Estadística Descriptiva

Daniel Fraiman
dfraiman@udesa.edu.ar

Tipos de variables y su distribución

Variables

Ejemplos:

- sexo: M – F
- edad: 0,1,2,3,…
- religión
- nro. de hijos
- nivel de educación

Variables cualitativas y cuantitativas

- Una variable se dice cualitativa si su rango es un conjunto de categorías sin un orden preestablecido.
- Cuando los diferentes valores que toma la variable difieren en magnitud (existe una distancia) se dice que la variable es cuantitativa.

Entre medio de las cualitativas y las cuantitativas

- Escala ordinal: Hay un orden natural entre las categorías pero no una noción de distancia entre los distintos valores.
- EJEMPLOS: Nivel educativo, nivel socioeconómico, etc.

Los métodos estadísticos que vamos a estudiar dependen del rango de la variable.

Toy example:

Hospital Fernández

Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora en ser atendido
Juan Perez	Arenales 843	07/09/18	Médica	07:32	08:04
Romina Paz	Malabia 1820	07/09/18	Emergencia	09:35	09:46

Hospital Argerich

Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora en ser atendido
Pedro Ast	Zapiola 2232	07/09/18	Emergencia	07:32	08:34
Graciela Fort	Castillo 156	07/09/18	Médica	09:45	10:26

Hospital Álvarez

Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora en ser atendido
Lía Sope	Bogotá 1566	07/09/18	Emergencia	05:56	06:02
Carlos Seguí	Caracas 578	07/09/18	Emergencia	10:35	11:12

Centro de atención	Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora en ser atendido
H. Fernández	Juan Perez	Arenales 843	07/09/18	Médica	07:32	08:04
H. Fernández	Romina Paz	Malabia 1820	07/09/18	Emergencia	09:35	09:46
H. Argerich	Pedro Ast	Zapiola 2232	07/09/18	Emergencia	07:32	08:34
H. Argerich	Graciela Fort	Castillo 156	07/09/18	Médica	09:45	10:26
H. Álvarez	Lía Sope	Bogotá 1566	07/09/18	Emergencia	05:56	06:02
H. Álvarez	Carlos Seguí	Caracas 578	07/09/18	Emergencia	10:35	11:12

Estudiamos la variable *Centro de atención*:
¿cómo representar gráficamente esta información?

Variable categórica

Centro de atención
H. Fernández
H. Fernández
H. Argerich
H. Argerich
H. Álvarez
H. Álvarez

Frecuencia
8000
6000
4000
2000
0
Alvarez
Zubizarreta
Penna
Argerich
Tornu
Santojanni
Fernandez
Velez_Sarsfield
Grierson
Durand
Pirovano
Rivadavia
Ramos_Mejia
Piñero

Durand
Argerich
Alvarez
Zubizarreta
Velez_Sarsfield
Tornu
Santojanni
Rivadavia
Ramos_Mejía
Piñero
Pirovano
Penna
Grierson
Fernandez

¿Cómo resumir esta información en una tabla ?

Hospital	Frecuencia
Alvarez	8000
Argerich	5600
Grierson	3200
Durand	2500
Fernandez	4300
Penna	6500
Piñero	1200
Pirovano	2400
Ramos_Mejía	1900
Rivadavia	2400
Santojanni	5200
Tornu	5300
Velez_Sarsfield	4280
Zubizarreta	7230

Variable categórica

Consulta
Médica
Emergencia
Emergencia
Médica
Emergencia
Emergencia

Para las variables categóricas no tenemos más alternativas

Un comentario tonto

Datos categóricos:
Elecciones Presidenciales 2007

Variables categóricas

Diagrama de sectores ("torta")

Diagrama de barras

Para las variables categóricas no tenemos más alternativas

Sin embargo, si la variable es cuantitativa veremos que tenemos muchas maneras de presentar los datos.

