de β_0 y β_1:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\bar{X}. \end{aligned}$$

Esas estimaciones de los parámetros dan la ecuación de regresión

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

Valores pronosticados y residuales

Cada cantidad obtenida de la ecuación de regresión $\hat{Y}_i$ es usada como:

1. la **estimación** de la media poblacional de Y para cada valor particular de X.
2. el **pronóstico** del valor Y el cual se podría obtener en alguna observación futura de X.

Valores pronosticados y residuales

Si los valores observados de Y_i son comparados con sus correspondientes valores $\hat{Y}_i$, una medida del grado de coincidencia entre el modelo y los datos es obtenida. Los residuales

$$e_i = Y_i - \hat{Y}_i,$$

miden la discrepancia entre los datos y el modelo ajustado.

Análisis de variaciones en Y

Cada valor observado Y_i puede ser escrito como:

$$Y_i = \hat{Y}_i + e_i,$$

donde $\hat{Y}_i$ es la parte de la observación Y_i **explicada** por el modelo, mientras que e_i refleja la parte **no explicada** por el modelo. La **suma de cuadrados de Y_i total no corregida**,

$$SS(Total_{nc}) = \sum_{i=1}^{n} Y_i^2$$

puede ser también particionada. Sustituyendo $\hat{Y}_i + e_i$ para cada Y_i y expandiendo el cuadrado, se tiene,

Análisis de variaciones en Y

$$
\begin{aligned}
\sum_{i=1}^{n} Y_i &= \sum_{i=1}^{n} \left(\hat{Y}_i + e_i\right)^2 \\
&= \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} e_i^2 \\
&= SS_{modelo} + SS_{res},
\end{aligned}
$$

donde

$$
\begin{aligned}
SS_{modelo} &= n\bar{Y}^2 + \hat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \\
SS_{res} &= SS(Total_{nc}) - SS_{modelo}.
\end{aligned}
$$

Análisis de variaciones en Y

$SS(Total_{nc})$ puede ser expresada en términos de la **suma de cuadrados corregida** restando la suma de cuadrados debida a la corrección de la media, el factor de corrección $n\bar{Y}^2$, es decir,

$$\begin{aligned} SS(Total_{nc}) - n\bar{Y}^2 &= \left[SS_{modelo} - n\bar{Y}^2\right] + SS_{res} \\ \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 &= \hat{\beta}_1^2 \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{n} e_i^2 \\ SS_{total} &= SS_{regr} + SS_{res}. \end{aligned}$$

Análisis de variaciones en Y

Los grados de libertad asociados con cada suma de cuadrados son determinados por el tamaño de la muestra n y el número de parámetros $p^{'}$ en el modelo. Los grados de libertad asociados con SS_{modelo} son $p^{'} = 2$; los grados de libertad asociados con SS_{regr} son siempre uno menos para reflejar la sustracción del factor de corrección, el cual tiene un grado de libertad. SS_{res} contendrá $(n - p^{'})$ grados de libertad no contabilizados por el SS_{modelo}.

Análisis de variaciones en Y

Una medida de la contribución de la variable independiente en el modelo es el **coeficiente de determinación**, denotada por R^2:

$$R^2 = \frac{SS_{regr}}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}.$$

Es la proporción de la suma de cuadrados de Y_i (corregida) atribuible a la información obtenida de la variable independiente. El rango del coeficiente de determinación es el intervalo $[0, 1]$ y es el cuadrado del producto de la correlación entre Y_i y $\hat{Y}_i$.

Análisis de variaciones en Y

Si el modelo es correcto, la media cuadrada de los residuales es un estimador insesgado de σ^2, la varianza entre los errores aleatorios. La media cuadrada de los regresores es un estimador insesgado de $\sigma^2 + \beta_1^2 \left(\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \right)$. Ellas son referidas como las **medias cuadradas esperadas** y son denotadas por

$$\mathbb{E}\left[MS_{res}\right] \quad \text{y} \quad \mathbb{E}\left[MS_{regr}\right].$$

Precisión de las estimaciones

Puesto que, $\bar{Y}$, $\hat{Y}$, e, $\hat{\beta}_0$ y $\hat{\beta}_1$ son v.a.. Medidas de precisión, varianzas o errores estándar de los estimadores son una base para juzgar la confiabilidad de las estimaciones. Los $\hat{\beta}_j$, las $\hat{Y}_i$ y los e son todos **funciones lineales** de Y_i, así sus varianzas pueden ser determinadas usando la definición de varianza de una función lineal. Sea $U = \sum a_i Y_i$ una función lineal arbitraria de las v.a. Y_i, donde las $a_i's$ son constantes. La fórmula para la varianza de U es

$$\mathrm{Var}\left(U\right) = \sum a_i^2 \mathrm{Var}\left(Y_i\right) + \sum \sum_{i \neq j} a_i a_j \mathrm{Cov}\left(Y_i, Y_j\right),$$

Precisión de las estimaciones

donde la doble suma es sobre las $n(n-1)$ posibles pares de términos donde i y j no son iguales. $\mathrm{Cov}\,(\cdot,\cdot)$ denota la covarianza entre las dos variables indicadas dentro del paréntesis. Cuando las v.a. son independientes, entonces todas las covarianzas son cero. Si además, las varianzas de las v.a. son iguales; la varianza de la función lineal se reduce a

$$\mathrm{Var}\,(U) = \sigma^2 \left(\sum_{i=1}^{n} a_i^2 \right).$$

Precisión de las estimaciones

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum \left(X_i - \bar{X}\right)^2}.$$

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum \left(X_i - \bar{X}\right)^2}\right)\sigma^2.$$

$$\mathrm{Var}\left(\hat{Y}_i\right) = \left[\frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum_{j \neq i} \left(X_j - \bar{X}\right)^2}\right]\sigma^2.$$

$$\mathrm{Var}\left(\hat{Y}_{pronostico}\right) = \left[1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum \left(X_i - \bar{X}\right)^2}\right]\sigma^2.$$

Pruebas de significancia e intervalos de confianza

La hipótesis más común es que el verdadero valor del coeficiente de regresión lineal, β_1, es cero. En algunos casos, la naturaleza del fenómeno sugerirá otros valores para la hipótesis nula. La importancia de la prueba de significancia es la de proteger contra los efectos de las variaciones aleatorias en las estimaciones. Una pregunta importante sería ¿la diferencia entre el valor observado $\hat{\beta}_1$ y el valor hipotético del parámetro, es tan grande que puede ser atribuido a variaciones aleatorias? Si es así, la hipótesis nula es rechazada.

Pruebas de significancia e intervalos de confianza

Si los errores aleatorios en el modelo se distribuyen normal, las Y_i y cualquier función lineal de Y_i será distribuido normal. Así, $\hat{\beta}_1$ es distribuido normal, con media β_1 y varianza $\mathrm{Var}\left(\hat{\beta}_1\right)$. Si la hipótesis nula es $\beta_1 = m$ cierta, entonces $\hat{\beta}_1 - m$ se distribuye normal con media cero. Así,

$$t = \frac{\hat{\beta}_1 - m}{s\left(\hat{\beta}_1\right)},$$

se distribuye como t-Student con grados de libertad determinados por los gardos de libertad en la estimación de σ^2.

Pruebas de significancia e intervalos de confianza

El estadístico-F puede ser usado como una alternativa a la prueba t para hipótesis de dos colas sobre los coeficientes de regresión. Si la hipótesos nula es que $\beta_1 = 0$ es cierta, ambos MS_{regr} y MS_{res} estiman a σ^2. Cuando β_1 se desvia de cero, MS_{regr} será cada vez más grande (en promedio) que MS_{res}. Por consiguiente, el cociente de MS_{regr} y MS_{res} será mayor que la unidad y sugieré que β_1 es distinto de cero. Dicho cociente se distribuye F cuando el supuesto de normalidad en los reisduales en cierto y la hipótesis nula es cierta.

Pruebas de significancia e intervalos de confianza

Las estimaciones de los intervalos de confianza de los parámetros son más informativos que las estimaciones puntuales, ya que ellos reflejan la precisión de las estimaciones. Los intervalos de confianza al 95% de β_1 y β_0 son,

$$\hat{\beta}_1 \pm t_{(0.025,\nu)} s\left(\hat{\beta}_1\right)$$

y

$$\hat{\beta}_0 \pm t_{(0.025,\nu)} s\left(\hat{\beta}_0\right),$$

donde ν es el grado de libertad asociado con s^2.

Violación a los supuestos

Se asume que

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde los errores aleatorios ϵ_i son variables aleatorias independientes e identicamente distribuidas normal con media cero y varianza constante; y las X_i's son n observaciones de la variable independiente las cuales son medidas sin error. Bajo esos supuestos, los estimadores de mínimos cuadrados para β_0 y β_1 son los mejores (mínima varianza) entre todos los posibles estimadores insesgados.

Violación a los supuestos

Cuando sólo el sup. de normalidad es violado, los estimadores de mínimos cuadrados continuan siendo los de mínima varianza entre todos los estimadores lineales insesgados. El sup. de normalidad no es necesario para la partición de la variación total o la estimación de la varianza. Sin embargo, es necesario para las pruebas de signif. y para la const. de los int. de conf. y de pronósticos.
Cuando los errores estan correlacionados, los estimadores de mínimos cuadrados continuan siendo insesgados, pero ya no son los mejores estimadores. En este caso, los estimadores de las varianzas son sesgados.

Violación a los supuestos

Si la variabilidad de los errores incrementa con la variable independiente o con la media de la variable respuesta, se dice que existen varianciones no constantes. Dichos efectos en los estimadores de mínimos cuadrados son similares a los de los errores correlacionados, es decir, los estimadores no son los más eficientes y las estimaciones de las varianzas no son válidas. Cuando la variable independiente es medida con error o cuando en el modelo se cometen errores al omitir variables independientes importantes, los estimadores serán sesgados. En tal caso, los estimadores de la varianza también son sesgados.

