Resumen Inferencia y Modelos Estadísticos

Introducción

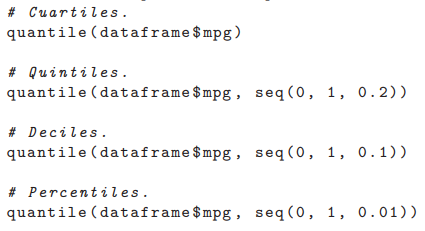
* Población: conjunto de individuos o elementos de los cuales se busca una conclusión
* Muestra: Subconjunto de la población

Cap. 2: Conceptos básicos

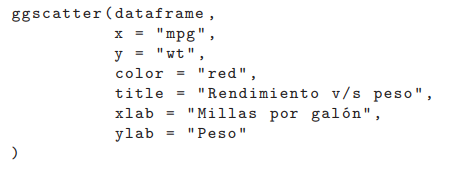
* Nociones de datos
  + Los datos se almacenan en matrices
  + Existen tipos de variables, de las cuales no todas pueden tomar los mismos valores.
    - Numéricas: pueden tomar muchos valores numéricos
      * Continuas: cualquier valor en un intervalo del conjunto de Reales
      * Discretas: valores enteros no negativos
    - Categóricas: solo pueden tomar un valor en un conjunto acotado. Cada valor se denomina nivel.
      * Nominales: no existe un orden natural entre los niveles.
      * Ordinales: existe un orden natural entre los niveles.
  + Dos variables pueden ser:
    - Independenties
    - Dependientes
      * Asociación positiva: directamente proporcional
      * Asociación negativa: inversamente proporcional
  + Parámetro: cualquier número que describa una población en forma resumida (ej, promedio)
  + Estadístico: cualquier cantidad cuyo valor puede ser calculado a partir de datos muestrales (ej, media) (es una estimación del parámetro)
* Nociones de R
  + Importación de datos
    - Importar una matriz de datos (data frame) desde un txt o csv hay que considerar:
      * La primera fila son para nombres de las columnas o variables
      * La primera columna para nombres de las observaciones (únicos), solo se permite el uso de puntos y guiones bajos. Los nombres no deben empezar con un dígito.
    - Distingue mayúsculas
    - Sin filas en blanco
    - Sin comentarios
    - Si no hay valores, debe contener un NA
    - Formato fecha mm/dd/aaaa
    - setwd(“directorio”) permite establecer el directorio de trabajo de R
    - head(): muestra por consola las primeras 6 filas de la data frame
    - tail(): muestra por consola las últimas 6 filas de la data frame
  + Importación de paquetes:
    - Antes de utilizar un paquete, este debe ser instalado con install.packages(“nombre\_paquete”)
    - Para utilizar el paquete se debe colocar library(nombre\_paquete) o require(nombre\_paquete)
  + Construcción de una data frame:
    - As.Date() para dar formato de fecha
    - nombre <- c(“Pedro”, “Juan”, “Diego”) : crea un vector que contiene los nombres.
    - fecha\_nacimiento <- as.Date(c(" 2 0 0 8 -1 -2 5 " , " 20 0 6 -1 0 -4 ", " 2 0 0 8 -3 -2 7 ") crea un vector con fechas
    - Dataframe <- data.frame(nombre, fecha\_nacimiento) crea un dataframe

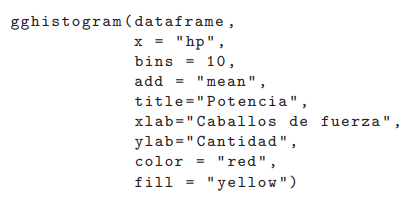
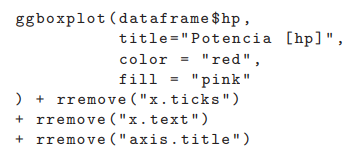
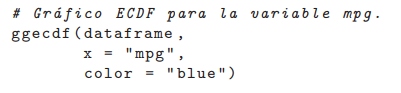
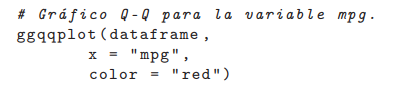
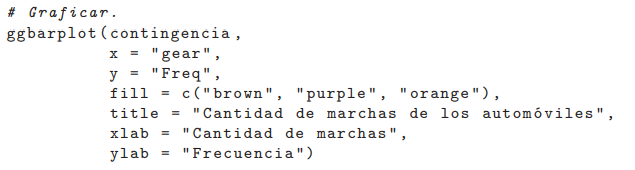
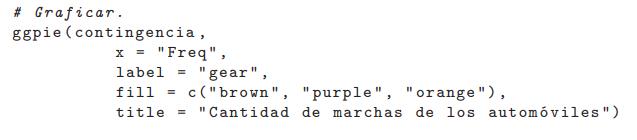
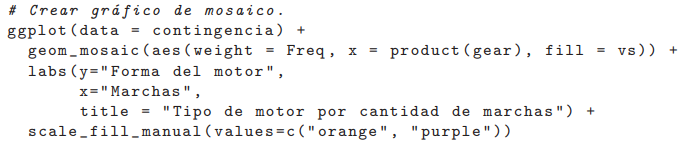
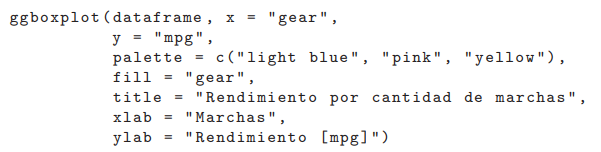
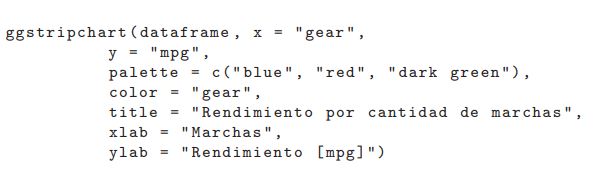
Cap. 3: Exploración de datos

