Estadística como ciencia
Plataformas de trabajo
Vectores
Progresiones y secuencias
Funciones
Subvectores
Abordando 'NA'
Factores
Listas
Matrices

Estadística Descriptiva

Estructura de datos

MSc. Javier Linkolk López Gonzales

Facultad de Ingeniería y Arquitectura

21 de abril de 2020

Tabla de Contenido

1. Estadística como ciencia
2. Plataformas de trabajo
3. Vectores
4. Progresiones y secuencias
5. Funciones
6. Subvectores
7. Abordando 'NA'
8. Factores
9. Listas
10. Matrices

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

Es un conjunto de métodos utilizados para analizar datos. Esta ciencia puede ser aplicada en, prácticamente, todas las áreas del conocimiento humano. Algunos significados de Estadística son:

❶ Recolección de datos.
❷ Medidas resultantes de un conjunto de datos.
❸ Métodos utilizados en la recolección e intepretación de datos.

Las 3 grandes áreas de la Estadística:

Las 3 grandes áreas de la Estadística:

❶ Probabilidad. Estudio de la aleatoriedad e incerteza.

Las 3 grandes áreas de la Estadística:

❶ Probabilidad. Estudio de la aleatoriedad e incerteza.

❷ Estadística Descriptiva. Utiliza métodos para la recolección, organización, presentación, análisis y síntesis de datos obtenidos en una población o muestra. (Exploración)

Las 3 grandes áreas de la Estadística:

❶ Probabilidad. Estudio de la aleatoriedad e incerteza.

❷ Estadística Descriptiva. Utiliza métodos para la recolección, organización, presentación, análisis y síntesis de datos obtenidos en una población o muestra. (Exploración)

❸ Estadística Inferencial. Proceso de estimar información sobre una población a partir de los resultados observados en una muestra. (Previsión)

DATO $\neq$ OBSERVACIÓN

Tipos de variables:

CUALITATIVAS y CUANTITATIVAS

POBLACIÓN || MUESTRA

- Colección de todos los elementos de interés.
- Convencionalmente, denotado por **N**.
- Los números obtenidos a partir de la población son llamados **parámetros**.

- Subconjunto de la población.
- Convencionalmente, denotado por **n**.
- Los números obtenidos al trabajar con la muestra son llamados **estadísticos**.

POBLACIÓN:

La población es ❶ difícil de definir y ❷ difícil de observar.

MUESTRA:

La muestra es mucho más fácil de recolectar.

❶ Menos tiempo, ❷ Menos costo.

- Por eso se diseñan test estadísticos para trabajar con data incompleta.

- Casi siempre se trabajará con datos muestrales y se tomarán decisiones e inferencias basadas en ellos.

- Dado que las pruebas estadísticas generalmente se basan en datos de muestra, las muestras son clave para obtener información estadística precisa.

ALEATORIEDAD – – – – – – – – – – REPRESENTATIVIDAD

- Se recoge una muestra aleatoria cuando cada miembro de la muestra se elige de la población estrictamente al AZAR.
- Una muestra representativa es un subconjunto de la población que refleja con precisión los miembros de toda la población.

Ejemplo*

Entonces...

¿Cómo sacar una muestra que sea aleatoria y representativa?

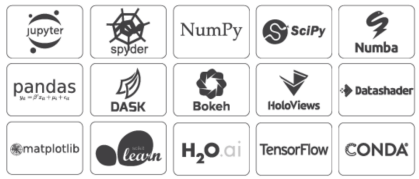
La forma más segura sería obtener acceso a la base de datos de estudiantes y contactar a las personas de manera aleatoria. Sin embargo, tales encuestas son casi imposibles de realizar sin la ayuda de la universidad.

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

- (r project) `https://cran.r-project.org/bin/windows/base/`
- (r studio) `https://rstudio.com/products/rstudio/download/`

- `https://www.anaconda.com/distribution/`



Recursos:

- Why use R? Five reasons

  `https://www.r-bloggers.com/why-use-r-five-reasons/`

- R Markdown Cheat Sheet

  `https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf`

- Support Markdown

  `https://wordpress.com/support/markdown-quick-reference/`

- StatsModels

  `https://www.statsmodels.org/stable/index.html`

- Latex

  `http://www.ptep-online.com/ctan/symbols.pdf`

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

Un vector es una secuencia ordenada de datos. En R se encuentran diferentes tipos de datos, entre ellos, están:

1. logical: TRUE, FALSE.
2. integer: números enteros $(\mathbb{Z})$.
3. numeric: números reales $(\mathbb{R})$.
4. complex: números complejos $(\mathbb{C})$.
5. character: palabras.

Todos los objetos deben ser del mismo tipo. Caso contrario, se usan las LISTAS.

Código en R para vectores:

- c(), define un vector.
- scan(), define un vector.
- fix(x), modifica visualmente el vector x.
- rep(i, n), define un vector constante, que contiene el dato i repetido n veces.

