Análisis de Datos para la Toma de Decisiones

Clase 2

Christian Araya

Instituto de Estadística
PUCV

Marzo 2020

Recordatorio

Tablas de Frecuencias

No agrupadas:

- Clases: respuestas observadas en la población o muestra C_i.
- Frecuencia absoluta: número de elementos que pertenecen a la clase C_i. La denotamos: n_i.
- Frecuencia relativa: número de elementos que pertenecen a la clase C_i, relativo al total de elementos (de la población o muestra). La denotamos: f_i. Recordar que usamos el número decimal, no porcentaje.
- Frecuencia absoluta acumulada: la denotamos como N_i. Sumamos las frecuencias absolutas hasta la clase correspondiente.
- Frecuencia relativa acumulada: la denotamos como F_i. Ídem, pero con la frecuencia relativa.
- Acumulamos cuando tiene sentido hacerlo, es decir, si como mínimo estamos trabajando con una variable en una escala ordinal.

Agrupadas para datos continuos o discretos*

- Cuando existen muchos valores posibles para datos discretos* o un rango para datos continuos, debemos generar intervalos para calcular frecuencias. Estos intervalos, si se hacen de manera arbitraria, podrían presentar inconvenientes al momento de realizar alguna interpretación.
- Definiremos entonces:

1. LI_i: límite inferior del i-ésimo intervalo (clase).
2. LS_i: límite superior del i-ésimo intervalo (clase).
3. a_i: amplitud del i-ésimo intervalo (clase). Todas quedan iguales a una constante a.
4. m_i: marca de clase. Se obtiene como: $m_i = (LI_i + LS_i)/2$

Tablas de Frecuencias

Reglas:

- ¿Cómo calculamos la cantidad de clases necesarias?. Existe una regla como buena aproximación inicial: Sturges.

$$k = 3{,}3 \cdot log_{10}(n) + 1$$

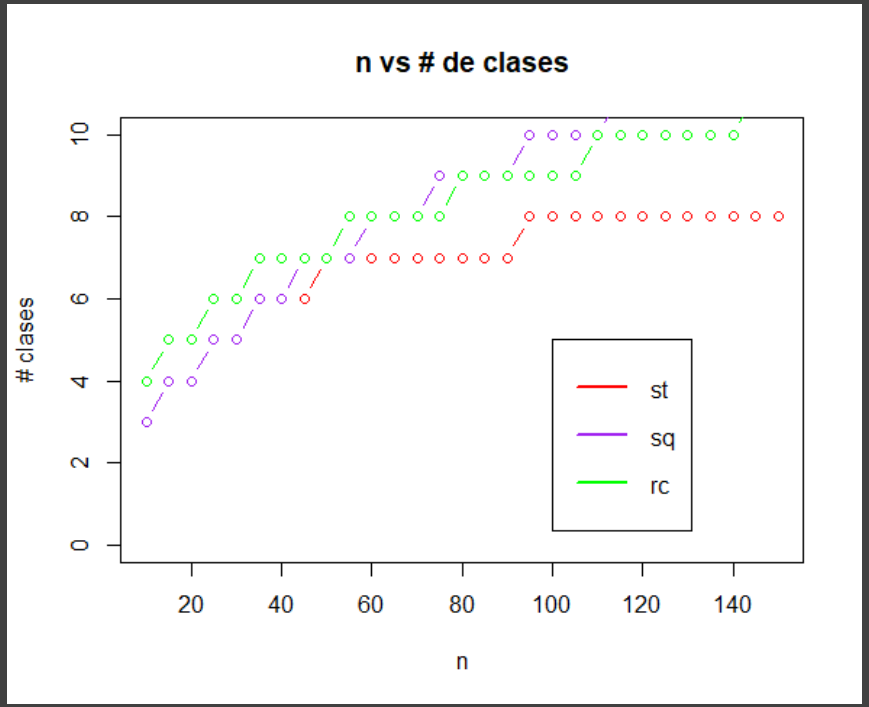
- En general se recomienda utilizar un número k impar, **pero la utilizaremos aproximando al entero más cercano.**
- Importante: el estudio/problema/ejercicio puede definir a priori el número de clases a considerar.
- Rango de la muestra: $R = (maxvalor) - (minvalor)$
- Amplitud del intervalo: $a_i = (R + u)/k$
- **¿Qué es u?** es la unidad de trabajo (entero => 1; un decimal => 0.1; dos decimales => 0.01).