* Datos numéricos
  + Estadísticas descriptivas
    - Cuando una medida se aplica a una muestra, corre=sponde a un estimador puntual.
    - Distribución de frecuencia: representa cuántas veces aparece cada valor para una variable.
    - Media (media aritmética o promedio)
      * Media muestral ()
      * Media poblacional (μ­x)
      * mean(nombre\_dataframe$variable) para calcular el promedio de una variable
      * sapply(nombre\_dataframe, mean) para calcular el promedio de todas las variables
    - Mediana: valor central de los valores ordenados
      * median()
    - Moda: valor que más se repite
      * Unimodal
      * Bimodal
      * Multimodal
      * Paquete modeest tiene la función mfv() para calcular la moda
    - Varianza y desviación estandar:
      * Se calcula en base a la desviación de las observaciones.
        + Distancia entre una observación y la media del conjunto de datos
      * La desviación estandar es útil cuando se necesita saber cuán cercanos son los datos a la media
      * var() para calcular varianza y sd() para desvianción estandar
    - Rango:
      * Muestra el mínimo y máximo de una variable
    - Rango intercuartil (IQR):
      * Cada fragmento de datos dvidido en partes iguales se denomina cuantil.
        + Percentil: 100 subconjuntos de igual tamaño
        + Deciles: 10 subconjuntos
        + Quintiles: 5 subconjuntos
        + Cuartiles: 4 subconjuntos
      * Los cuatiles se nombran de forma ascendente (percentil 1 es el del valor más pequeño)
      * quantile(nombre\_dataframe$variable) para calcuar cuantiles



* + - * seq(inicio, término, incremento) genera una secuencia de números equiespaciados
      * IQR() para calcular un rango intercuartil
    - Desviación absoluta promedio (MAD):
      * Es el promedio de la desviación de cada observación con respecto a la mediana
      * mad()
  + Funciones de interés en R
    - summary() entrega la media, mediana, el primer y tercer cuartil, el mínimo y máximo.
    - El paquete pastecs tiene la función stat.desc(), que entrega la media, varianza y desviación estandar.
  + Estimadores robustos
    - Valores atípicos o outliers
      * Observaciones fuera de rango o muy extremas con respecto al resto de datos.
    - La mediana es una buena medida de tendecia central y el IQR buena medida de disperción.
    - MAD aún más robusta que IQR
  + Representación gráfica de datos numéricos
    - Gráficos creados con el paquete ggpubr
    - Gráfico de dispersión
      * Cada punto del gráfico corresponde a una observación