Regresión lineal múltiple
Teoría, definiciones, ejemplos y ejercicios

El modelo

El modelo lineal aditivo para relacionar una variable dependiente con p variables independientes es

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i.$$

Hay $p+1$ parámetros β_j; por convención se usará $p^{'} = p+1$, además se asume que $n > p'$. En notación matricial el modelo puede escribirse como

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

donde

El modelo

$\mathbf{Y}$: vector de $n \times 1$ de las observaciones Y_i;

$\mathbf{X}$: matriz de $n \times p'$ con la primer columna de 1 y p columnas de las observaciones X_i;

β : vector de $p' \times 1$ con los parámetros a estimar; y

ϵ : vector de $n \times 1$ con los errores aleatorios.

Los vectores $\mathbf{Y}$ y ϵ son vectores aleatorios. $\mathbf{X}$ es la matriz de constantes conocidas. Un modelo para el cual $\mathbf{X}$ tiene todas las columnas linealmente independientes es un modelo de **rango completo**.

Ecuaciones normales y su solución

En notación matricial las ecuaciones normales son escritas como

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Y}.$$

Si $\mathbf{X}'\mathbf{X}$ tiene inversa (es decir, $\mathbf{X}$ es una matriz de rango completo), entonces las ecuaciones normales tienen solución única dada por

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}.$$

Estimaciones y residuales

El vector de las medias estimadas de la variable dependiente $\mathbf{Y}$ se define como

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}.$$

Es útil expresar a $\hat{\mathbf{Y}}$ como una función lineal de $\mathbf{Y}$ mediante

$$\begin{aligned}
\hat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} \\
&= \mathbf{X}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\right] \\
&= \left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{Y} \\
&= \mathbf{P}\mathbf{Y},
\end{aligned}$$

donde $\mathbf{P}$ es una matriz de $n \times n$ determinada enteramente por las $\mathbf{X}$'s, la cual es simétrica e idempotente. La relación anterior, muestra que $\hat{\mathbf{Y}}$ es una función lineal de $\mathbf{Y}$ con los coeficientes dados por $\mathbf{P}$.

Estimaciones y residuales

El vestor de residuales $\mathbf{e}$ refleja la discrepancia entre lo observado $\mathbf{Y}$ y estimado $\hat{\mathbf{Y}}$:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Así como $\hat{\mathbf{Y}}$, $\mathbf{e}$ puede expresarse como una función lineal de $\mathbf{Y}$ sustituyendo $\mathbf{PY}$ por $\hat{\mathbf{Y}}$, esdecir,

$$\mathbf{e} = \mathbf{Y} - \mathbf{PY} = (\mathbf{I} - \mathbf{P})\,\mathbf{Y}.$$

Propiedades de los estimadores

Los coeficientes de regresión estimados $\hat{\beta}$, los valores ajustados $\hat{\mathbf{Y}}$ y los residuales $\mathbf{e}$ son todos funciones lineales de $\mathbf{Y}$. Sabemos que

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

donde

$$\mathbb{E}\left[\epsilon\right] = 0 \quad \text{y} \quad \mathrm{Var}\left[\epsilon\right] = \mathbf{I}\sigma^2.$$

Entonces

$$\mathbb{E}\left[\mathbf{Y}\right] = \mathbf{X}\beta \quad \text{y} \quad \mathrm{Var}\left[\mathbf{Y}\right] = \mathbf{I}\sigma^2.$$

Propiedades de los estimadores

Cuando ϵ se distribuye normal, $\mathbf{Y}$ se distribuye normal multivariada. Así,

$$\mathbf{Y} \sim N\left(\mathbf{X}\beta, \mathbf{I}\sigma^2\right).$$

Este resultado esta basado en el supuesto que el modelo lineal usado es el modelo correcto; asumiendo lo anterior, la densidad conjunta de $\mathbf{Y}$ esta dada por

$$f\left(\mathbf{Y}\right) = \left(2\pi\right)^{-n/2} \sigma^{-n} e^{-\left(1/2\sigma^2\right)\left(\mathbf{Y}-\mathbf{X}\beta\right)'\left(\mathbf{Y}-\mathbf{X}\beta\right)}.$$

Propiedades de los estimadores

Expresando a $\hat{\beta}$ como

$$\hat{\beta} = \left[\left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}' \right] \mathbf{Y}$$

así los estimadores de los coef. de regresión son funciones lineales de $\mathbf{Y}$. Si el modelo $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ es correcto, entonces

$$\mathbb{E}\left[\hat{\beta}\right] = \beta \quad \text{y} \quad \mathrm{Var}\left[\hat{\beta}\right] = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \sigma^2.$$

Cuando ϵ se distribuye normal

$$\hat{\beta} \sim N\left(\beta, \left(\mathbf{X}'\mathbf{X} \right)^{-1} \sigma^2 \right).$$

Propiedades de los estimadores

El vector de medias estimadas $\hat{\mathbf{Y}}$ se define como

$$\hat{\mathbf{Y}} = \left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{Y} = \mathbf{P}\mathbf{Y},$$

. Si el modelo $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ es correcto, entonces

$$\mathbb{E}\left[\hat{\mathbf{Y}}\right] = \mathbf{X}\beta \quad \text{y} \quad \mathrm{Var}\left[\hat{\mathbf{Y}}\right] = \mathbf{P}\sigma^2.$$

Cuando ϵ se distribuye normal

$$\hat{\mathbf{Y}} \sim N\left(\mathbf{X}\beta, \mathbf{P}\sigma^2\right).$$

Propiedades de los estimadores

El vector de residuales se define como

$$\mathbf{e} = (\mathbf{I} - \mathbf{P})\,\mathbf{Y}.$$

Por consiguiente,

$$\mathbb{E}\left[\mathbf{e}\right] = 0 \quad \text{y} \quad \mathrm{Var}\left[\mathbf{e}\right] = (\mathbf{I} - \mathbf{P})\,\sigma^2.$$

Si ϵ se distribuye normal

$$\mathbf{e} \sim N\left(0, (\mathbf{I} - \mathbf{P})\,\sigma^2\right).$$

Propiedades de los estimadores

El pronóstico de una obs. futura, $Y_0 = \mathbf{x}_0'\beta + \epsilon_0$ en un vector dado $\mathbf{x}_0'$, esta dado por

$$\hat{\mathbf{Y}}_0 = \mathbf{x}_0'\hat{\beta}.$$

Se obs. que

$$\hat{\mathbf{Y}}_0 \sim N\left(\mathbf{x}_0'\hat{\beta}, \mathbf{x}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_0\sigma^2\right).$$

Este resultado es usado para construir intervalos de confianza para la media, $\mathbf{x}_0'\hat{\beta}$.

Partición de la varianza

El vector de obs. en la variable dependiente $\mathbf{Y}$ es particionado en $\hat{\mathbf{Y}}$ y $\mathbf{e}$, es decir,

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}.$$

Esta partición es utilizada para la suma total de cuadrados de la variable dependiente, es decir,

$$\mathbf{Y}'\mathbf{Y} = \sum Y_i^2,$$

dando SS_{total}. Así,

$$\begin{aligned}
\mathbf{Y}'\mathbf{Y} &= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathbf{Y}}'\mathbf{e} + \mathbf{e}\hat{\mathbf{Y}} + \mathbf{e}'\mathbf{e} \\
&= \mathbf{Y}'\mathbf{P}'\mathbf{P}\mathbf{Y} + \mathbf{Y}'\mathbf{P}'(\mathbf{I}-\mathbf{P})\mathbf{Y} + \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{P}\mathbf{Y} \\
&\quad + \mathbf{Y}'(\mathbf{I}-\mathbf{P})'(\mathbf{I}-\mathbf{P})\mathbf{Y} \\
&= \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{e}'\mathbf{e}.
\end{aligned}$$

Partición de la varianza

Así,

$$SS_{modelo} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \hat{\beta}\mathbf{X}'\mathbf{Y}$$

y

$$SS_{res} = \mathbf{e}'\mathbf{e} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} - SS_{modelo}.$$

Fuente de variación	Grados de libertad	Suma de cuadrados	
Total n.c.	$tr\,(\mathbf{I}) = n$	$\mathbf{Y}'\mathbf{Y}$	
Modelo	$tr\,(\mathbf{P}) = p'$	$\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$	$\hat{\beta}'\mathbf{X}'\mathbf{Y}$
Residuales	$tr\,(\mathbf{I} - \mathbf{P}) = n - p'$	$\mathbf{e}'\mathbf{e}$	$\mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}$

Partición de la varianza

La suma de cuadrados explicada por las variables independientes es llamado la suma de cuadrados de los regresores y etiquetada como SS_{regr}. Así

$$SS_{regr} = SS_{modelo} - SS_{\mu},$$

donde SS_{μ} es llamado comunmente el **factor de corrección**.

$$SS_{\mu} = \mathbf{Y}' \left(\frac{1}{n} \mathbf{J} \right) \mathbf{Y},$$

donde $\mathbf{J} = \mathbf{1}\mathbf{1}'$, así la matriz $\mathbf{J}/n$ es una matriz idempotente.