Tablas de Frecuencias

Reglas:

- Determinamos el rango de la tabla: $R_T = a \cdot k$
- Distancia entre rangos: $\Delta = R_T - R$
- Finalmente, iniciamos la tabla: $LI_1 = minimovalor - \Delta/2$
- Y: $LS_1 = LI_1 + a$

1. **Nota:** algunos autores sugieren redondear la amplitud a un valor entero para no trabajar con decimales en exceso.
2. Recordar que la elección de un número de intervalos impares tiene relación con la posibilidad de observar atisbos de distribuciones presentes en los datos.
3. Existen varias reglas en Estadística para calcular el número de intervalos. La más popular es la de Sturges, pero todas entregan resultados similares.
   Ejemplo: Rice ($2n^{1/3}$) y se aproxima; Raíz cuadrada $\sqrt{n}$, ídem.

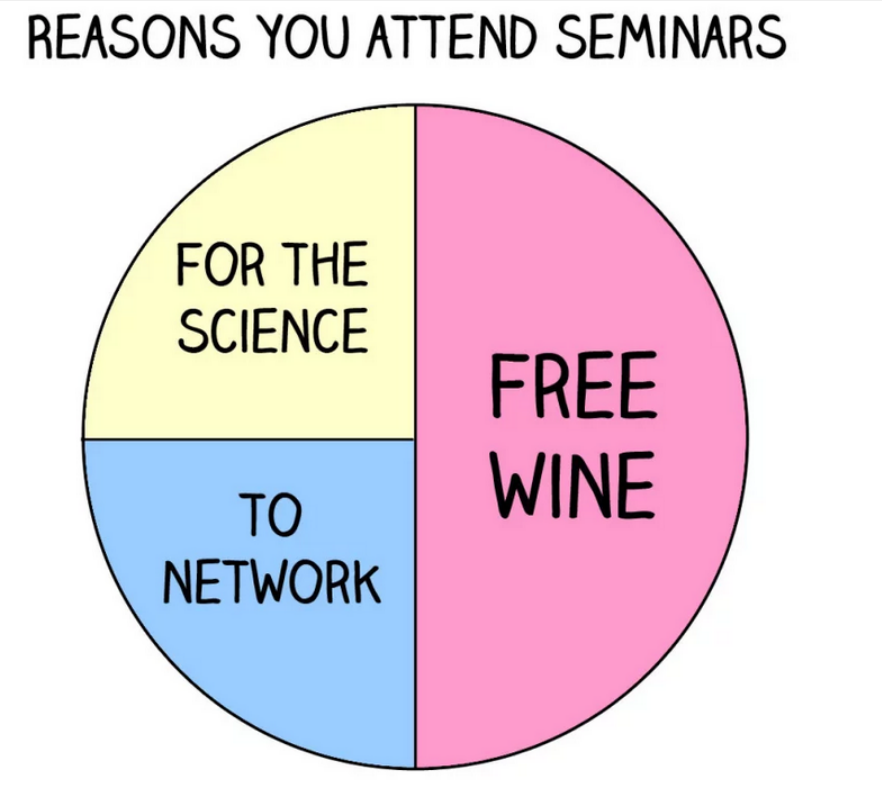


Gráficos en Estadística Descriptiva Univariada

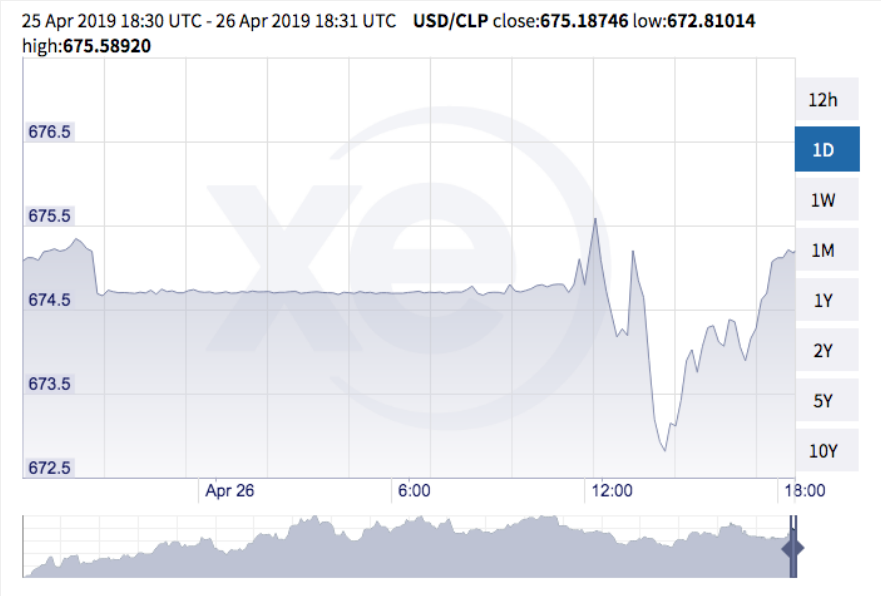
Introducción

- A través de un gráfico, representamos y resumimos de una manera distinta los datos disponibles.
- Ayudan a visualizar características de los datos que no son visibles fácilmente en una tabla de frecuencias.
- Se ha masificado el uso y confección de gráficos (ejemplo: publicaciones científicas, periódicos, matinales, Buzzfeed, etc.) por el desarrollo de herramientas para graficar. La idea es comprender qué tipo de gráfico es el apropiado para el estudio que se está aplicando y las variables involucradas en el análisis.

Gráficos



Fuente: Buzzfeed.



Fuente: Xe.com.

Barras

- Se utiliza para variables cualitativas (nominales u ordinales).
- Las barras van separadas para indicar que es un caso cualitativo.
- Si se emplea para graficar variables cuantitativas discretas (con pocos niveles), cuidar que las barras estén separadas. No se debe confundir con un histograma.
- En escala nominal, el orden de las categorías es arbitrario (recordar que no existe jerarquía). En el caso ordinal, se mantiene el orden en la escala.
- Se muestran las frecuencias absolutas (cantidad real), pero también se utilizan para mostrar los porcentajes.

Circular

- Otra forma de representación para datos cualitativos (o variables discretas, con pocos niveles).
- Se recomienda no usar versión 3D porque distorsiona la percepción de los segmentos posteriores.