* + - * Sirve para ver dependencias entre variables
    - Gráfico de puntos
      * Útil cuando solo se estudia una variable y la muestra es pequeña
      * Gráfico de disperción para una variable
      * Suele añadirse una señal para la media
    - Histograma
      * Útil con muestras grandes
      * Rango de valores se divide en intervalos
      * Reflejan densidad de datos
      * 
      * Permite visualizar la distribución de frecuencias
      * Distribución desviada a la izquierda o asimetría negativa: observaciones concentradas en la izquierda. Es analogo a la derecha
      * Simetrica: cuando las observaciones se aglomeran hacia el centro
    - Gráfico de caja
      * Su construcción considera 5 estadísticos para representar el conjunto de datos y facilita la identificación de datos atípicos
      * 
      * Los extremos de inferior y superior de la caja corresponden al 1er y 3er cuartil
      * La linea al interior corresponde a la mediana
      * Su altura corresponde al rango intercuartil
      * Las barras fuera de la caja son llamadas bigotes
        + Capturan datos fuera de la caja y a no mas de 1,5 veces el IQR
      * Cualquier punto fuera de la caja es atípico
    - Función acumulativa de distribución empírica (ECDF)
      * Para muestras grandes se aproxima a la distribución de probabilidad real de la población
      * Ordena el conjunto de datos de manera no decreciente y luego asigna una probabilidad de a cada dato individual. Luego suma las probabilidades de cada dato y los anteriores a él
      * 
    - Gráfico cuantil-cuantil (Q-Q)
      * Permite verificar si la distibución de datos se acerca a la distribución normal de probabilidad
      * 
* Datos categóricos
  + Tablas de contingencia, matriz de confusión o tabla de frecuencias
    - Cada fila representa la cantidad de veces que ocurre una combinación de variables
    - Tabla de frecuencias relativas: Se usan porcentajes o proporciones
    - Tabla de contingencia para una variable
      * xtabs(formula) muestra el nombre de la variable tabulada al imprimir los resultados
      * marginSums() permite calcular los totales por filas
      * addmargins() permite calcular los totales e incorporarlos en la tabla
    - Tabla de contingencia para dos variables
      * Para determinar proporciones, se debe dividir el valor de una celda por el total de su fila o columna
    - Tabla de contingencia para más de dos variables
      * Se contruye una subtabla por cada nivel de la tercera variable, cada una de las variables muestra las otras dos variables
  + Representación gráfica de datos categóricos
    - Gráfico de barras
      * Para representar una variable categórica
      * Cada barra es tan larga como la proporción de valores presentes en cada nivel de la variable
      * 
    - Gráfico de torta
      * Alternativa para representar una variable categórica
      * 
    - Gráficos de barras segmentadas y barras agrupadas
      * Permiten visualizar la tabla de proporciones entre 2 variables y encontrar posibles relaciones entre ellas
    - Gráfico de mosaico
      * Divide un área en regiones para representar la cantidad de observaciones de cada región
      * Se requiere el paquete ggmosaic
      * 
* Datos agrupados
  + Estadísticas descriptivas para datos agrupados
    - Se utiliza el paquete splyr
    - gruop\_by()
    - summarise(cantidad\_observaciones, diferentes\_estadisticas\_descriptivas\_calculables)
    - pipe: operador %> % cuya función es entregar un valor o el resultado de una expresión a la siguiente llamada a una función. x %> % f ⬄ f(x)
  + Representación gráfica de datos agrupados
    - Utilizados para comparar diferentes grupos de observaciones de acuerdo a una característica categórica
    - Gráfico de cajas
      * Recibe una variable categórica para el eje x y otra numérica para el eje y
    - Gráfico de tiras
      * Se utiliza cuando se tienen pocas observaciones en cada grupo
* Cap. 4: Variables aleatorias
  + Distribuciones discretas parte 1
    - Variable aleatoria: variable o proceso con resultado numérico. Denotada con letra mayúscula y valores con letra minúscula.
      * Distribución de probabilidad: probabiliadad de que ocurran diferentes valores.
    - Valor esperado (E(X) o μ): resultado promedio de una variable aleatoria.
    - Varianza general (Var(X) o σ2): qué tan alejado podría estar un valor obtenido del valor esperado
    - Paquete DiscreteRV
    - Útil cuando se desea conocer la distribución del tiempo de ejecución de un programa
  + Combinaciones lineales de variables aleatorias
    - Para representar un fenómeno como una combinación de dos o más variables aleatorias
  + Distribuciones continuas
    - Función de densidad de probabilidad (distribución o densidad): curva continua (campana de gauss)
    - Distribución normal (distribución gaussiana)
      * Muchas variables se acercan a esta distribución
      * Es unimodal y simétrica con forma de campana
      * Se usa para modelar diversos fenomenos y se ajusta mediante dos parámetros
        + μ: la media, que desplaza el centro de la curva a lo largo del eje x
        + σ: la desviación estándar, que modifica su extención
      * N(μ, σ)
      * dnorm(rango\_valores(vector), μ, σ) calcula la densidad de la disvtribución normal
      * técnicas de estandarización: para determinar cuan usal es un dete. valor en una escala única
        + valor z: cuan encima o debajo de la media se encuentra una observación x

para encontrar percentiles correspondientes a la función de distribución

* + - * + pnorm(vector\_valores, mean - 0, sd = 1): prob de que se tome un valor mayor al entregado en q
        + qnorm(prob\_acumulada, mean - 0, sd = 1)
        + rnorm(n, mean - 0, sd = 1): genera un vector con n observaciones dentro de la distribución normal
        + Regla

Cerca de 68 % de las observaciones se encuentran a una distancia de una desviación estándar de la media.