Partición de la varianza

Por lo que la suma de cuadrados atribuible a las variables independientes del modelo es

$$SS_{regr} = \mathbf{Y}'\left(\mathbf{P} - \mathbf{J}/n\right)\mathbf{Y}.$$

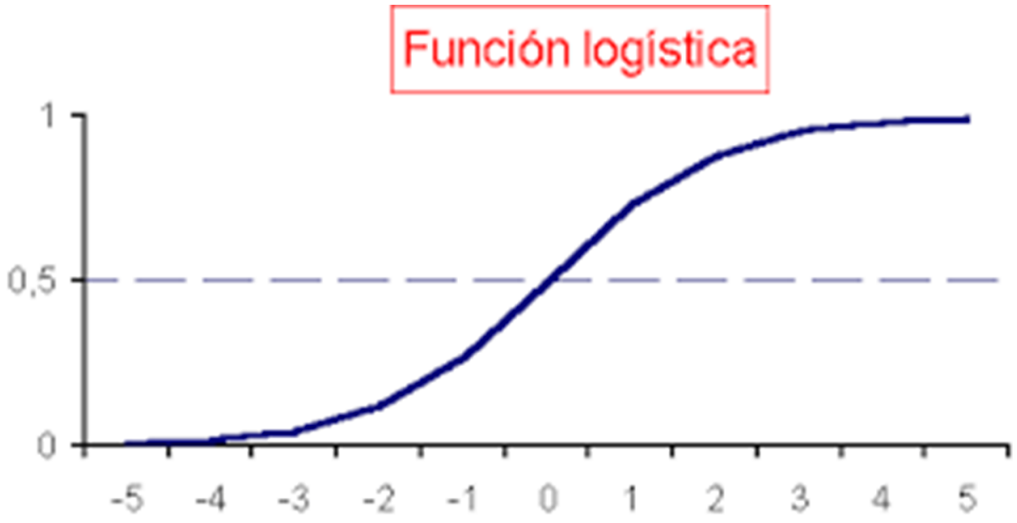
Así, la suma total de cuadrados es particionada en tres componentes ortogonales:

$$\begin{aligned} \mathbf{Y}'\mathbf{Y} &= SS_{\mu} + SS_{regr} + SS_{res} \\ &= \mathbf{Y}'\left(\mathbf{J}/n\right)\mathbf{Y} + \mathbf{Y}'\left(\mathbf{P} - \mathbf{J}/n\right)\mathbf{Y} + \mathbf{Y}'\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y} \end{aligned}$$

Regresión logística
Teoría, definiciones, ejemplos y ejercicios

Definición

La regresión logística es un tipo de análisis de regresión utilizado para pronosticar el resultado de una variable categórica en función de las variables independientes. Es útil para modelar la probabilidad de que un evento ocurra como función de otros factores. El análisis de regresión logística se un miembro de los Modelos Lineales Generalizados que usa como función liga la función **logit** y función de distribución de la variable dependiente **Bernoulli**.



El modelo

La regresión logística analiza datos distribuidos binomialmente de la forma

$$y_i \sim \mathrm{B}(p_i, n_i), \quad \text{para} \quad i = 1, \ldots, m,$$

donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas.
El modelo es entonces obtenido a base de lo que cada ensayo y el conjunto de variables independientes puedan informar acerca de la probabilidad final. Estas variables independientes son un vector x_i k-dimensional y el modelo toma entonces la forma

$$p_i = \mathbf{E}\left(\frac{y_i}{n_i}\middle| x_i\right).$$

Los logits de las probabilidades binomiales desconocidas son modeladas como una función lineal de los x_i,

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}. \tag{1}$$

Interpretación

El modelo de regresión logística escrito como la ecuación (1), tal vez con esa expresión el modelo no resulte muy elocuente, pero tras una transformaciones resulta de mayor interés:

- Dado el valor de las variables independientes, se puede calcular la probabilidad de ocurrencia del evento de interés:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}.$$

- Los odds para los individuos de referencia o control es

$$\exp(\beta_0).$$

- La cantidad $\exp(\beta_i)$ coincide con el cociente de probabilidades del aumento del valor x_i en una unidad con respecto a aquellos individuos que presentan los valores de todas las demás variables iguales.

Análisis de discriminante

Análisis discriminante

Análisis de discriminante (AD), es una técnica estadística multivariada que se utiliza comúnmente para construir un modelo predictivo de la discriminación de grupos basados en variables de predicción observadas y clasificar cada observación en uno de los grupos.

En AD múltiples atributos cuantitativos se utilizan para discriminar la variable de clasificación. Los objetivos comunes de AD son :

- investigar diferencias entre grupos,
- discriminar grupos efectivamente,
- identificar variables de discriminación importantes,
- clasificar nuevas observaciones.

Análisis discriminante

Parametric discriminant analysis, use a method of generalized squared distance. Accounting for prior probabilities, each observation is classified in the group to which it has the smallest generalized squared distance.
El análisis de discriminante, utiliza un método de distancia al cuadrado generalizada. Contabilización de las probabilidades previas, cada observación se clasifica en el grupo al que tiene la distancia al cuadrado más pequeño generalizada. La distancia al cuadrado generalizada para una observación x a un grupo t es

$$D_t^2(x) = d_t^2(x) + g_1(t) + g_2(t)$$

donde d_t^2 es la distancia Mahalanobis al cuadrado de x a el grupo t, g_1 depende de si es agrupado o no y se utiliza la matrices de covarianza, g_2 depende de la pronanilidades a priori estimadas.

Análisis discriminante

La distancia Mahalanobis al cuadrado d_t^2, de x a el grupo t esta definida como

$$d_t^2(x) = (x - \mathbf{m_t})'\mathbf{V_t^{-1}}(x - \mathbf{m})$$

donde $\mathbf{V_t} = \mathbf{S_t}$ si se usa la matriz de covarianzas entre grupos, o $\mathbf{V_t} = \mathbf{S_p}$ si se utiliza la matriz de covarianzas de los grupos, y $\mathbf{m_t}$ es el vector de las medias de los grupos t.

El análisis de discriminante asume que las matrices de covarianzas de los grupos son homogéneas.

Proc Discrim

PROC DISCRIM can be used for many different types of analysis including

- canonical discriminant analysis
- assessing and confirming the usefulness of the functions (empirical validation and crossvalidation)
- predicting group membership on new data using the functions (scoring)
- linear and quadratic discriminant analysis
- nonparametric discriminant analysis

The default in PROC DISCRIM is to use pooled covariances in calculating the generalized squared distances. If you prefer to use within-class covariances, specify POOL = NO in the DISCRIM statement this option is used for for quadratic discriminant analysis.

Proc Discrim Ejemplo

Pathological Gambling Example
Researchers want to determine whether a questionnaire measurement can adequately classify people identified as binge gamblers, steady gamblers, and nongamblers. A 12-item questionnaire based on DSM-IV diagnostic criteria for pathological gambling was administered to 100 people divided into the three groups.
The variable type is a class variable that identifies each observation as

- Binge
- Steady
- Control.

Análisis discriminante lineal

```
 proc discrim data = sasuser.gamblegrp list;
class type;
var dsm1-dsm12;
run;
```

Análisis discriminante cuadrático

```
 proc discrim data = sasuser.gamblegrp list pool = NO;
class type;
var dsm1-dsm12;
run;
```

Proc Discrim Resultados

Análisis discriminante lineal

The DISCRIM Procedure

Total Sample Size	100	DF Total	99
Variables	12	DF Within Classes	97
Classes	3	DF Between Classes	2

Number of Observations Read	100
Number of Observations Used	100

Class Level Information

type	Variable Name	Frequency	Weight	Proportion	Prior Probability
Binge	Binge	33	33.0000	0.330000	0.333333
Control	Control	48	48.0000	0.480000	0.333333
Steady	Steady	19	19.0000	0.190000	0.333333

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
12	-1.27952

The DISCRIM Procedure

Generalized Squared Distance to type

From type	Binge	Control	Steady
Binge	0	7.04994	10.75149
Control	7.04994	0	5.19242
Steady	10.75149	5.19242	0

Linear Discriminant Function for type

Variable	Label	Binge	Control	Steady
Constant		-12.08684	-4.18302	-8.98956
dsm1	Wished stop thkg re gambling	-0.72609	-0.14839	1.21042
dsm2	Wished stop thkg re get money	0.87634	0.28493	0.39280
dsm3	Felt need to bet more and more	0.43956	0.26780	0.10346
dsm4	Rely on others for funds	3.89561	0.86096	1.07942
dsm5	Gamble to escape	1.06698	0.52868	0.85271
dsm6	Lie about how much I gamble	0.38691	0.35853	0.22443
dsm7	Relaxing difficult if not gambling	0.07963	0.53608	0.37026
dsm8	Win back money next day	-0.26089	0.08036	-1.39281
dsm9	Felt I should cut back on gambling	1.21308	1.00883	1.65032
dsm10	Illegal acts to pay for gambling	0.42288	0.21331	0.46620
dsm11	Danger of losing relationship	0.68002	0.68806	0.82384
dsm12	Danger of losing job	0.46495	0.43634	1.05590

The DISCRIM Procedure
Classification Summary for Calibration Data: SASUSER.GAMBLEGRP
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into type

From type	Binge	Control	Steady	Total
Binge	33	0	0	33
	100.00	0.00	0.00	100.00
Control	8	36	4	48
	16.67	75.00	8.33	100.00
Steady	0	0	19	19
	0.00	0.00	100.00	100.00
Total	41	36	23	100
	41.00	36.00	23.00	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for type

	Binge	Control	Steady	Total
Rate	0.0000	0.2500	0.0000	0.0833
Priors	0.3333	0.3333	0.3333	

Proc Discrim Resultados

Análisis discriminante cuadrático

The DISCRIM Procedure

Total Sample Size	100	DF Total	99
Variables	12	DF Within Classes	97
Classes	3	DF Between Classes	2

Number of Observations Read	100
Number of Observations Used	100

Class Level Information

type	Variable Name	Frequency	Weight	Proportion	Prior Probability
Binge	Binge	33	33.0000	0.330000	0.333333
Control	Control	48	48.0000	0.480000	0.333333
Steady	Steady	19	19.0000	0.190000	0.333333

Within Covariance Matrix Information

type	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
Binge	12	-5.61076
Control	12	-1.26247
Steady	12	-10.13772

Generalized Squared Distance to type

From type	Binge	Control	Steady
Binge	-5.61076	3.68643	134.36281
Control	25.36712	-1.26247	71.47387
Steady	38.64178	4.61032	-10.13772

The DISCRIM Procedure
Classification Summary for Calibration Data: SASUSER.GAMBLEGRP
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into type

From type	Binge	Control	Steady	Total
Binge	32	1	0	33
	96.97	3.03	0.00	100.00
Control	1	45	2	48
	2.08	93.75	4.17	100.00
Steady	0	0	19	19
	0.00	0.00	100.00	100.00
Total	33	46	21	100
	33.00	46.00	21.00	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for type

	Binge	Control	Steady	Total
Rate	0.0303	0.0625	0.0000	0.0309
Priors	0.3333	0.3333	0.3333	

Análisis de componentes principales

Componentes principales

Principal component analysis, is a technique to reduce the dimensionality of data by forming new variables that are linear composites of the original variables. These new variables describe the main components or dimensions that are present in the original data set.