- Cada porción es proporcional al % de cada clase, de modo que se puede comprender la relación de cada clase con respecto al total.

Histograma

- En data continua, es uno de los diagramas más conocidos de Estadística. Horizontal: clases; vertical: frecuencias. Las barras se dibujan de manera adyacente.
- Hay versiones que utilizan los límites de cada clase y otras, la marca de clase al centro del intervalo.
- Permite visualizar la distribución de frecuencia (se percibe la forma que describen los datos).
- Hay versiones que utilizan el % como la altura de las barras o la frecuencia relativa.

Polígonos de Frecuencia

- Conecta puntos a través de segmentos lineales. Los puntos se corresponden con la marca de clase.
- Parte y termina en o.
- Asocian la frecuencia con la marca de clase. En algunos casos, complementan visualmente al histograma.
- También se utiliza para comparar grupos cuando existen al interior de la muestra.
- Además existen versiones con las frecuencias relativas.

Ojiva

- Este tipo de gráfico de líneas se utiliza para representar las frecuencias acumuladas (absolutas o relativas).
- Sirve para determinar la cantidad de valores que se encuentran bajo un determinado nivel.

Otras formas de resumir un set de datos

Introducción

- Hemos resumido la información con ayuda de tablas y gráficos. Ahora buscamos resúmenes expresados en un número con reglas de cálculo establecidas.
- La idea es que dicho número represente todos los datos disponibles.

Medidas de Tendencia Central

- Estos números definen una idea del ”centro” de los datos. Ejemplos típicos son: **media, mediana y moda**.
- Si las tres medidas proporcionan valores similares, concluimos que tenemos una distribución simétrica.

Media

- Promedio o media aritmética.
- Para la población, se denota como μ, mientras que para la muestra como $\overline{X}$.
- A partir de los datos a granel, se calcula como:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Para la población, reemplazamos n por N, es decir por el total de elementos en ella.
- Como desventaja, se ve afectada por la presencia de valores extremos.
- Recordar que cuando calculamos el promedio, no damos cuenta de la dispersión de los datos.