Variables numéricas

Centro de atención	Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora en ser atendido
H. Fernández	Juan Perez	Arenales 843	07/09/18	Médica	07:32	08:04
H. Fernández	Romina Paz	Malabia 1820	07/09/18	Emergencia	09:35	09:46
H. Argerich	Pedro Ast	Zapiola 2232	07/09/18	Emergencia	07:32	08:34
H. Argerich	Graciela Fort	Castillo 156	07/09/18	Médica	09:45	10:26
H. Álvarez	Lía Sope	Bogotá 1566	07/09/18	Emergencia	05:56	06:02
H. Álvarez	Carlos Seguí	Caracas 578	07/09/18	Emergencia	10:35	11:12

Estudiamos la variable *Hora de llegada*:
¿cómo representar gráficamente esta información?

Variable numérica

Hora de llegada
07:32
09:35
07:39
09:45
07:42
08:35
07:25
07:53
08:12

HISTOGRAMA

La idea es presentar la información de la variable de interés de alguna forma gráfica perdiendo poca información.

Es practicamente una aproximacion a la FUNCION DE DENSIDAD de los datos que se estan mostrando.

Distintas formas de un histograma

Pirámides de población 1970, 2005 y 2050

Fuente: (INE)

http://www.censo2010.indec.gov.ar/definitivostotalxpais.asp

4,1+4,1+4,3+….+0,2+0,01+4,2+4,3+4,4…+0.1=100

uente: INDEC. Censo Nacional de Población, Hogares y Viviendas 2010.

Preguntas típicas:

- ¿Qué porcentaje de los datos se encuentran ….?

El histograma es la mejor técnica gráfica para mostrar cómo están distribuidos los datos de una población. Se pierde muy poca información (lo que está dentro de cada intervalo de clase).

Volviendo a nuestro ejemplo

Centro de atención	Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora en ser atendido
H. Fernández	Juan Perez	Arenales 843	07/09/18	Médica	07:32	08:04
H. Fernández	Romina Paz	Malabia 1820	07/09/18	Emergencia	09:35	09:46
H. Argerich	Pedro Ast	Zapiola 2232	07/09/18	Emergencia	07:32	08:34
H. Argerich	Graciela Fort	Castillo 156	07/09/18	Médica	09:45	10:26
H. Álvarez	Lía Sope	Bogotá 1566	07/09/18	Emergencia	05:56	06:02
H. Álvarez	Carlos Seguí	Caracas 578	07/09/18	Emergencia	10:35	11:12

Podemos definir una nueva variable, T.

T= Tiempo de espera
T=Hora en ser atendido-Hora de llegada

El tiempo de espera promedio es 33.2 minutos

Información un poco pobre…

Estudiemos la distribución.

Emergencias Hosp. Alvarez
Frecuencia
0
100
200
300
0
50
100
150
200
250
Tiempo de espera [m]

Consulta médica Hosp. Alvarez
Frecuencia
0
200
400
600
800
10
15
20
25
30
35
40
Tiempo de espera [m]

- En emergencias hay pacientes que los atienden muy rápido y otros que esperan mucho. distribucion parecida a la exponencial
- En cambio cuando llegan con turno más o menos se "respeta" el horario. distribucion parecida a la normal

¿Qué acabamos de hacer?

Estudiamos la *distribución* del tiempo de espera *condicional* a cada una de las categorías de consulta.

Estudiamos la dependencia entre la variable *tiempo de espera* y la variable *consulta*.

Pregunta al margen: conociendo esta información.

Si saben que un nuevo paciente esperó 2 horas en ser atendido, ¿entró por emergencias, por consulta médica o por otra consulta?

Y si un nuevo paciente esperó 5 minutos en ser atendido, ¿emergencias, consulta médica o por otra consulta?

Y si un nuevo paciente esperó 40 minutos en ser atendido, ¿emergencias, consulta médica o por otra consulta? ¿Con qué chance?

Acaban de resolver un problema de clasificación.

Alvarez: ¿cómo bajar el tiempo de espera?

¿Agregando más médicos alcanzará?

¿Disminuyendo la demanda?

Construyamos un nuevo hospital o sala de emergencias para bajar la demanda, ¿dónde?