In principal components analysis, the goal is to create linear combinations of the original variables that maximally capture the variance in the data. In PCA, you typically create as many components as you have original variables, and the sum of the variance of the components will always account for 100% of the variance in the data.

Componentes principales propiedades

Principal components analysis exhibit the following properties:

- **Decreasing variance:** Much of the variance in the original set of p variables tends to be concentrated in the first few PCs, and little of the variance (information) in the original set of p variables tends to be concentrated in the last PCs.
- **Orthogonality:** The eigenvectors are orthogonal; therefore the principal components represent jointly perpendicular directions through the space of the original variables.
- **Independence:** Principal components are jointly uncorrelated. They can be used instead of the original variables in situations where having uncorrelated variables is desirable.

Elección de componentes

You can explain 100% of the variance in the data using all the principal components, this is not advantageous because you have the same number of PCs as you have input variables. If you want to use PCA for variable reduction, you need some decision rule or criterion for determining how many components to retain.

- **Scree plot of eigenvalues:** is a graphical tool for deciding how many components to retain. This plot shows the number of the eigenvalues $1, 2, \ldots, p$ on the X axis and the eigenvalues on the Y axis.
- **Proportion of variance explained by each component:** decide that a component is only retained if it explains a certain proportion of the variance.
- **Cumulative variance explained by components:** decide to retain components that together account for some percentage of the variance in the data
- **Eigenvalue > 1**: is often called the Kaiser-Guttman criterion.

Componentes principales en SAS Enterpise Guide

SAS Enterprise Guide use PROC PRINCOMP procedure to perform principal component analysis on continuous data and outputs standardized or unstandardized principal component scores.

You can specify the following analysis options:

Specifying your analysis:

- **Correlations**, computes the principal components from the correlation matrix.
- **Covariances**, computes the principal components from the covariance matrix.
- **Uncorrected correlations**, computes the principal components from the correlations matrix; however, the intercept is omitted from the model, so the correlations matrix and the standard deviations might not be corrected for the mean.
- **Uncorrected covariances**, computes the principal components from the covariance matrix; however, the intercept is omitted from the model, so the covariance matrix and the standard deviations might not be corrected for the mean.

Specifying the number of principal components to be computed: The default setting is the number of variables. **Changing the prefix for the principal components:** Specify the prefix for the principal components. **Specifying the**

singularity criterion: Specify the singularity criterion where $0 < p < 1$. The default is $1E\check{}08$. **Specifying the divisor**

for the variance: by default, the divisor for the variance calculation is the degrees of freedom. You can choose one of the following options:

Divisor	Formula	
Degrees of freedom	*n-i*	(before partialing)
	n-p-i	(after partialing)
Number of observations	*n*	
Sum of weights - 1	$\left(\sum_{j=1}^{n} w_j\right) - i$	(before partialing)
	$\left(\sum_{j=1}^{n} w_j\right) - p - i$	(after partialing)

Componentes principales en SAS Enterpise Guide

You can select the following plots:

Create scree and variance plots : creates a scree plot of eigenvalues and a plot of the proportion variance.
Create a scatter matrix principal component scores plot: creates a scatter plot matrix of the principal component scores.
Create a pattern profile plot: creates a pattern profile plot. There is one profile for each component. The value on the Y-axis is the correlation between the variable and the principal component.
Create a scatter plot of the principal components scores: creates a scatter plot for each principal components score.
Create pattern component plots : creates a scatter plot for each principal components score.
Create a scatter plot of the principal components scores: creates pairwise component pattern plots. Each observation on the plot is the correlation between the variable and the two corresponding components on the plot..

Resultados análisis de componentes principales

Principal Components Analysis

The PRINCOMP Procedure

Observations	60
Variables	16

Simple Statistics

	PREC	JANT	JULT	OVR65	POPN	EDUC	HOUS	DENS	NONW	WWDRK	POOR	HC	NOX	SO2	HUMID	MORT
Mean	37.36666667	33.98333333	74.58333333	8.798333333	3.263166667	10.97333333	80.91333333	3876.050000	11.87000000	46.08166667	14.37333333	37.85000000	22.65000000	53.76666667	57.66666667	940.3584333
StD	9.98467753	10.16889852	4.76317679	1.464551955	0.135252327	0.84529940	5.14137312	1454.102361	8.92114798	4.61304310	4.16009561	91.97767323	46.33328964	63.39046784	5.36993093	62.2062785

Correlation Matrix

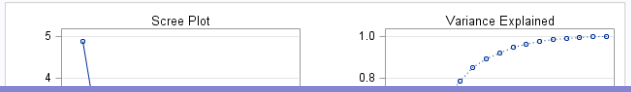
		PREC	JANT	JULT	OVR65	POPN	EDUC	HOUS	DENS	NONW	WWDRK	POOR	HC	NOX	SO2	HUMID	MORT
PREC	Average annual precipitation in inches	1.0000	0.0922	0.5033	0.1011	0.2634	-.4904	-.4908	-.0035	0.4132	-.2973	0.5066	-.5318	-.4873	-.1069	-.0773	0.5095
JANT	Average January temperature in degrees F	0.0922	1.0000	0.3463	-.3981	-.2092	0.1163	0.0149	-.1001	0.4538	0.2380	0.5653	0.3508	0.3210	-.1078	0.0679	-.0300
JULT	Average July temperature in degrees F	0.5033	0.3463	1.0000	-.4340	0.2623	-.2385	-.4150	-.0610	0.5753	-.0214	0.6193	-.3565	-.3377	-.0993	-.4528	0.2770
OVR65	Percent population aged 65 or older (1960)	0.1011	-.3981	-.4340	1.0000	-.5091	-.1389	0.0650	0.1620	-.6378	-.1177	-.3098	-.0205	-.0021	0.0172	0.1124	-.1746
POPN	Average household size	0.2634	-.2092	0.2623	-.5091	1.0000	-.3951	-.4106	-.1843	0.4194	-.4257	0.2599	-.3882	-.3584	-.0041	-.1357	0.3573
EDUC	Median school years completed by those over 22	-.4904	0.1163	-.2385	-.1389	-.3951	1.0000	0.5522	-.2439	-.2088	0.7032	-.4033	0.2868	0.2244	-.2343	0.1765	-.5110
HOUS	Percent housing units which are sound & with all facilities	-.4908	0.0149	-.4150	0.0650	-.4106	0.5522	1.0000	0.1819	-.4103	0.3387	-.6807	0.3868	0.3483	0.1180	0.1219	-.4268
DENS	Population per sq. mile in urbanized areas, 1960	-.0035	-.1001	-.0610	0.1620	-.1843	-.2439	0.1819	1.0000	-.0057	-.0318	-.1629	0.1203	0.1653	0.4321	-.1250	0.2655
NONW	Percent non-white population in urbanized areas, 1960	0.4132	0.4538	0.5753	-.6378	0.4194	-.2088	-.4103	-.0057	1.0000	-.0044	0.7049	-.0259	0.0184	0.1593	-.1180	0.6437
WWDRK	Percent employed in white collar occupations	-.2973	0.2380	-.0214	-.1177	-.4257	0.7032	0.3387	-.0318	-.0044	1.0000	-.1852	0.2037	0.1600	-.0685	0.0607	-.2848
POOR	Percent of families with income < $3000	0.5066	0.5653	0.6193	-.3098	0.2599	-.4033	-.6807	-.1629	0.7049	-.1852	1.0000	-.1298	-.1025	-.0965	-.1522	0.4105
HC	Relative hydrocarbon pollution potential	-.5318	0.3508	-.3565	-.0205	-.3882	0.2868	0.3868	0.1203	-.0259	0.2037	-.1298	1.0000	0.9838	0.2823	-.0202	-.1772
NOX	Relative nitric oxides pollution potential	-.4873	0.3210	-.3377	-.0021	-.3584	0.2244	0.3483	0.1653	0.0184	0.1600	-.1025	0.9838	1.0000	0.4094	-.0459	-.0774
SO2	Relative sulphur dioxide pollution potential	-.1069	-.1078	-.0993	0.0172	-.0041	-.2343	0.1180	0.4321	0.1593	-.0685	-.0965	0.2823	0.4094	1.0000	-.1026	0.4259
HUMID	Annual average % relative humidity at 1:00pm	-.0773	0.0679	-.4528	0.1124	-.1357	0.1765	0.1219	-.1250	-.1180	0.0607	-.1522	-.0202	-.0459	-.1026	1.0000	-.0885
MORT	Total age-adjusted mortality rate per 100,000	0.5095	-.0300	0.2770	-.1746	0.3573	-.5110	-.4268	0.2655	0.6437	-.2848	0.4105	-.1772	-.0774	0.4259	-.0885	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.87859562	2.11202119	0.3049	0.3049
2	2.76657442	0.47409874	0.1729	0.4778
3	2.29247568	0.94081534	0.1433	0.6211
4	1.35166034	0.12815293	0.0845	0.7056
5	1.22350741	0.13676893	0.0765	0.7821
6	1.08673848	0.42526222	0.0679	0.8500
7	0.66147626	0.18205081	0.0413	0.8913
8	0.47942545	0.07192460	0.0300	0.9213
9	0.40750085	0.16268096	0.0255	0.9467
10	0.24481989	0.05072219	0.0153	0.9620
11	0.19409770	0.03769574	0.0121	0.9742
12	0.15640196	0.03959183	0.0098	0.9840
13	0.11681013	0.02752574	0.0073	0.9913
14	0.08928439	0.04332239	0.0056	0.9968
15	0.04596200	0.04129258	0.0029	0.9997
16	0.00466942		0.0003	1.0000