Buenos Aires, Argentina

Thursday 7:00 PM

Light Drizzle

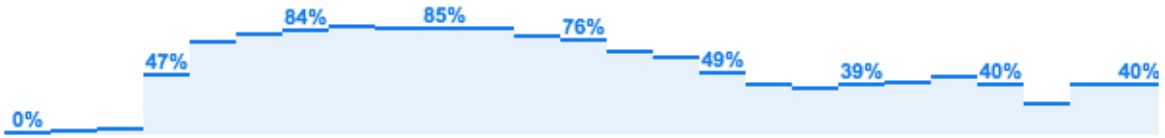
18 °C | °F

Precipitation: 0%

Humidity: 83%

Wind: 31 km/h

Temperature | Precipitation | Wind



8 PM | 11 PM | 2 AM | 5 AM | 8 AM | 11 AM | 2 PM | 5 PM

Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu
19° 17°	19° 17°	20° 14°	22° 15°	23° 13°	18° 12°	21° 13°	23° 15°

Media

- Para datos tabulados, utilizamos la marca de clase:

$$\overline{X} = \frac{\sum_{i=1}^{k} m_i * n_i}{n}$$

- Se tiene una extensión similar para el caso de la población (en las k clases generadas, hemos repartido los N elementos de la población en ese caso).

Masa (kg)	Li	Ls	mi	ni	fi	Ni	Fi
45 – 49	45	49	47	2	0,05	2	0,05
50 – 54	50	54	52	4	0,10	6	0,15
55 – 59	55	59	57	7	0,18	13	0,33
60 – 64	60	64	62	10	0,25	23	0,58
65 – 69	65	69	67	4	0,10	27	0,68
70 – 74	70	74	72	6	0,15	33	0,83
75 – 79	75	79	77	7	0,18	40	1,00

Mediana

- Es el valor que se encuentra al centro de los datos, **cuando estos se ordenan de menor a mayor.**
- Datos a granel, n impar:

$$M_e = X_{(n+1)/2}$$

- Datos a granel, n par:

$$M_e = (1/2)(X_{n/2} + X_{n/2+1})$$

Mediana

- La mediana es más estable que el promedio ante la presencia de datos extremos. Sin embargo, presenta la desventaja que el valor obtenido puede no encontrarse en el set de *datos reales*.

Mediana

- Para el caso de datos tabulados, la mediana se calcula del siguiente modo:

$$M_e = L_i + (\frac{n/2 - N_{i-1}}{n_i})a_i$$

- L_i: límite inferior de la clase que contiene la mediana.
- n: número total de observaciones en la distribución de frecuencias.
- a_i: amplitud de clase.
- N_{i-1}: la frecuencia acumulada anterior a la clase que contiene la mediana.
- n_i: número de observaciones en la clase que contiene la mediana.
- La clase que contiene a la mediana es la primera clase cuya **frecuencia acumulada** es mayor o igual a la mitad de los datos.

Masa (kg)	Li	Ls	mi	ni	fi	Ni	Fi
45 – 49	45	49	47	2	0,05	2	0,05
50 – 54	50	54	52	4	0,10	6	0,15
55 – 59	55	59	57	7	0,18	13	0,33
60 – 64	60	64	62	10	0,25	23	0,58
65 – 69	65	69	67	4	0,10	27	0,68
70 – 74	70	74	72	6	0,15	33	0,83
75 – 79	75	79	77	7	0,18	40	1,00

Moda

- Es el valor o clase que ocurre con mayor frecuencia o el que más veces se repite.
- En un contexto con variables cualitativas o discretas con pocos niveles, es directo al mirar la tabla.
- Cuando hemos armado una tabla de frecuencia con intervalos, existe una fórmula:

$$M_o = LI_o + (\frac{\Delta_1}{\Delta_1 + \Delta_2})a$$

- $\Delta_1 = n_i - n_{i-1}$
- $\Delta_2 = n_i - n_{i+1}$
- Un conjunto de datos es bimodal, si dos clases tienen la misma frecuencia y es la más alta. Si ocurre más de dos veces, es multimodal.
- Notar que la Moda es la única medida de tendencia central que aplica para variables cuantitativas y cualitativas.