¿Dónde lo construimos?, ¿Qué información tenemos?

Hospital Álvarez

Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora en ser atendido
Lía Sope	Bogotá 1566	07/09/18	Emergencia	05:56	06:02
Carlos Seguí	Caracas 578	07/09/18	Emergencia	10:35	11:12

Ballester
Av. Gral. Paz
Av. Leopoldo Lugones
PALERMO
RECOLETA
Buenos Aires
PCIA DE BUENOS AIRES
BUENOS AIRES
Ciudadela
Au 25 de Mayo
Avellaneda
BUENOS AIRES
PCIA DE BUENOS AIRES
Justo

Nuevo Hospital
Ballester
Av. Gral. Paz
Av. Leopoldo Lugones
PALERMO
RECOLETA
Buenos Aires
PCIA DE BUENOS AIRES
BUENOS AIRES
Ciudadela
Au 25 de Mayo
Avellaneda
BUENOS AIRES
PCIA DE BUENOS AIRES
Justo

Acaban de “resolver” un problema de clasificación no supervisada.

Encontraron el centro de un cluster.

Cambiemos el problema para ver si aprendimos que es esto de la distribución de los datos.

Supongamos que existe una única ambulancia con el equipamiento adecuado para un tipo de accidente (por ejemplo, quemaduras graves).

- **Llegar rápido es fundamental para que el paciente tenga un mejor pronóstico.**
- **Los datos históricos de los accidentes están marcados en el mapa**

¿Dónde ponemos la ambulancia?

¿Cómo serían los histogramas de tiempo en llegar?

RESUMEN

Resumen:

Variables categóricas

Diagrama de sectores ("torta")

Diagrama de barras

Consulta	Frecuencia
Consulta médica	35202
Emergencias	5627
Otros	15238

Variables numéricas

El histograma es la mejor técnica gráfica para mostrar cómo están distribuidos los datos. Se pierde muy poca información (lo que está dentro de cada intervalo de clase).

Entender la forma de esta distribución es fundamental.

Variables

Categóricas

- Nominal — Ej: Partido político
- Ordinal — Ej: Nivel educativo

	Hogares	%
Hogares que no sufrieron hambre	6.535.432	82.5
Hogares que sufrieron hambre	1.381.945	17.5
Total de hogares	7.917.377	100

Numéricas

- Discretas — Ej: número de hijos
- Continuas — Ej: Temperatura

¿Cómo resumir la información en una "tabla"?

Medidas de Resumen

Medidas de Resumen: { centralidad o posición / dispersión o variabilidad

centralidad

dispersión

Medidas de posición: $\begin{cases} \text{promedio (media) muestral} \\ \text{mediana muestral} \\ \text{media } \alpha \text{ podada} \end{cases}$

Promedio muestral $\equiv \overline{x} = \frac{x_1+x_2+\cdots+x_n}{n}$

Ej.: Datos $1, 3, 4, 5, 3, 2 \rightarrow \overline{x} = \frac{1+3+4+5+3+2}{6} = \frac{18}{6} = 3.$

Ej.: Datos 1,1,1,1,1,1,3,3,3,4,4,5,5,5,5 $\rightarrow$

$\overline{x} = \frac{1\times6+3\times3+4\times2+5\times4}{6+3+2+4} = \frac{43}{15} \approx 2.867$

x	Frecuencia (F)	Frecuencia Relativa (FR)
1	6	$\frac{6}{15}$
3	3	$\frac{3}{15}$
4	2	$\frac{2}{15}$
5	4	$\frac{4}{15}$

Mediana muestral $\equiv \widetilde{x}$

$\widetilde{x}$ es el dato que se encuentra en el medio una vez ordenados.

Datos $x_1, x_2, x_3, \ldots, x_n$.

Definimos $x_{(1)}$ =dato más chico, $x_{(2)}$ =segundo dato más chico,...,$x_{(n)}$ =dato más grande.