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

Una progresión aritmética es una sucesión de números tales que, la diferencia $\mathbf{d}$, de cualquier par de términos sucesivos de la secuencia es constante:

$$a_n = a_1 + (n-1) \cdot d.$$

- seq(i,f, by=d), progresión aritmética que va desde i hasta f, con una diferencia d.
- seq(i,f, length.out=n), progresión aritmética que va desde i hasta f, con una diferencia d y una longitud n. Esto es: $d = (b - a)/(n - 1)$
- seq(i,by=d, length.out=n), progresión aritmética que empieza en i, con una diferencia d y una longitud n. Esto es: $d = (b - a)/(n - 1)$.
- i:f, secuencia de números enteros $(\mathbb{Z})$ consecutivos, entre dos números i y f.

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

Una de las mejores formas de lograr tener mayor alcance en análisis de datos es escribir funciones. Estas permitirán automatizar diversas tareas comunes de una forma más poderosa y general que *copy and paste*.

Cuando se desea aplicar una función a cada uno de los elementos que componen el vector de datos, la función **sapply** ayuda a evitar programar con bucles en R:

- **sapply(nombre-vector, FUN=nombre-función)**, para aplicar dicha función a todos los elementos sin excepción, del vector.
- **sqrt(x)**, Calcula un nuevo vector con las raíces cuadradas de cada uno de los elementos del vector **x**.

Dado un vector de datos **x** se puede calcular diferentes medidas estadísticas respecto al mismo vector:

- **length(x)**, calcula la longitud del vector **x**.
- **max(x)**, calcula el máximo del vector **x**.
- **min(x)**, calcula el mínimo del vector **x**.
- **sum(x)**, calcula la suma de las entradas del vector **x**.
- **mean(x)**, calcula la media aritmética del vector **x**.
- **sort(x)**, ordena el vector **x** en orden natural de los objetos que lo componen.

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

vector[i], proporciona la $i-ésima$ entrada del vector.

- Los índices en **R** comienzan en 1.
- **vector[length(vector)]**, reporta la última entrada del vector.
- **vector[x:y]**, siendo **x** e **y** números naturales, reporta el subvector con las entradas del vector original desde la posición $x-ésima$ hasta la $y-ésima$.
- **vector[-x]**, siendo **x** un número, este subvector está conformado por todas las entradas del vector menos la entrada $x-ésima$.

Condicionales:

- **which(x condición)**, para obtener los índices de las entradas del vector **x** que satisface la condición.
- **which.min(x**, da la primera posición en la que el vector **x** toma su valor mínimo.
- **which(x==min(x))**, da todas las posiciones en las que el vector **x** toma sus valores mínimos.
- **which.max(x**, da la primera posición en la que el vector **x** toma su valor máximo.
- **which(x==max(x))**, da todas las posiciones en las que el vector **x** toma sus valores máximos.

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

Afortunadamente, existen funciones que abordan el tratamiento cuando existen NA = Not Avalaible dentro del vector de datos.

- sum(x, na.rm = TRUE).
- mean(x, na.rm = TRUE).
- is.na(x).
- which(is.na(x)).
- na.omit(x).

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

Factor, es como un vector, pero con una estructura más rica que permite usarlo para clasificar observaciones.

levels: atributo del factor

Cada elemento del factor es igual a un nivel. Los niveles clasifican las entradas del factor. Se ordenan por orden alfabético.

Para definir un factor, primero se define un vector y se lo transforma por medio de una de las funciones **factor()** o **as.factor()**.

Función **factor()**

factor(vector, levels=...), define un factor a partir del vector y dispone de algunos parámetros que permiten modificar el factor creado.

levels, permite especificar los niveles e incluso añadir niveles que no aparecen en el vector.
labels, permite modificar los nombres de los niveles.
levels(factor), obtiene los niveles del factor.

Factor ordenado

Es un factor donde los niveles siguen un orden.
ordered(vector, levels=...), define un factor ordenado y tiene los mismos parámetros que **factor**.

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

Lista formada por diferentes objetos, no necesariamente del mismo tipo, cada cual con un nombre interno:

- **list(...)**, función que crea una lista.
- Para obtener una componente específica, se usa el comando **list$componente**.
- También se puede indicar el objeto por su posición usando dobles corchetes **list[[i]]** .

Información de una list

- **str(list)**, permite conocer la estructura interna de una lista.
- **names(list)**, reporta los nombres de la lista.

Tabla de Contenido

1 Estadística como ciencia
2 Plataformas de trabajo
3 Vectores
4 Progresiones y secuencias
5 Funciones
6 Subvectores
7 Abordando 'NA'
8 Factores
9 Listas
10 Matrices

Cómo definirlas

- **matrix(vector, nrow=n, byrow=valor-lógico)**, define una matriz de n filas formada por las entradas del vector.
- **R** muestra las matrices indicando como $[i,]$ la fila $i-ésima$ y $[,j]$ la columna $j-ésima$.

☞ Todas las entradas de la matriz tienen que ser del mismo tipo de datos.

Cómo construirlas

- **rbind(vector1, vector2,...)**, construye la matriz de filas vector1, vector2.
- **cbind(vector1, vector2,...)**, construye la matriz de columnas vector1, vector2.
  ☞ Los vectores deben tener la misma longitud.
- **diag(vector)**, construye una matriz diagonal con un vector dado.

Funciones para matrices:

- **diag(matriz)**.
- **nrow(matriz)**.
- **ncol(matriz)**.
- **dim(matriz)**.
- **sum(matriz)**.
- **prod(matriz)**.
- **mean(matriz)**.