Masa (kg)	Li	Ls	mi	ni	fi	Ni	Fi
45 – 49	45	49	47	2	0,05	2	0,05
50 – 54	50	54	52	4	0,10	6	0,15
55 – 59	55	59	57	7	0,18	13	0,33
60 – 64	60	64	62	10	0,25	23	0,58
65 – 69	65	69	67	4	0,10	27	0,68
70 – 74	70	74	72	6	0,15	33	0,83
75 – 79	75	79	77	7	0,18	40	1,00

Todas las medidas de Tendencia Central son importantes, ninguna reemplaza a la otra. La idea es utilizarlas en conjunto para potenciar la información que se extrae desde los datos.

Dispersión

- Son números que proporcionan información sobre la dispersión de los datos y complementan a las medidas de tendencia central.
- Podemos comprender qué tan homogéneos (o heterogéneos) son los datos.

Varianza

- Mide la desviación de los datos con respecto al promedio, pero considera los términos de la diferencia al cuadrado (ver la fórmula).

$$V(X) = \sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N}$$

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

$$V(X) = \sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N}$$

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

Varianza

- Su valor es siempre positivo.
- Tiene dimensiones de [unidad de medida al cuadrado]. Por ello, tiene la desventaja de no tener interpretación directa.

Varianza

- Hay dos relaciones de interés: la primera, es una equivalencia que se desprende de la fórmula de S^2:

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} = (\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)/(n-1)$$

- La segunda, la forma de obtener la medida desde data tabulada:

$$S^2 = (\sum_{i=1}^{k} m_i^2 n_i - n\bar{X}^2)/(n-1)$$

Desviación Estándar

- Corresponde a la raíz cuadrada positiva de la varianza.
- Tiene sentido su interpretación debido a que posee las mismas unidades de medida que la variable en análisis.

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

Coeficiente de Variación

- Permite comparar variabilidad entre conjuntos de datos, sin importar la unidad de medida.
- Elevado CV = datos heterogéneos; bajo CV = datos homogéneos.

$$CV = 100\,\% \cdot \frac{S}{\bar{X}}$$

Importante

- Podemos comparar distintas varianzas o desviaciones estándar, pero es apropiado hacerlo cuando los promedios de la muestra o población son aproximadamente los mismos. Recordar que estamos dando cuenta de la variabilidad de los datos con respecto a la media.
- Una forma de comparar conjuntos de datos con medias distintas o con unidades de medida diferentes (ejemplo: peso en Kg versus estatura en m), es el coeficiente de variación.

Otras medidas de posición

Cuantiles

- Los cuantiles son medidas de posición. Dividen los datos en grupos, bajo los cuales se ubica una determinada proporción del conjunto de datos.
- Los datos deben ser de tipo cuantitativo (discreto o continuo).
- Por ejemplo, la **mediana** es un cuantil que divide la distribución de los datos en dos partes de igual frecuencia acumulada (50 %).
- Los **cuartiles** dividen el set en 4 cuartos.
- Los **quintiles** dividen en 5 partes.
- Los **deciles** en 10.
- Los puntos **percentiles** dividen al conjunto de datos en 100 partes.

Otras medidas de posición

Cuantiles

- Cuartil i, con i entre 1 y 3:

$$Q_i = X_{\frac{i(n+1)}{4}}$$

Quintiles

- Quintil i, con i entre 1 y 4:

$$K_i = X_{\frac{i(n+1)}{5}}$$

Otras medidas de posición

Deciles

- Decil *i*, con i entre 1 y 9:

$$D_i = X_{\frac{i(n+1)}{10}}$$

Percentiles (general)

- Todas pueden ser obtenidas desde la fórmula del percentil.
- Percentil *i*, con i entre 1 y 99:

$$P_i = X_{\frac{i(n+1)}{100}}$$

Percentiles (general)

- Para datos tabulados se utiliza:

$$P_j = LI_i + \{\frac{(n \cdot j/100) - N_{i-1}}{n_i}\}a$$

Masa (kg)	Li	Ls	mi	ni	fi	Ni	Fi
45 – 49	45	49	47	2	0,05	2	0,05
50 – 54	50	54	52	4	0,10	6	0,15
55 – 59	55	59	57	7	0,18	13	0,33
60 – 64	60	64	62	10	0,25	23	0,58
65 – 69	65	69	67	4	0,10	27	0,68
70 – 74	70	74	72	6	0,15	33	0,83
75 – 79	75	79	77	7	0,18	40	1,00