$\widetilde{x}=\begin{cases} x_{((n+1)0.5)} & \text{si n es impar} \\ \frac{1}{2}(x_{(n0.5)} + x_{(n0.5+1)}) & \text{si n es par} \end{cases}$ —>"Tomo el promedio de los 2 que estan en la mitad"

Ej.: Datos $x_1 = 1, x_2 = 8, x_3 = 6, x_4 = 5, x_5 = 2$

$\rightarrow x_{(1)} = 1, x_{(2)} = 2, x_{(3)} = 5, x_{(4)} = 6, x_{(5)} = 8 \rightarrow \widetilde{x} = 5$

Mediana muestral $\equiv \widetilde{x}$

Ej.: Datos 1,1,1,1,1,1,3,3,4,4,4,5,5,5,5,5

x	Frecuencia (F)
1	6
3	2
4	3
5	5

Buscamos: $x_{(8)} = 3$ y $x_{(9)} = 4 \rightarrow \widetilde{x} = \frac{x_{(8)}+x_{(9)}}{2} = 3.5$

- $\widetilde{x}$ es una medida robusta (insensible a puntos atípicos).

Ejemplo: Una pequeña empresa donde los sueldos (en pesos) de sus empleados son los siguientes:

62865 62820 62835 62860 62810 62800 329000

La media es

(62865+62820+62835+62860+62810+62800+329.000)/7= $100855,1

La mediana es 22835

62800 62810 62820 62835 62660 62865 329000

Medidas de posición: $\overline{x}_\alpha$

Media donde quitamos datos de las puntas o colas, con el fin de eliminar los outliers

Media α podada $\equiv \overline{x}_\alpha$

$\overline{x}_\alpha =$ es el promedio de los datos una vez que eliminamos el $\alpha 100\%$ de los datos más chicos y el $\alpha 100\%$ de los datos más grandes.

Ej.: Datos $x_{(1)} = 2, x_{(2)} = 6, x_{(3)} = 8, x_{(4)} = 10, x_{(5)} = 12, x_{(6)} = 14, x_{(7)} = 15, x_{(8)} = 21, x_{(9)} = 32, x_{(10)} = 53$

$\overline{x}_{0.1}$=es el promedio luego de haber eliminado el $x_{(1)}$ y el $x_{(10)}$.

$$\overline{x}_{0.1} = \frac{x_{(2)} + x_{(3)} + x_{(4)} + \cdots + x_{(7)} + x_{(8)} + x_{(9)}}{8}.$$

Definición

$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \cdots + x_{(n-[n\alpha])}}{n-2[n\alpha]}$ donde $[a]$ = parte entera de a.

Medidas de posición: $\overline{x}_\alpha$

Definición

$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \cdots + x_{(n-[n\alpha])}}{n-2[n\alpha]}$ donde $[a]$ = parte entera de a.

- Si $n\alpha$ es un entero $\overline{x}_\alpha$ la definición es correcta, en caso contrario se interpola.
- $\overline{x}_\alpha$ es una medida robusta (insensible a una cantidad $n\alpha$ de puntos atípicos en cada lado).

Propiedades de la media

- Sólo sirve para datos cuantitativos (escala intervalo)
- Es sensible a outliers

Propiedades de la mediana

- Sirve para escala de intervalo y también para datos ordinales.
- No es sensible a outliers

Comparación Media-Mediana

1. Para distribuciones simétricas, la media y la mediana coinciden
2. Para distribuciones asimétricas, la media se mueve hacia el lado de la cola.
3. Sensibilidad a outliers
4. Tipo de datos para los que sirven

Cuartiles y Percentiles

[Lo que esta entre corchetes, siempre totmamos la parte enterea]

primer cuartil muestral

$$q_1 = x_{([0.25(n+1)])}$$

tercer cuartil muestral

$$q_3 = x_{([0.75(n+1)])}$$

Cuartiles y Percentiles

El percentil p es un número que deja al p% de los datos por debajo de él y al (100-p)% por encima

EN LA DISTRIBUCION DEL INGRESO SE REPORTAN LOS PERCENTILES

Medidas de dispersión:
- rango muestral
- desvío estándar muestral
- distancia intercuartil
- MAD

- Todas las medidas de dispersión son ≥ 0.

dispersión

Medidas de dispersión:

Rango Muestral: RM

$RM =$ valor máximo-valor mínimo$= x_{(n)} - x_{(1)}$.

