Curso de R y estadística básica

[Felipe de J. Muñoz González]

fmunoz@lcg.unam.mx

Descripción de Datos y Probabilidad

Descargar Presentación

Descripción de Datos en Estadística

¿Tipos de Datos?

datum se refiere a la información concreta, cualquier pieza de información colectada.

Un **"Data set"** o set de datos es una colección de datos relacionadas de alguna forma.

Se definen 5 tipos de datos:

- Cuantitativos
- Cualitativos
- Logicos
- Faltantes
- Otros tipos

¿Tipos de Datos?

- Datos cuantitativos

Son datos que se pueden medir o son asociados a alguna cantidad.

Se subdividen en:

- Datos discretos
- Datos continuos (datos escalares o de intervalos)

Nota Cuando no se sabe que tipo de dato es, considerese continuo

¿Tipos de Datos?

- Datos cuantitativos (Ejemplo)

Ejemplo Precipitaciones anuales en ciudades de EE.UU. El vector contiene la cantidad promedio de lluvia (en pulgadas) para cada una de las 70 ciudades de los Estados Unidos.

```
> str(precip)
```

```
> precip[1:4]
```

Ejercicio

Describir los datos dentro de los dataset "rivers" y "discoveries"

¿Tipos de Datos?

- Datos cuantitativos

- Gráficas de puntos

Una de las cosas básicas que debe de manejarse cuando se describen los datos son gráficas que nos permitan tener mas información.

1. Graficas de puntos (Strip charts). Existen 3 metodos:
   - overplot
   - jitter
   - stack

```
> stripchart(precip, xlab = "rainfall")
```

```
> stripchart(rivers, method = "jitter", xlab = "length")
```

```
> stripchart(discoveries, method = "stack", xlab = "number")
```

¿Tipos de Datos?

- Datos cuantitativos

- Histogramas

1. Histogramas (Bar Graphs)

Normalmente se usan para datos continuos y se requiere decidir un conjunto de clases o compartimientos que dividen la linea real en un conjunto de cajas a los cuales caen los valores.

```
> hist(precip, main = "Histograma de lluvias en U.S.A")
```

```
> hist(precip, freq = FALSE, main = "") #Frecuencias Relativas
```

Consideraciones:

- La gráfica depende de los "bins" elegidos

¿Tipos de Datos?

- Datos cuantitativos

- Histogramas

Ejercicios

- Genera dos histogramas de los datos de precipitación, el primero con 10 divisiónes y el segundo con 200

¿Tipos de Datos?

- Datos cuantitativos

- Ejercicios

Ejercicios

- Genera dos histogramas de los datos de precipitación, el primero con 10 divisiónes y el segundo con 200

```
> hist(precip, breaks = 10, main = "")
```

```
> hist(precip, breaks = 200, main = "")
```

¿Tipos de Datos?

- Datos cuantitativos

- Gráficas de tallo

Definición

Las Gráficas de tallo tienen dos partes básicas: tallos y hojas. El último dígito de los valores de datos se toma como una hoja y el (los) dígito (s) principal (es) se toma (n) como tallos.

Ejemplo UKDriverDeaths serie de datos en el tiempo que contiene las muertes en accidentes automovilisticos o con lesiones fuertes en Reino Unido de Enero de 1969 a Diciembre de 1984. ?UKDriverDeaths.

```
> install.packages("aplpack")

> library(aplpack)

> stem.leaf(UKDriverDeaths, depth = FALSE)
```

¿Tipos de Datos?

- Datos cuantitativos

- Gráficas de Índice

Estas se realizan utilizando la función **plot** y son buenas para visualizar datos que han sido ordenados, cuando los datos fueron medidos a traves del tiempo.

Es una gráfica de dos dimensiones que tiene una variable índice (x) y una variable medida (y).

Existen los siguientes métodos:

- picos (spikes). code: (type = "h")
- puntos (points) code: (type = "p"=)

Ejemplo Mediciones anuales (En pies) del lago Huron de 1875-1972. Los datos son en el tiempo. ?LakeHuron

```
> plot(LakeHuron, type = "h")
```

```
> plot(LakeHuron, type = "p")
```

¿Tipos de Datos?