Eigenvectors

		PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7	PRIN8	PRIN9	PRIN10	PRIN11	PRIN12	PRIN13	PRIN14	PRIN15	PRIN16
PREC	Average annual precipitation in inches	0.336695	-.113583	-.001523	0.325508	0.129198	0.178372	-.001818	0.507584	0.267262	0.364516	0.224912	0.031821	-.321569	-.160368	-.276767	0.000641
JANT	Average January temperature in degrees F	0.057553	0.463602	-.179848	0.330008	-.082768	0.035248	-.374027	0.079430	-.175167	0.149127	-.330297	0.017996	-.321423	0.167313	0.438537	0.015783
JULT	Average July temperature in degrees F	0.320576	0.167142	-.168470	0.034811	0.392329	-.169747	-.098632	0.092914	-.268224	-.209435	0.535602	-.320780	0.189444	0.301876	0.076581	0.012014
OVR65	Percent population aged 65 or older (1960)	-.155296	-.344587	0.194812	0.524002	0.037212	-.062652	0.252196	0.111259	0.045270	0.154776	0.040660	0.134671	0.386721	0.224860	0.463674	0.049176
POPN	Average household size	0.285398	-.077239	-.005597	-.548713	-.238237	-.060219	-.074733	0.019995	0.147847	0.547269	0.069628	0.082694	0.106726	0.403473	0.192635	0.027553
EDUC	Median school years completed by those over 22	-.293791	0.157534	-.348163	-.142394	0.144797	0.173399	0.223083	0.108308	0.190344	-.214853	0.288011	0.590650	-.219628	0.240931	0.107399	0.053072
HOUS	Percent housing units which are sound & with all facilities	-.345269	0.065490	0.011322	-.161661	0.196922	0.132930	-.425926	0.553857	0.023774	-.009991	-.230393	0.025950	0.456467	0.077793	-.190867	-.018829
DENS	Population per sq. mile in urbanized areas, 1960	-.033467	0.018034	0.437103	0.061519	0.419022	0.193545	-.463218	-.484874	0.259985	0.090483	0.159172	0.182904	-.050557	0.062221	-.000738	-.000922
NONW	Percent non-white population in urbanized areas, 1960	0.313794	0.363090	0.009912	-.109637	-.012986	0.230334	0.087002	0.036584	0.196036	-.059273	0.081472	0.125505	0.398717	-.575829	0.374897	-.003822
WWDRK	Percent employed in white collar occupations	-.196136	0.229178	-.269244	0.009618	0.397236	0.300026	0.403410	-.235466	0.006511	0.458158	-.180606	-.315283	0.112013	0.054630	-.110728	-.025718
POOR	Percent of families with income < $3000	0.344207	0.247404	-.080187	0.292025	-.135515	-.091335	0.054628	-.193519	-.096085	0.018679	-.157998	0.453383	0.356646	0.213157	-.498608	0.011057
HC	Relative hydrocarbon pollution potential	-.250798	0.393455	0.210785	0.072368	-.224042	-.215113	0.055164	0.036276	0.239851	0.071916	0.205519	-.190739	0.018653	0.022133	-.121164	0.689174
NOX	Relative nitric oxides pollution potential	-.226728	0.395319	0.280157	0.063750	-.209517	-.189961	0.116704	0.077655	0.160630	0.074453	0.229263	-.090337	-.007618	0.090345	-.059258	-.711571
SO2	Relative sulphur dioxide pollution potential	-.015787	0.142727	0.515166	-.169239	0.101190	0.172927	0.225641	0.170582	-.675295	0.150071	0.061418	0.230742	-.130243	-.048209	-.008128	0.101360
HUMID	Annual average % relative humidity at 1:00pm	-.109189	-.081814	-.102580	0.140143	-.497992	0.680707	-.167622	-.112043	-.171548	-.024116	0.338435	-.157114	0.117947	0.105115	-.070437	-.007615
MORT	Total age-adjusted mortality rate per 100,000	0.294026	0.079898	0.333203	-.055631	0.020337	0.344009	0.257812	0.133479	0.284885	-.418962	-.329016	-.211220	-.085398	0.414501	0.022710	0.028864

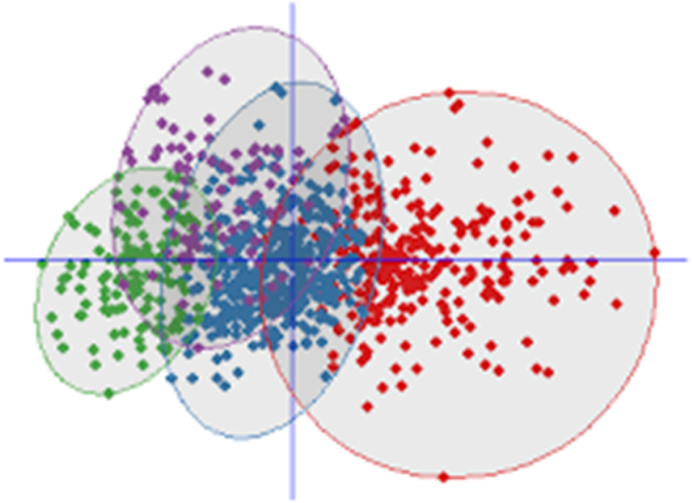


Análisis de clústers

Definición

Análisis de grupos o agrupamiento es la tarea de agrupar un conjunto de objetos de tal manera que objetos en el mismo grupo (llamado cluster) sean más similares (en algún sentido u otro) entre ellos que a aquellos en otros grupos (clusters).

El análisis de grupos no es en sí un algoritmo especifico, sino la tarea general a resolver. Se puede hacer el agrupamiento utilizando varios algoritmos que difieren significativamente en su idea de qué constituye un grupo y cómo encontrarles eficientemente.



Tipos de algoritmos

Típicamente los modelos de cluster incluyen:

- Modelos de conectividad
- Modelos de centroide
- Modelos de distribución
- Modelos de densidad
- Modelos de sub-espacios
- Modelos de grupo
- Agrupamiento suave
- Arupación con partición estricta con ruido
- Agrupación con solapamiento
- Agrupamiento jerárquico

Proc

El PROC CLUSTER agrupa jerárquicamente las observaciones en un conjunto de datos, haciendo uso de uno de los 11 métodos. Los datos pueden ser coordenadas o distancias. Si los datos son coordenadas, el PROC CLUSTER calcula distancias euclidianas. Si quieres distancias no euclidianas, utiliza el PROC DISTANCE para calcular un conjunto de datos a distancia apropiado que puede ser utilizado como insumo para el PROC CLUSTER.

Los métodos de la agrupación son: la vinculación promedio, el método centroide, vinculación completo, vinculación densidad, máxima verosimilitud para las mezclas de esféricas distribuciones normales multivariadas con varianzas iguales pero posiblemente proporciones de mezcla desiguales, método flexibilidad-beta, análisis de similitud de McQuitty , el método de la mediana, la única vinculación , la vinculación de densidad de dos etapas, y el método de mínima varianza de Ward.

Ejemplo - Análisis de clúster para los datos Iris

Los datos Iris publicados por Fisher (1936) han sido ampliamente utilizados para los ejemplos en el análisis discriminante y el análisis cluster. La longitud sépalo, ancho sépalo, pétalo de longitud y la anchura de pétalos se miden en milímetros, en 50 ejemplares de cada una de las tres especies, setosa, versicolor y virginica.
Este ejemplo analiza los datos iris usando el método de Ward y vinculación densidad de dos etapas e ilustra cómo el **FASTCLUS PROC** se puede utilizar en combinación con **CLUSTER PROC** para analizar grandes conjuntos de datos.

Cluster Analysis of Fisher's Iris Data

- Ward's Method
- Two-Stage Density Linkage
- Preliminary Analysis by FASTCLUS
- Clustering Clusters by Ward's Method

MEDIA

La media ($\bar{x}$ o m_x) es el más conocido e intuitivo, siendo su objeto localizar alrededor de qué punto se sitúan todas las observaciones. Su cálculo es bien sencillo:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

MEDIANA

La mediana (med), como la media, es un parámetro de localización, siendo su objeto resumir en una sola cantidad los valores muestrales. Se define como el número tal que $F(med) = 1/2$, siendo F la función de distribución muestral. Intuitivamente, es el valor numérico que queda en el centro cuando se ordena toda la muestra.

VARIANZA

La varianza (s_x^2) mide la dispersión alrededor de la media; cuanto más pequeña sea, más concentrados estarán los puntos alrededor de $\bar{x}$:

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Para evitar tener que trabajar con las unidades cuadradas de la varianza, se extrae su raíz cuadrada, con lo que se obtiene la **DESVIACIÓN TÍPICA** (s_x):

$$s_x = \sqrt{s_x^2}.$$

En ocasiones, la varianza muestral introduce un sesgo; una forma de corregir este defecto es calculando la **CUASI-VARIANZA** (S_x^2):

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

a la cual se asocia la **CUASI-DESVIACIÓN TÍPICA** (S_x):

COEFICIENTE DE VARIACIÓN

El coeficiente de variación (CV) se suele utilizar para comparar la dispersión relativa de varias muestras:

$$CV = \frac{s_x}{\bar{x}}.$$

COEFICIENTE DE SIMETRÍA

El coeficiente de simetría (γ_1) mide si la muestra se distribuye de igual manera a ambos lados de la media:

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3}.$$

- Si $\gamma_1 < 0$, la simetría es negativa, siendo mayor la dispersión hacia la izquierda de la media.
- Si $\gamma_1 = 0$, la muestra es simétrica alrededor de la media.
- Si $\gamma_1 > 0$, la simetría es positiva, siendo mayor la dispersión hacia la derecha de la media.