- Es una medida muy sensible a puntos atípicos.

Desvío Estándar Muestral: S

¿Qué hacemos con estas distancias (palitos verdes)?

Desvío Estándar Muestral: S

La varianza de n observaciones $x_1, x_2, \ldots, x_n$ es*

$$S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots (x_n - \bar{x})^2}{n}$$

El desvío estándar de n observaciones $x_1, x_2, \ldots, x_n$ es

$$S = \sqrt{S^2}$$

* En realidad hay que dividir por $n - 1$ en vez de n.

- S es la distancia "promedio" de un dato a la media de todos los datos.
- Es una medida sensible a puntos atípicos.

Regla empírica: Si el histograma de los datos tiene forma de campana (aproximadamente) entonces

- Aproximadamente el 68% de los datos están a distancia a lo sumo S de la media.
- Aproximadamente el 95% están a distancia a lo sumo 2S de la media.
- Prácticamente todos están a distancia a lo sumo 3S de la media.

Medidas de dispersión:

Distancia Intercuartil: IQR

IQR=tercer cuartil muestral-primer cuartil muestral= $q_3 - q_1$.
Donde $q_3 = x_{([0.75(n+1)])}$ y $q_1 = x_{([0.25(n+1)])}$.

Ej.: n=19, Datos ordenados:
$1, 1, 2, 3, \underline{4}, 5, 6, 7, 7, 9, 10, 12, 14, 14, \underline{17}, 19, 23, 27, 39.$

$q_3 = x_{([0.75(19+1)])} = x_{([0.75*20])} = x_{([15])} = x_{(15)} = 17.$
$q_1 = x_{([0.25(19+1)])} = x_{([0.25*20])} = x_{([5])} = x_{(5)} = 4.$
$IQR = q_3 - q_1 = 17 - 4 = 13.$

- El IQR nos dice en que rango se encuentra el 50% de los datos centrales.
- Notar que $q_2 = \widetilde{x}$.
- Es una medida poco sensible a puntos atípicos (como $\overline{x}_{0.25}$).

Mucho mas robusto ante outliers

Desvío absoluto mediano: MAD

$$MAD = mediana|x_i - \widetilde{x}|.$$

Pasos para calcular el MAD:
1) ordenamos los datos
2) calculamos $\widetilde{x}$
3) calculamos $|x_i - \widetilde{x}|$ (ojo!! tiene un módulo)
4) ordenamos $|x_i - \widetilde{x}|$
5) calculamos la mediana de 4).

- Es una medida robusta (como $\overline{x}_{0.49}$).

Ej: $x_1 = 1.2, x_2 = 2.3, x_3 = 0.8, x_4 = 3.1, x_5 = 1.3$

1) $x_{(1)} = 0.8, x_{(2)} = 1.2, \underline{x_{(3)} = 1.3}, x_{(4)} = 2.3, x_{(5)} = 3.1.$

2) $\widetilde{x} = 1.3$ mediana

3)

x	$\lvert x_i - 1.3 \rvert$
1.2	0.1
2.3	1
0.8	0.5
3.1	1.8
1.3	0

4) ordenamos la segunda columna 0,0.1,<u>0.5</u>,1,1.8

5) MAD=0.5

Algunos comentarios sobre la distribución Normal (forma de campana)

IQR y MAD estandarizados

Muestra aleatoria de una distribución $N(\mu, \sigma^2)$:

- Ahora tenemos una interpretación para S, $\frac{IQR}{1.35}$, $\frac{MAD}{0.675}$.
- Además, el IQR y el MAD tienen la ventaja de ser robustos.

Supongamos que la variable altura de los hombres tiene una distribución Normal.

=

Si tomamos una muestra aleatoria (o toda la población) y realizamos un histograma con todas la alturas medidas, éste tendrá forma de campana.