- Datos cualitativos

Datos **no numericos** o que no representan cantidades numericas.

Ej. Nombre, genero, grupo etnico, estado socioeconomico, numero de seguridad social, licencia, ...

Algunos datos parecen ser cuantitativos pero no lo son por que no representan cantidades numericas medibles ni conservan reglas matemáticas.

Ej. Tamaño del pie de una persona (si sumas el tamaño del pie de dos personas no tiene sentido)

La información cuantitativa que se puede utilizar para subdividir información en diversas categorias se le llama **factor**

¿Tipos de Datos?

- Datos cualitativos

- Presentación de Datos

Tablas Una forma de mostrar resumenes de datos estadisticos es con el uso de las tablas.

```
> str(state.abb)
```

Frecuencias absolutas

```
> Tbl <- table(state.division)
> Tbl
```

Frecuencias Relativas

```
> Tbl/sum(Tbl)
```

```
> Tbl/sum(Tbl)
```

¿Tipos de Datos?

- Datos cualitativos

- Descripción

Los datos de state.region enumera cada uno de los 50 estados y la región a la que pertenece, ya sea en el noreste, sur, norte central u oeste.

```
> str(state.region)
```

```
> state.region[1:5]
```

```
> str(state.abb)
```

Frecuencias absolutas

```
> Tbl <- table(state.division)
> Tbl
```

Frecuencias Relativas

```
> Tbl/sum(Tbl)
```

```
> prop.table(Tbl) # same thing
```

¿Tipos de Datos?

- Datos cualitativos

- Gráficas de Barras

Un gráfico de barras es el análogo de un histograma para datos categóricos. Se muestra una barra Para cada nivel de un factor, con las alturas de las barras proporcionales a las frecuencias de observaciones Pertenecientes a las respectivas categorías. Una desventaja de los gráficos de barras es que los niveles están ordenados alfabéticamente (por defecto), lo que a veces puede oscurecer los patrones en la pantalla.

```
> barplot(table(state.region), cex.names = 0.5)
```

```
> barplot(prop.table(table(state.region)), cex.names = 0.5)
```

¿Tipos de Datos?

- Datos cualitativos

- Diagramas de Pareto

Un diagrama pareto es muy parecido a un gráfico de barras excepto que las barras se reordenan de tal manera que disminuyen en altura, pasando de izquierda a derecha. La reorganización es útil porque puede revelar visualmente la estructura (si es que hay) en la velocidad de las barras disminuyen - esto es mucho más difícil cuando las barras se mezclan.

```
> library(qcc)
```

```
> pareto.chart(table(state.division), ylab = "Frequency")
```

¿Tipos de Datos?

- Datos cualitativos

- Gráfica de puntos

Estos se parecen mucho a un gráfico de barras que se ha girado en su lado con las barras reemplazadas por puntos en líneas horizontales. No transmiten más (o menos) información que el gráfico de barras asociado, pero la fuerza reside en la economía de la pantalla. Los gráficos de puntos son tan compactos que es fácil graficar interacciones multi-variables muy complicadas en un gráfico.

```
x <- table(state.region)
```

```
> dotchart(as.vector(x), labels = names(x))
```

¿Tipos de Datos?

- Datos cualitativos

- Gráfica de pastel

These can be done with R and the R Commander, but they fallen out of favor in recent years because researchers have determined that while the human eye is good at judging linear measures, it is notoriously bad at judging relative areas.

```
> slices <- c(10, 12,4, 16, 8)
> lbls <- c("US", "UK", "Australia", "Germany", "France")
```

```
> pie(slices, labels = lbls, main="Pie Chart of Countries"
```

¿Tipos de Datos?