COEFICIENTE DE CURTOSIS

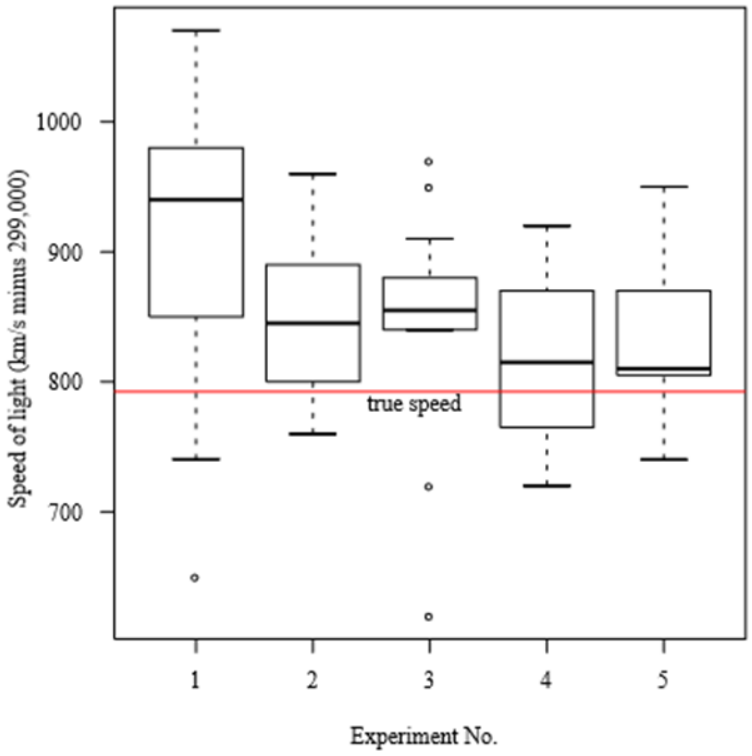
El coeficiente de curtosis (γ_2) mide la concentración de la muestra alrededor de la media:

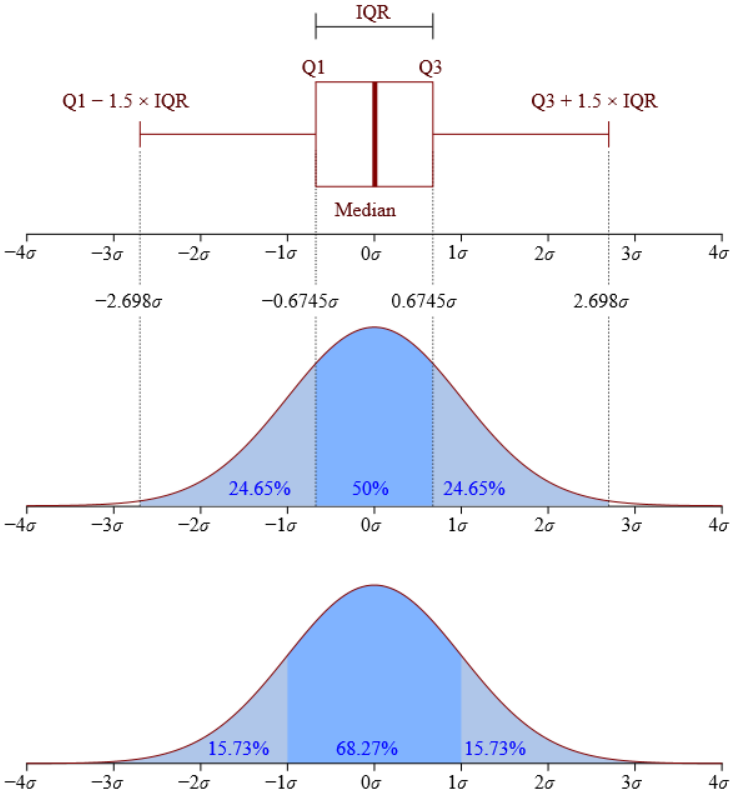
$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4} - 3.$$

- Si $\gamma_2 < 0$, el histograma de la muestra está más aplastado que el de la distribución normal.
- Si $\gamma_2 = 0$, el histograma de la muestra está igual de aplastado que el de la distribución normal.
- Si $\gamma_2 > 0$, el histograma de la muestra está menos aplastado que el de la distribución normal.

Box plot

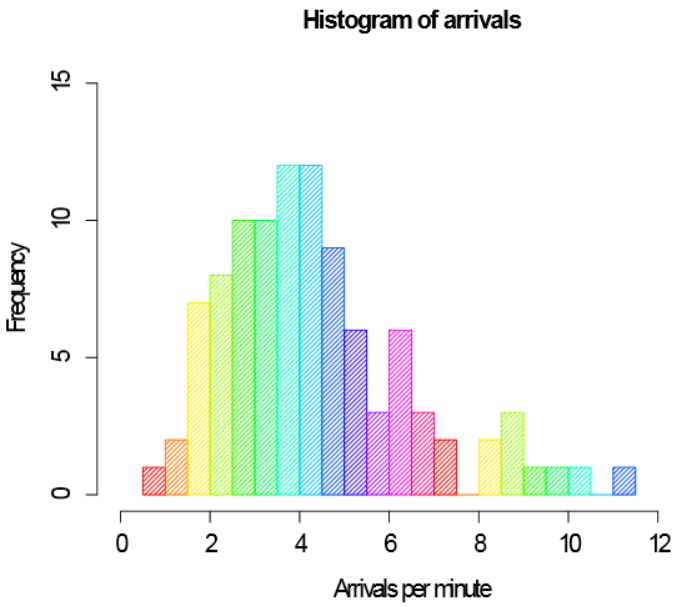
Un box plot es una manera conveniente de representar gráficamente los grupos de datos numéricos a través de sus cuartiles.

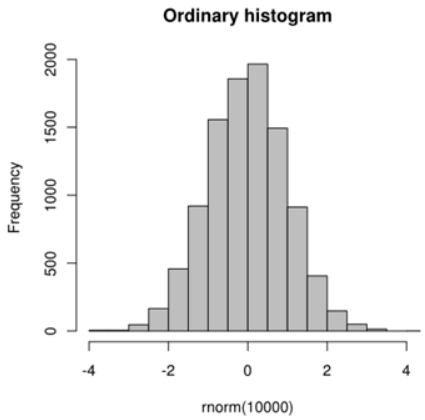


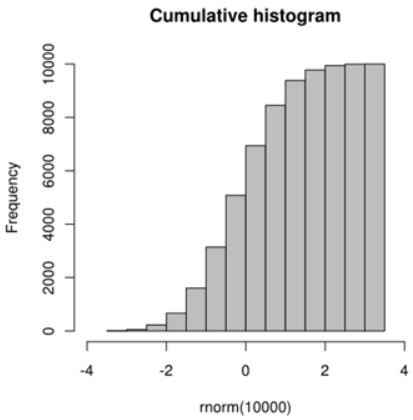


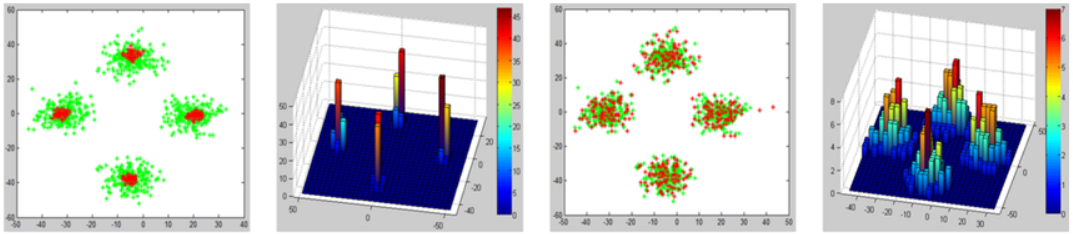
Histograma

Un histograma es una representación gráfica de la distribución de los datos numéricos.

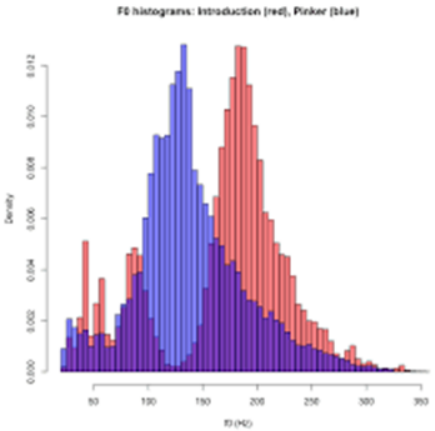








(a) (b) (c) (d)

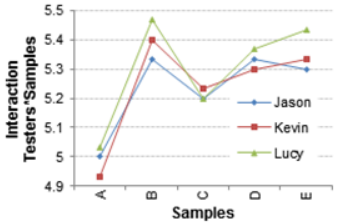


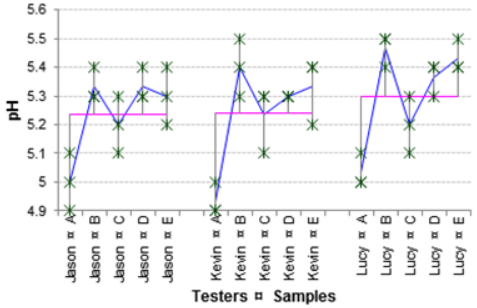
Multi-vari chart

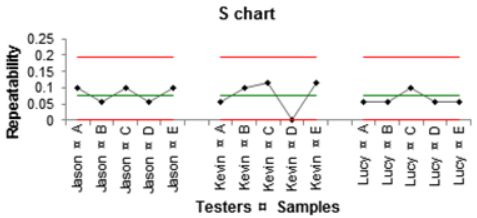
En el control de calidad , los gráficos multi-vari son una forma visual de presentar la variabilidad a través de una serie de gráficos.