Población Inglesa

JOHN TIENE ALTURA B, ¿ SE PUEDE CONSIDERAR QUE ES ALTO?

Población de Pigmeos

TWA TIENE ALTURA C, ¿ SE PUEDE CONSIDERAR QUE ES ALTO?

¿Cómo se puede mostrar cuan típico o atípico es un valor?.

En la pregunta esta implícito que la escala no es relevante, simplemente nos interesa saber cuan diferente es ese dato del resto de sus "vecinos".

JOHN TIENE ALTURA B

TWA TIENE ALTURA C

Ambos son muy altos (en relación a su población).

Definir un numero que refleje esto:

$$Z_{john} = \frac{(B - \text{promedio})}{\text{desvío estándar}} \qquad Z_{twa} = \frac{(C - \text{promedio})}{\text{desvío estándar}}$$

$$Z_{john} \approx Z_{twa}$$

Estandarización

$$Z = \frac{(\text{valor} - \text{promedio})}{\text{desvío estándar}}$$

¿Por qué hay que restar el promedio?

¿Por qué se divide por el desvío estándar?

Boxplot

El boxplot es otra técnica gráfica para mostrar cómo están distribuidos los datos.

Se construye a partir de q_1, q_2, y q_3

Se pueden reconocer rápidamente los puntos atípicos

Boxplot

Nota, por mas que el largo del bigote llegue a 1.5 * IQR, llega hasta donde se encuentre el ultimo dato. Por eso, no siempre va a ser simétrico a ambos lados el largo del bigote.

Presentación de la información: boxplot

Ej.: n=19, Datos ordenados:
$1, 1, 2, 3, \underline{4}, 5, 6, 7, 7, 9, 10, 12, 14, 14, \underline{17}, 19, 23, 27, 39.$

$q_3 = x_{([0.75(19+1)])} = x_{([0.75*20])} = x_{([15])} = x_{(15)} = 17.$
$q_1 = x_{([0.25(19+1)])} = x_{([0.25*20])} = x_{([5])} = x_{(5)} = 4.$
$IQR = q_3 - q_1 = 17 - 4 = 13 \mapsto 1.5IQR = 19.5.$
El bigote superior va hasta el dato más grande que sea inferior o igual a 17+19.5=$\underline{36.5}$→ bigote superior en $\underline{27}$.
El bigote inferior va hasta el dato más chico que sea mayor o igual a 4-19.5=$\underline{-1.5}$→ bigote inferior en $\underline{1}$.

QQ plot y acumulada empírica

Función de distribución (acumulada) empírica

Si $x_1, x_2, \ldots, x_n$ es una muestra de datos, definimos su función de distribución acumulada empírica como la función $F_n : \mathbb{R} \to \mathbb{R}$ dada por

$$F_n(x) = \frac{\#\{x_i \leq x\}}{n}.$$

O sea, si ordenamos la muestra: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ entonces

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ 1/n & \text{si } x_{(1)} \leq x < x_{(2)} \\ 2/n & \text{si } x_{(2)} \leq x < x_{(3)} \\ \ldots & \\ 1 & \text{si } x \geq x_{(n)} \end{cases}$$

ACUMULADA EMPIRICA
"cantidad de datos menores a un valor"

Acumulada empirica

El 80% de los datos son mas chicos que 0.89

QQ-plot

Compara los quantiles de la muestra vs los quantiles teoricos de otras distribuciones conocidas, para ver si se parece a alguna distribucion conocida

El qq-plot es una técnica gráfica que busca determinar si la variable de estudio tiene una determinada distribución.

Se compara la acumulada de los datos (acumulada empírica) con la acumulada de la distribución en duda.

QQ-plot

Tenemos una muestra de datos $x_1, x_2, \ldots, x_n$ y queremos saber si podemos pensarlos como v. a. i. i. d. con alguna distribución continua, por ejemplo $N(0,1)$.