- Datos cualitativos

- Gráfica de pastel

These can be done with R and the R Commander, but they fallen out of favor in recent years because researchers have determined that while the human eye is good at judging linear measures, it is notoriously bad at judging relative areas.

```
> slices <- c(10, 12,4, 16, 8)
> lbls <- c("US", "UK", "Australia", "Germany", "France")
```

```
> pie(slices, labels = lbls, main="Pie Chart of Countries"
```

Distribuciones de datos

Centroide: Conjunto de datos está asociado con un número que representa una tendencia media o general de los datos.

La **Dispersión** de un conjunto de datos está asociada con su variabilidad; Los conjuntos de datos con una dispersión grande tienden a cubrir un gran intervalo de valores, mientras que los conjuntos de datos con dispersión pequeña tienden a agruparse fuertemente alrededor de un valor central.

Forma: Forma exhibida por una pantalla gráfica asociada. La forma puede decirnos mucho sobre cualquier estructura subyacente a los datos, y puede ayudarnos a decidir qué procedimiento estadístico debemos usar para analizar los.

Distribuciones de datos

Forma

Simetría y **asimetría**

- **positivamente sesgada**
- **negativamente sesgada**

La **curtosis** (o apuntamiento) es una medida de forma que mide cuán escarpada o achatada está una curva o distribución.

Observaciones Extremas

Problemas que pueden implicar estimaciones exageradas e "inestabilidad" estadística. Podemos considerar que estos datos pueden ser:

- Error tipográfico (typoo)
- Observaciónes que no eran para el estudio. (Ej. Complicaciones medicas)
- Indican un fenomeno o una tendencia más profunda

Estadistica descriptiva

Rangos intercuantiles y MAD

suceptibilidad de la media, mediana a valores extremos.

Rango intercuartil (**IQR**) definido por IQR = q*{0.75}* - *q*{0.25}

Otro método más robusto que el IQR es la Media de la desviación absoluta (**MAD**).

1. Calculamos la media (prom)
2. mediana(|x{i} - prom(X)|) , para toda i

Estadistica descriptiva

Utilizando R. Calcula las siguientes cosas del vector

```
x<-round(runif(20, min=1, max=100))
```

- rango
- media
- mediana
- quantiles/quintiles/septiles
- varianza
- desviación estandar

Nota Rcmdr

Statistics > Summaries > Numerical Summaries

calculamos los cuantiles automaticamente

Estadistica descriptiva

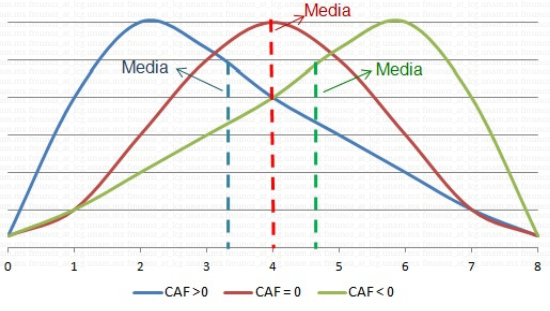
Medidas de Forma

La **asimetría** (Fisher) de la muestra, se define por la fórmula

$$CA_F = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^3}{Ns^3}$$

siendo $\overline{x}$ la media y s la desviación típica

donde S es la desviación estandar (o tipica)



Estadistica descriptiva

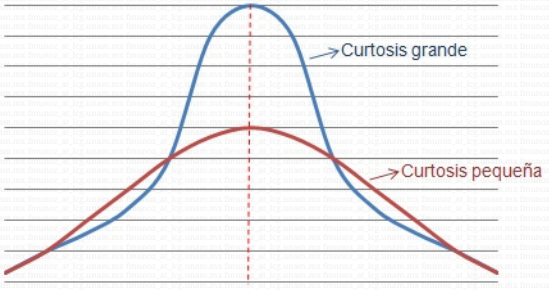
Medidas de Forma

La **curtosis** de la muestra, se define por la fórmula

$$Curtosis = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{Ns^4} - 3$$

siendo $\overline{x}$ la media y s la desviación típica

donde S es la desviación estandar (o tipica)



Estadistica descriptiva

Medidas de Forma

Asimetria

```
> library(e1071)
> skewness(discoveries)
```

```
> 2 * sqrt(6/length(discoveries))
```

Nota si 2 * sqrt(6/n) < skewness(x) => existe un sesgo dado el signo del calculo.