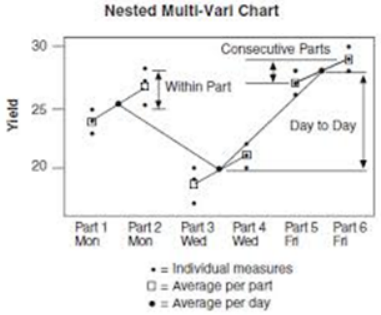
Measurement variation	
S(Repeatability)	0.080
S(Reproductibility)	0.030
S(Interaction)	0.000
S(Samples)	0.164
S(Total measurement)	0.085
Confidence level	0.95
Precision (±)	**0.167**

Increment diagnosis	
Increment	0.1
Probable error	0.058
Minimal increment	0.013
Maximal increment	0.127
Measurement increment	sufficient



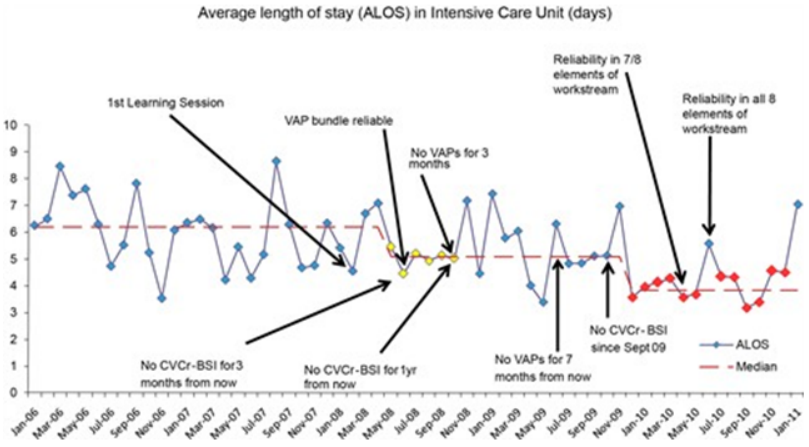


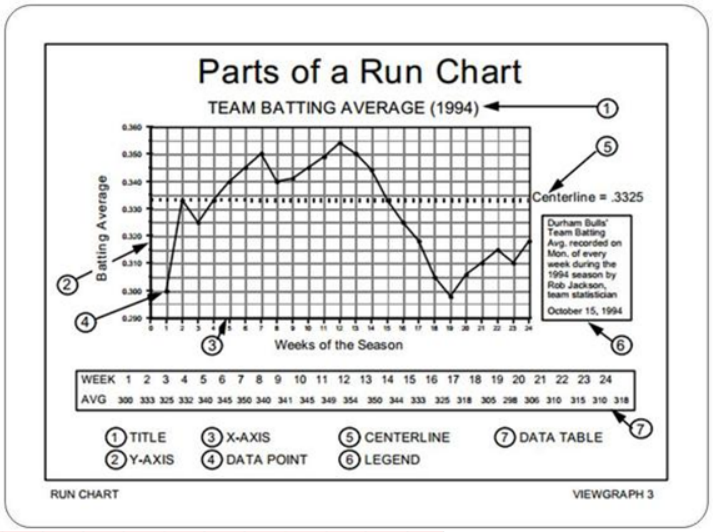


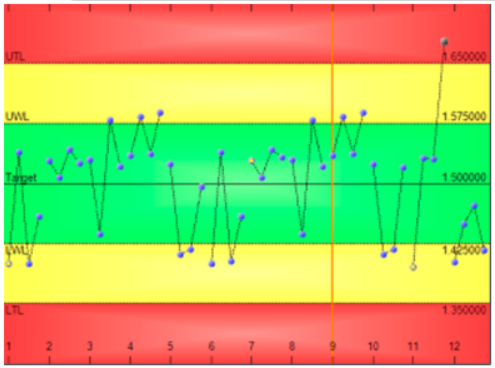


Run chart

Un diagrama de comportamiento, es un gráfico que muestra los datos observados en una secuencia de tiempo.

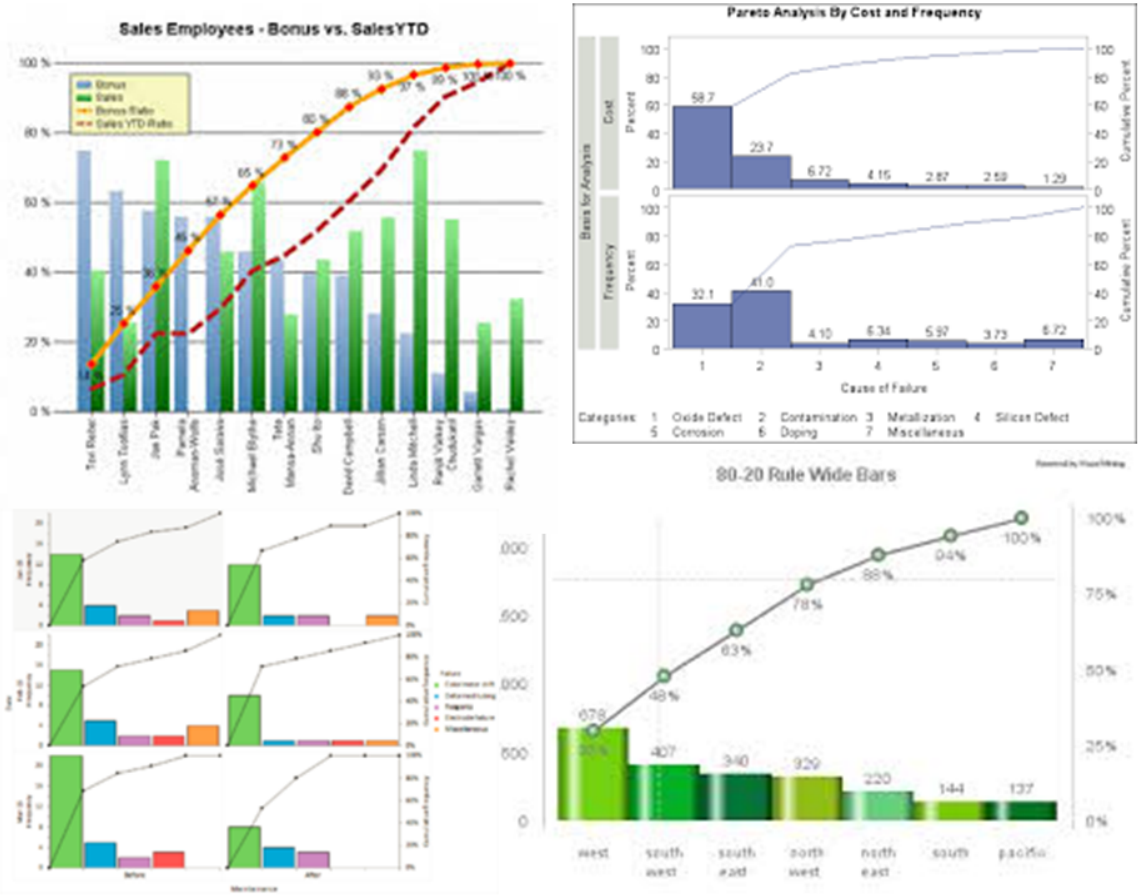






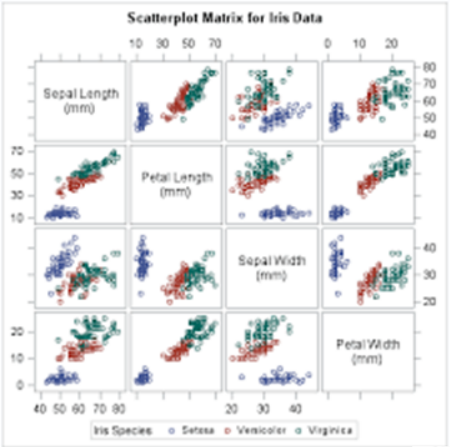
Pareto chart

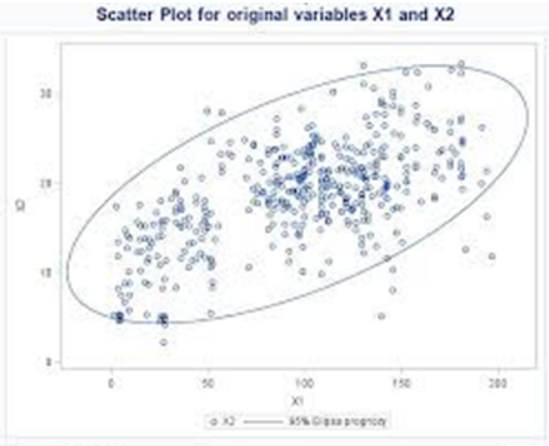
Un diagrama de Pareto, es un tipo de gráfico que contiene barras y un gráfico de línea, donde están representados los valores individuales en orden descendiente por barras y el total acumulado está representado por la línea.

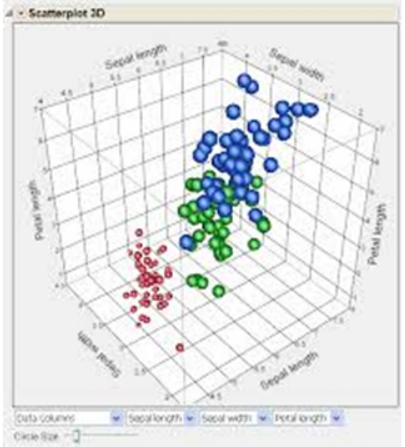


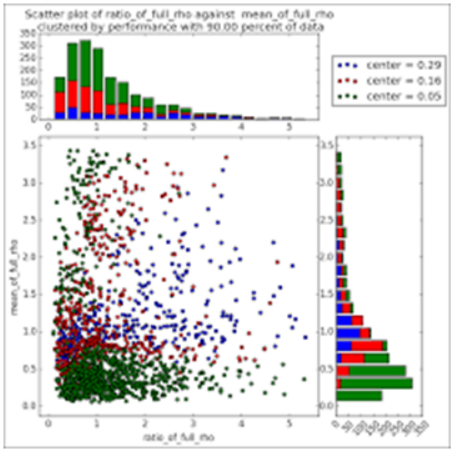
Scatter plot

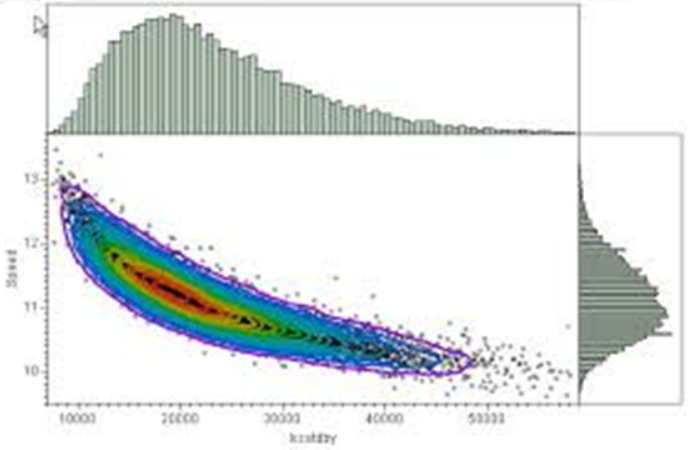
Un gráfico de dispersión, es un tipo de diagrama utilizando coordenadas cartesianas para mostrar los valores de típicamente dos variables para un conjunto de datos. Si los puntos están codificados por color puede aumentar el número de variables que se muestran.