Al ordenar la muestra $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ tenemos que:

$$x_{(1)} = \text{ cuantil } \frac{1}{n+1} \text{ de la muestra}$$

$$x_{(2)} = \text{ cuantil } \frac{2}{n+1} \text{ de la muestra}$$

$$\cdots$$

$$x_{(n)} = \text{ cuantil } \frac{n}{n+1} \text{ de la muestra}$$

Entonces, para cada $i = 1, \ldots, n$, $x_{(i)}$ es el cuantil $\frac{i}{n+1}$ de la muestra. Queremos ver si *se parece* al cuantil $\frac{i}{n+1}$ de una $N(0,1)$.

Notamos $\Phi^{-1}\left(\frac{i}{n+1}\right)$ al cuantil $\frac{i}{n+1}$ de una $N(0,1)$ ya que es el valor x tal que $\Phi(x) = \frac{i}{n+1}$.
Para ver si

$$\Phi^{-1}\left(\frac{i}{n+1}\right) \approx x_{(i)} \qquad \forall i = 1, \ldots, n$$

realizamos un q-q plot, o sea un gráfico de los puntos

$$\left(\Phi^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right)_{i=1,\ldots,n}.$$

En el caso de que los datos, coincidan con la distribucion con la que lo comparo (en este caso la normal), al graficarlo, deberia darme algo "parecido" a una recta.
Nota = Siempre compara en el caso de la normal, los datos vs normal (0, 1)

datos: rnorm(100,20,4)
Sample Quantiles
Theoretical Quantiles
10
15
20
25
30
-2
-1
0
1
2

datos: rnorm(10000,20,4)
Sample Quantiles
Theoretical Quantiles
5
10
15
20
25
30
35
-4
-2
0
2
4

Sea $Y \sim N(\mu, \sigma^2)$ y notemos $X = \frac{Y-\mu}{\sigma} \sim N(0,1)$.

¿Qué relación cumplen los cuantiles de X e Y?

Si graficamos los puntos obtenemos $\left(F_X^{-1}\left(\frac{i}{n+1}\right), F_Y^{-1}\left(\frac{i}{n+1}\right)\right)$.

¿qué obtenemos?

$$y = \sigma x + \mu.$$

datos: rnorm(10000,20,4)
Sample Quantiles
Theoretical Quantiles
$y = \sigma x + \mu.$

rexp(100)
Frequency
x
Cuantiles muestrales
Cuantiles Teoricos Exp(1)

Resumen:

Datos categóricos:

+ una tabla

Datos numéricos:

+ todas las medidas de resumen (posición y variabilidad)

Y AHORA…

¿Qué podemos hacer si tenemos datos en dimensión 2 (2D)?

Medidas de Resumen: { centralidad o posición / dispersión o variabilidad

Persona	Ingreso	Gasto en alquiler
1	42521	14025
2	25200	8200
3	32126	9500

Medidas de centralidad

La distribución Normal en 2D

elipses

Buscamos el punto que esté en el “centro”

El punto que minimiza la suma de las distancias a los datos:

distancias al cuadrado $(\bar{x}, \bar{y}) = (\ \overline{ingreso},\ \overline{gasto\ en\ alquiler}\ )$

distancias en módulo

¿Cuál es la diferencia?

¿Se pueden tomar otras distancias?

Distancia edit

¿sensibilidad a ptos atípico?

Medidas de variabilidad

1D

2D

Promedio de distancias

Promedio de distancias al cuadrado y luego raiz

Puntos atípicos, distancias y profundidades

Círculos

$$x^2 + y^2 = 1 \quad , \quad (x - \mu_x)^2 + (y - \mu_y)^2 = 1$$

Elipses

$$(\frac{x}{2})^2 + (\frac{y}{4})^2 = 1 \quad , \quad (\frac{x - \mu_x}{\sigma_x})^2 + (\frac{y - \mu_y}{\sigma_y})^2 = 1$$