Curtosis

```
> kurtosis(UKDriverDeaths)
```

```
> 4 * sqrt(6/length(UKDriverDeaths))
```

Nota abs(4 * sqrt(6/n)) < kurtosis(x) => presenta curtosis

Estadistica descriptiva

Grafica de caja

Estas gráficas son buenas para visualizar mucha información descriptiva de nuestros datos al mismo tiempo:

Centroide (estimada por la mediana)

Dispersión

Forma

Observaciones extremas

Outliers Observaciones que pasan 1.5 veces el tamaño de la caja para cualquier extremo.

Para observar los valores outliers

```
> boxplot.stats(rivers)$out #1.5 default
```

```
> boxplot.stats(rivers, coef = 3)$out #coef=3
```

```
> boxplot(rivers, horizontal=T)
```

Estadistica descriptiva

Z-value

Valor estandarizado, cuando queremos comparar datos en escala que es independiente a la medida.

Dado X=x[1], x[2], x[3], ... ,x[n] los z-scores son z[1], z[2],..n se ven definidos como

z[i]=(x[i]-median(x))/s

donde s es la sd()

```
> ?scale
```

Datos Multivariados y DataFrames

Datos multivariados

Introducción

Los estudios estadísticos requieren mas de un factor o medición asociado a cada objeto, para esto utilizamos otra **estructura de datos**.

Para esto existen dos tipos de estructuras en R:

- Matrices
- DataFrames

Ambas son estructuras arreglas en dos dimensiones en forma rectangular y estaremos considerando que (a menos que se indique lo contrario):

- Las lineas son objetos
- Las columnas contienen diferentes mediciones o factores

Ejemplo:

```
> x <- 5:8
> y <- letters[3:6]
> z <- 1:4*pi
> A <- data.frame(v1 = x, v2 = y, v3=z)
```

Datos multivariados

Acceso a DataFrames

```
> A[3, ]
```

```
> A[1, ]
```

```
> A[, 2]
```

```
> names(A)
```

```
> A$v1
```

Datos multivariados

Matrices

```
> A = matrix(
+   c(2, 4, 3, 1, 5, 7), # the data elements
+   nrow=2,              # number of rows
+   ncol=3,              # number of columns
+   byrow = TRUE)        # fill matrix by rows

> A                      # print the matrix
```

```
> dimnames(A) = list(
+   c("row1", "row2"),         # row names
+   c("col1", "col2", "col3")) # column names
```

```
> A[, 2]
```

DataFrames vs Matrices:

Las **Matrices** son solamente arreglos numericos de dos dimensiones mientras que los **DataFrame** contienen diferentes tipos de valores

Probabilidad

Probabilidad

Espacio muestral

Moneda

```
> S<-data.frame(pos=c("H","T"))
```

Dado

```
> S<-data.frame(pos=c(1:6))
```

Espacio Muestral de una moneda

```
> install.packages("prob", dependencies=TRUE)
> library(prob)
> tosscoin(1, makespace=TRUE)
> tosscoin(3, makespace=TRUE)
```

Espacio Muestral de un dado

```
> rolldie(1, makespace=TRUE)
> rolldie(7, makespace=TRUE)
> rolldie(1, nsides=10, makespace=TRUE)
```

Probabilidad

Espacio muestral

Espacio Muestral de Cartas Inglesas

```
> cards(2, makespace=TRUE)
```

Espacio Muestral de Muestreo de urnas

```
> ?urnsamples
```

```
> urnsamples(x=c("roja","azul","amarilla","violeta","negra","
```

Probabilidad

Eventos

Evento con monedas

```
>  S <- tosscoin(2, makespace = TRUE)
>  S[c(2,4),]
```

Evento con cartas

```
> S <- cards()
> subset(S, suit == "Heart")
```

```
> subset(rolldie(3), X1 + X2 + X3 > 16)
```

Probabilidad

Subsets de datos

%in% #busqueda por elementos

```
> x <- 1:10
> y <- 8:12
> y %in% x
> y[y %in% x]
```

isin

```
> isin(x,y) #todo el vector
```

all

```
> x <- 1:10
> y <- c(3, 3, 7)
> all(y %in% x)

> isin(x, y)
```

¿Por que isin y all tienen esos resultados?