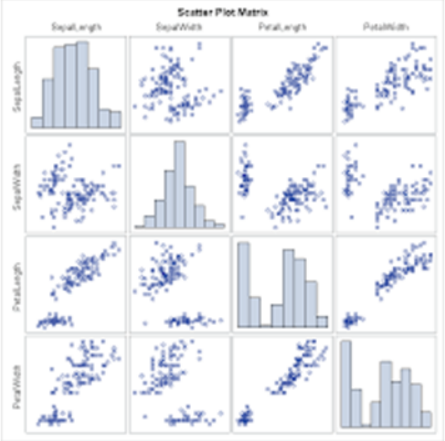






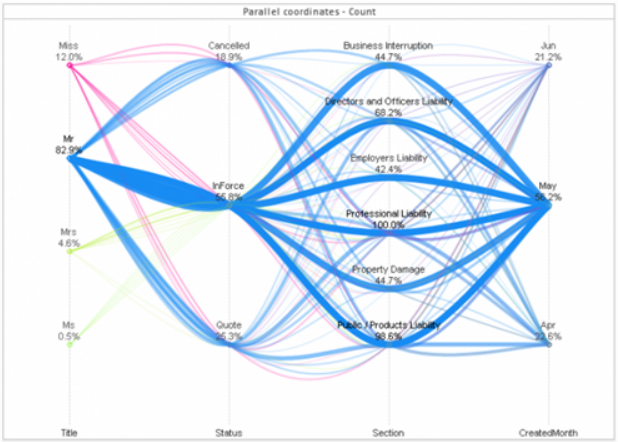


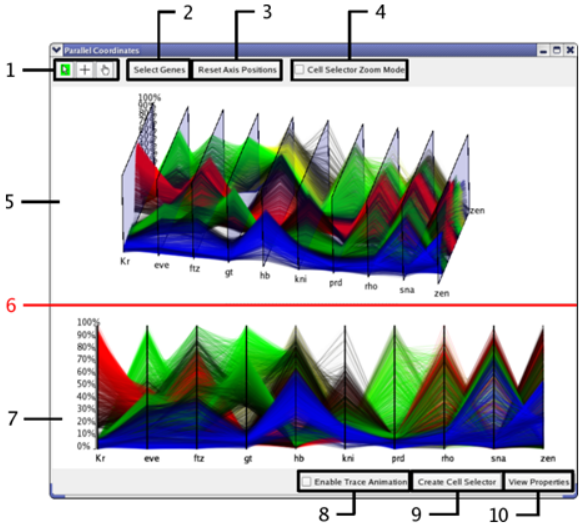


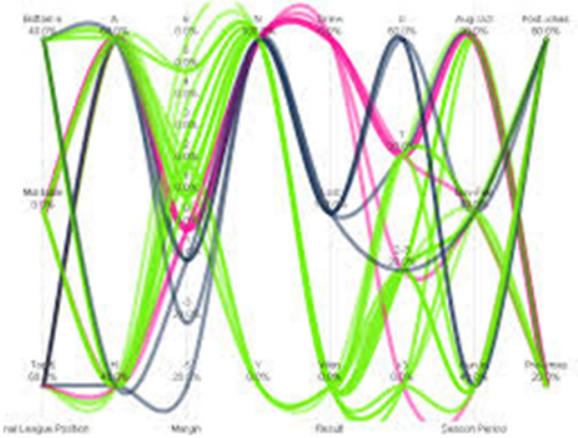


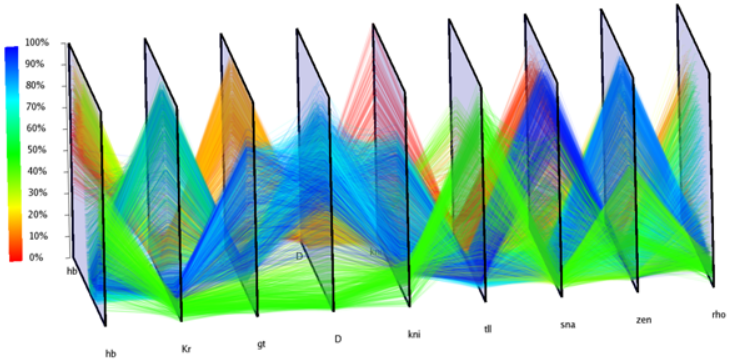
Parallel coordinates

Coordenadas paralelas, es una forma común de visualizar la geometría de altas dimensiones y el análisis de datos multivariados.







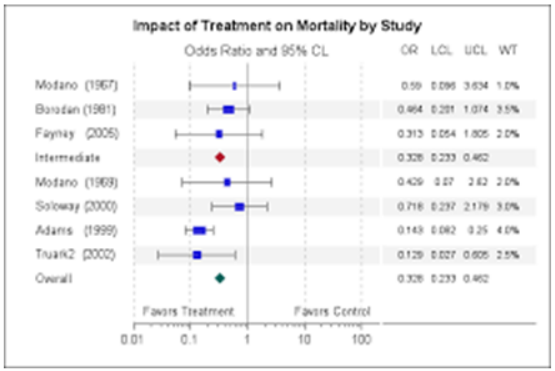


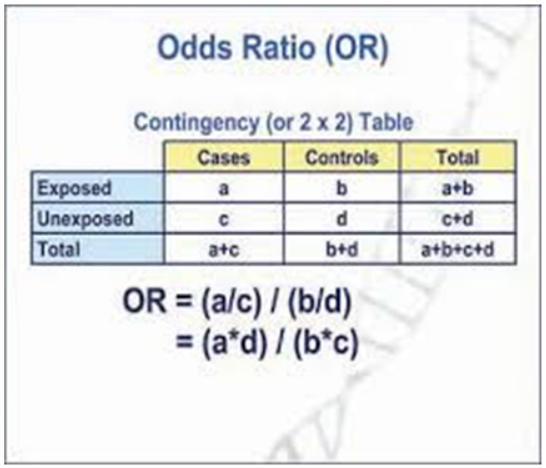
Odds ratio

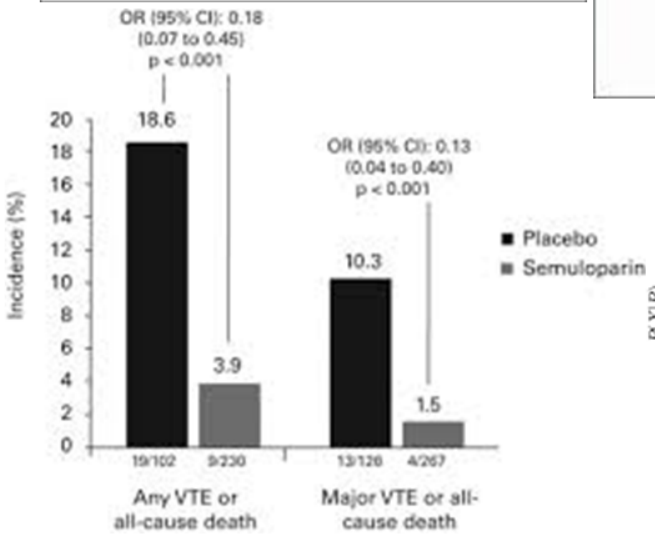
El odds ratio (generalmente abreviado OR) es una de las tres formas principales para cuantificar la intensidad de la presencia o ausencia de la propiedad A asociada con la presencia o ausencia de la propiedad B en una población determinada. Esta relación (OR) se puede calcular siguiendo estos pasos:

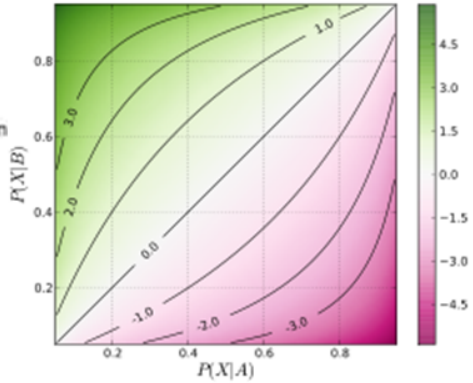
- Para un individuo dado que cuenta con B calcular la probabilidad de que el mismo individuo tienga A.
- Para un individuo dado que no cuenta con B calcular la probabilidad de que el mismo individuo tienga A.
- Dividir la probabilidad del paso 1 por la probabilidad del paso 2 para obtener el cociente de probabilidades (OR).

Odds ratio


Impact of Treatment on Mortality by Study
Odds Ratio and 95% CL
OR LCL UCL WT
Modano (1967)
Borodan (1981)
Faymiy (2005)
Intermediate
Modano (1969)
Soloway (2000)
Adams (1999)
Truark2 (2002)
Overall
Favors Treatment
Favors Control
0.01
0.1
1
10
100


Odds Ratio (OR)
Contingency (or 2 x 2) Table
Cases
Controls
Total
Exposed
a
b
a+b
Unexposed
c
d
c+d
Total
a+c
b+d
a+b+c+d
OR = (a/c) / (b/d)
= (a*d) / (b*c)


OR (95% CI): 0.18
(0.07 to 0.45)
p < 0.001
OR (95% CI): 0.13
(0.04 to 0.40)
p < 0.001
Incidence (%)
18.6
3.9
10.3
1.5
19/102
9/230
13/126
4/267
Placebo
Semuloparin
Any VTE or all-cause death
Major VTE or all-cause death


P(X|B)
P(X|A)

Contactos soporte analítico

- Guadalupe Rivera García, guadalupe.rivera@sat.gob.mx
- Edgar Díaz Ordoñez, edgar.diaz@sas.com
- Rita Guerrero Cervates, esmeralda.guerrero@sas.com
- Luis Javier Herrera Orduña, luisjavier.herrera@sas.com
- Mauricio Vallejo Castañon, mauricio.vallejo@sasmexico.com
- Daniel Arébalo Rojo, daniel.arevalo@sasmexico.com
- Yesenia Alejandro de la Cruz, yesenia.alejandro@sasmexico.com

Extensión 44580 o 22554.