Elipses rotadas

$$(\frac{x}{2})^2 + cxy + (\frac{y}{4})^2 = 1$$

$$(\frac{x - \mu_x}{\sigma_x})^2 - 2\rho\frac{(x - \mu_x)(y - \mu_y)}{\sigma_x\sigma_y} + (\frac{y - \mu_y}{\sigma_y})^2 = 1 - \rho^2$$

Normal de dimension 2

$$f(x,y) = a.e^{-\frac{1}{2}((x-\mu_x)^2+(y-\mu_y)^2)}$$

$$f(x,y) = a.e^{-\frac{1}{2}((\frac{x-\mu_x}{\sigma_x})^2+(\frac{y-\mu_y}{\sigma_y})^2)}$$

$$f(x,y) = a.e^{-\frac{1}{2(1-\rho^2)}((\frac{x-\mu_x}{\sigma_x})^2-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+(\frac{y-\mu_y}{\sigma_y})^2)}$$

$$¿\mu_x? \quad ¿\mu_y? \quad ¿\sigma_x = \sigma_y?$$

Teórica

$$\frac{1}{2(1-\rho^2)}\left(\left(\frac{-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)$$

A partir de datos estimamos los valores teóricos

$$\frac{1}{2(1-r^2)}\left(\left(\frac{x-\bar{y}}{S_x}\right)^2 - 2r\frac{(x-\bar{x})(y-\bar{y})}{S_xS_y} + \left(\frac{y-\bar{y}}{S_y}\right)^2\right)$$

Tomamos un punto $\vec{z} = (z_x, z_y)$

$$D(\vec{z}) \equiv \frac{1}{2(1-r^2)}\left(\left(\frac{z_x-\bar{y}}{S_x}\right)^2 - 2r\frac{(z_x-\bar{x})(z_y-\bar{y})}{S_xS_y} + \left(\frac{z_y-\bar{y}}{S_y}\right)^2\right)$$

$D(\vec{z})$ representa cuan cercano a la media se encuentra el punto $\vec{z}$.
$D(\vec{z})$ es la distancia *pesada* entre $\vec{z}$ y la media.

Profundidad:
Dada una muestra aleatoria $x^1, .., x^n$, in R^n, para cada z en el espacio podemos decir cuan lejos se encuentra, o cuan poco profundo es el punto o dato z.

El dato más profundo es que el se encuentra más en el centro, a medida que uno se aleja del centro la profundidad decrece.

Profundidad de Mahalanobis:

$D(\vec{z})$ representa cuan cercano a la media se encuentra el punto $\vec{z}$.
$D(\vec{z})$ es la distancia *pesada* entre $\vec{z}$ y la media.

$$Profundidad_M(\vec{z}) = \frac{1}{1 + D(\vec{z})}$$

Lo interesante de las profundidades es que nos dan una medida para determinar si un punto es atípico (outlier).

Un punto muy alejado (D(z) grande) o muy poco profundo (Profundidad(z) chica) es un candidato a outlier.

$$D(\vec{z}_1) > D(\vec{z}_2)$$

$$Profundidad_M(\vec{z}_1) < Pofundidad_M(\vec{z}_2)$$

Si la distribución de los datos no es Normal pero tiene un punto de simetría.

En esto caso también podemos medir la profundidad de cada punto, pero usando otra medida de profundidad (ya que aquí no hay elipses)

A partir de los *n* datos podemos construir muchos triángulos eligiendo 3 datos.

A partir de los *n* datos podemos construir muchos triángulos eligiendo 3 datos.

¿Y AHORA? ¿ COMO DEFINIMOS SI LA PROFUNDIDAD?
¿QUÉ CUMPLEN LOS DATOS MAS PROFUNDOS?

$$Profundidad_L(\vec{z}_1) \equiv \text{Número de triángulos que contienen a } \vec{z}$$

¿Y si los datos están en dimensión 3 o más?

Ojo con los datos atípicos (outliers)!!

En dimensiones bajas es fácil descubrirlos

En dimensiones altas es difícil descubrirlos

Solución: Métodos robustos

¿Por ejemplo?

Por eso buscamos desarrollar metodos que sean mas robustos en contra de estos puntos atipicos.

FIN