Otras funciones: **countrep** y **isrep**

Probabilidad

Union, Interseccion y diferencia

Elementos que existen en el Evento A, en el Evento B o en ambos union(A,B)

```
> S = cards()
> A = subset(S, suit == "Heart")
> B = subset(S, rank %in% 7:9)

> union(A, B)
```

Elementos que existen en el Evento A y en el Evento B intersect(A,B)

```
>  intersect(A, B)
```

Elementos que existen en el Evento A pero no en el Evento B setdiff(A,B)

```
> setdiff(A,B)
```

Nota setdiff no es simetrico y podemos calcular el complemento de todos los eventos Ej. setdiff(S,A)

Probabilidad

Probabilidades de frecuencias relativas

P(A) ≈ observados / posibles ≈ S_n/n

Utilizando la ley de Grandes Números:

S_n/n → IP(A) as n → ∞.

```
> probspace(1:6)
```

Ej. Moneda no balanceada

```
> probspace(tosscoin(1), probs = c(0.7, 0.3))
```

WARNING: RAM memory y probabilidades infinitecimales

- Espacio de probabilidad de tirar 100 monedas
- 50 Dados

Probabilidad

Conteo con urnas

	ordered = TRUE	ordered = FALSE
replace = TRUE	n^k	$\frac{(n-1+k)!}{(n-1)!k!}$
replace = FALSE	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Numeros Factoriales

```
> factorial(n)
```

Coeficiente binomial (Combinaciones)

```
> choose(n,k)
```

WARNING: RAM memory y probabilidades infinitecimales

- Espacio de probabilidad de tirar 100 monedas
- 50 Dados

Probabilidad

Problema del cumpleaños

¿Calcula la probablidad de que dos personas que esten en el mismo cuarto cumplan años el mismo dia?

```
> install.packages(pbirthday.ipsur)
> library(pbirthday.ipsur)
> g <- Vectorize (pbirthday.ipsur)
> plot (1:50 , g(1:50) ,
+ xlab = "Number of people in room",
+ ylab = "Prob(at least one match)",
+ main = "The Birthday Problem")
> abline(h = 0.5)
> abline(v = 23, lty = 2) # dashed line
```

Probabilidad

Probabilidad Condicional

```
> library(prob)
> S <- rolldie(2, makespace = TRUE) # assumes ELM
> head(S) # first few rows
```

```
> A <- subset(S, X1 == X2)
> B <- subset(S, X1 + X2 >= 8)
```

```
> prob(A, given = B)
> prob(B, given = A)
```

Probabilidad

Variables Aleatorias

Definición: Una variable aleatoria X es una función X:S -> R que asocia para cada w ∈ S exactamente X(ω) = x.

Se define como S todos los posibles resultados de el evento E

Ejemplo:

Definimos la variable aleatoria X como "numero de aguilas cuando se tira una moneda".

Por lo tanto si **S** es nuestro espacio muestral y **w** los sucesos posibles

w∈ S	AA	AS	SA	SS
X(w) = x	2	1	1	0

Probabilidad

Variables Aleatorias

Escribir una formula que define una variable aleatoria dentro de una función, agregando una columna a un data.frame.

```
> ?transform
> ?addrv
```

Ej. Tiramos un dado de 4 lados 3 veces y definimos nuestra variable U = X1 - X2 + X3

```
> S <- rolldie(3, nsides = 4, makespace = TRUE)
> S <- addrv(S, U = X1 - X2 + X3)
> head(S)
```

Ahora podemos preguntar, ¿Cual es la probabilidad de que U > 6?

```
Prob(S, U > 6)
```

That's all folks (for now)!

Slideshow created using remark.