Alrededor de 95 % de las observaciones se encuentran a una distancia de dos desviación estándar de la media.

Aproximadamente 99.7 % de las observaciones se encuentran a una distancia de tres desviación estándar de la media

* + - * Distribución chi-cuadrado (ji-cuadrado o X2)
        + Para caracterizar valores siempre positivos y desviados a la derecha
        + Parámetro: grados de libertad (v)

Estimación de observaciones usadas para calcular un estimador

* + - * + Cómo la cantidad de valores pueden cambiar en un conjunto de datos
        + μ = v, σ = 2v
        + dchisq(x, df).
        + pchisq(q, df, lower.tail).
        + qchisq(p, df, lower.tail).
        + rchisq(n, df).

x, q son vectores de cuantiles (enteros no negativos).

p es un vector de probabilidades.

n es la cantidad de observaciones.

df son los grados de libertad.

lower.tail es análogo al de la función pnorm

* + - * Distribución t de Student
        + Utilizada con muestras pequeñas
        + Parámetro: grados de libertad (v)
        + + v + semejante a la normal
        + Para v > 1, μ = 0.

dt(x, df).

pt(q, df, lower.tail).

qt(p, df, lower.tail).

rt(n, df).

* + - * Distribución F
        + df(x, df1, df2).
        + pf(q, df1, df2, lower.tail).
        + qf(p, df1, df2, lower.tail).
        + rf(n, df1, df2).
  + Distribuciones discretas parte 2
    - Distribución Bernoulli
      * Variable aleatoria de Bernoulli: en cada intento individual tiene solo dos resultados: éxito (p) o fracaso (1-p)
      * Proporción de la muestra:
      * μ = p
      * dbern(x, prob).
      * pbern(q, prob, lower.tail).
      * qbern(p, pro, lower.tail).
      * rbern(n, prob).
    - Distribución geométrica
      * Describe la cant de intentos que se deben realizar para obtener un éxito para variables indepe. e idénticamente distribuidas
        + Que las variables no se afectan unas a otras y c/u tiene la misma prob de éxito
      * dgeom(x, prob).
      * pgeom(q, prob, lower.tail).
      * qgeom(p, prob, lower.tail).
      * rbern(n, prob).
    - Distribución binomial
      * Describe la probabilidad de tener k éxistos en n intentos indepe.
      * Antes de decidir usar la distribución binomial, es necesario verificar cuatro condiciones:

1. Los intentos son independientes.
2. La cantidad de intentos (n) es fija.
3. El resultado de cada intento puede ser clasificado como éxito o fracaso.
4. La probabilidad de éxito (p) es la misma para cada intento.
   * + - dbinom(x, size, prob).
       - pbinom(x, size, prob).
       - qbinom(p, size, prob).
       - rbinom(n, size, prob).
         * x es un vector numérico.
         * p es un vector de probabilidades.
         * n es la cantidad de observaciones.
         * size corresponde al número de intentos.
         * prob es la probabilidad de éxito de cada intento
     + Distribución binomial negativa
       - Describe la prob. de encontrar el k-ésimo éxito al n-ésimo intento. Se necesitan 4 condiciones:
         * Los intentos son independientes.
         * El resultado de cada intento puede ser clasificado como éxito o fracaso.
         * La probabilidad de éxito (p) es la misma para cada intento.
         * El último intento debe ser un éxito
       - dnbinom(x, size, prob, mu).
       - pnbinom(q, size, prob, lower.tail).
       - qnbinom(p, size, prob, lower.tail).
       - rnbinom(n, size, prob, mu).
         * x, q son vectores de cuantiles (enteros no negativos).
         * p es un vector de probabilidades.
         * n es la cantidad de observaciones.
         * size corresponde al número (no negativo) de intentos.
         * prob es la probabilidad de éxito de cada intento.
         * lower.tail es análogo al de la función pnorm.
     + Distribución de Poisson
       - Útil para estimar la cant. de eventos en una población grande en un lapso de tiempo dado
       - dpois(x, lambda).
       - ppois(q, lambda, lower.tail).
       - qqpois(p, lambda, lower.tail).
       - rpois(n, lambda).
         * x, q son vectores de cuantiles (enteros no negativos).
         * p es un vector de probabilidades.
         * n es la cantidad de observaciones.
         * lambda es un vector no negativo de medias.
         * lower.tail es análogo al de la función pnorm.