Estadística Descriptiva[abc]

Ernesto Barrios

Departamento de Estadística

ITAM

[a] Material basado fundamentalmente en V. Aguirre y B. Artolia (2007).
[b] Gráficas y cálculos realizados con el lenguaje estadístico *R*.
[c] Notas elaboradas con el lenguaje tipográfico LaTeX.

Contenido

1. Tipos de variables y escalas de medición.
2. Análisis exploratorio univariado para datos cualitativos y cuantitativos.
3. Medidas descriptivas.
4. Problema de comparación.
5. Problema de asociación.

Tipo de Datos o Variables

- Variables
  - *Cualitativas*
    - Nominales. *Color de auto: rojo, negro, . . .*
    - Ordinales. *Calidad de un servicio: bueno, regular, malo*
  - *Cuantitativas*
    - Discretas. *Número de empleados por departamento: 1,2, . . .*
    - Continuas. *Ingreso mensual: $5000, $10000, . . .*

Tipo de Variables

Variables Cualitativas

Variables Cualitativas: Cuando la información tomada denotan cualidades o atributos.

Pueden clasificarse en un número fijo de clases o categorías *exhaustivas y excluyentes*. Así, los datos quedan clasificados en una y solo una categoría.

E. g., considere los empleados de la empresa ABC:

Variable	Categorías
Departamento	producción, ventas, contabilidad, ...
Turno	matutino, vespertino, nocturno
Escolaridad	primaria, secundaria, ...
Género	masculino, femenino

Tipo de Variables

Variables Cuantitativas

Variables Cuantitativas: Variables o respuestas con significado numérico obtenidas por conteo o medición.

Si las variables se obtienen por conteo, las variables se dicen *discretas*. Si se obtienen por medición, *continuas*. E. g., considere nuevamente a los empleados de la empresa ABC:

Variable	Valores posibles	Tipo
Antigüedad (años)	1,2, . . .	discreta
Sueldo mensual ($)	1000–50000	continua
Vacaciones (días)	6, 7, . . .	discreta
Peso (kg)	50–120	continua

Escalas de Medición

Dependiendo de la precisión de los datos será el tipo de análisis.

1. *Escala Nominal:* La más básica clasificación de los valores en categorías (exhaustivas excluyentes). No hay relación de orden. Operaciones aritméticas no tienen sentido. E. g., estado civil; zona donde vive; color de auto.
2. *Escala Ordinal:* Del tipo nominal pero las categorías pueden ordenarse de acuerdo al grado de posesión de cierto atributo (*mayor que, menor que*). E. g., nivel escolar: primaria, secundaria, etc.; nivel socioeconómico: bajo, medio y alto; calidad de servicio: bueno, regular y malo. Operaciones aritméticas sin sentido.

Escalas de Medición

3. *Escala de Intervalo:* Además del grado de posesión de cierto atributo es posible medir la intensidad de la posesión. Se acepta (arbitrariamente) una medida como *cero* u origen. Las operaciones de suma y resta son válidas. E. g., las escalas Celsius y Fahrenheit de temperatura.
4. *Escala de razón:* El *cero* indica "ausencia" del atributo. Todas las operaciones aritméticas son válidas. El cociente nos permite la comparación por proporciones (razones). E. g., costo mensual en publicidad; ingreso anual familiar, etc.

Escalas de Medición

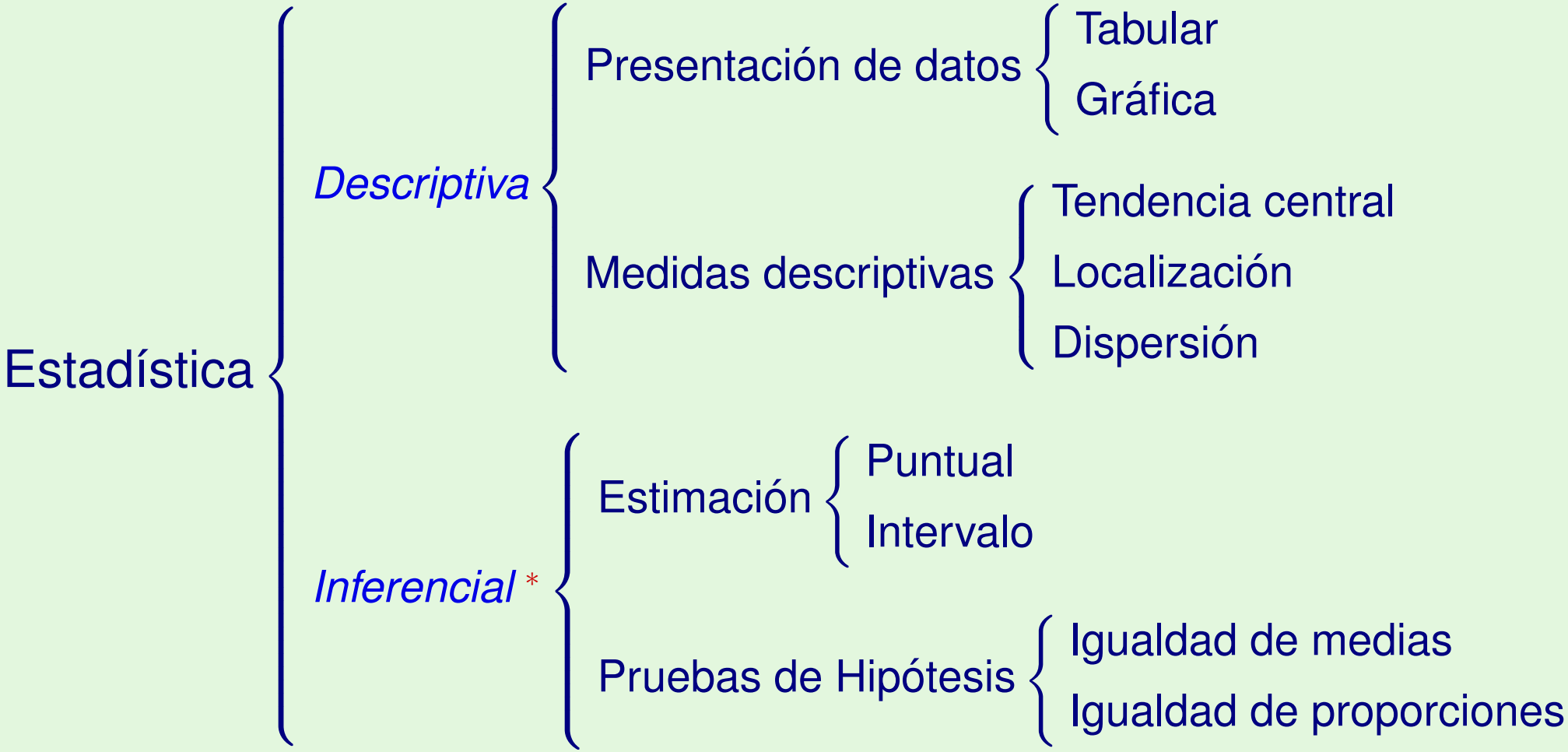
Razón $\subset$ Intervalo $\subset$ Ordinal $\subset$ Nominal

Tipos de Variable y Escalas de Medición

Ejemplos:

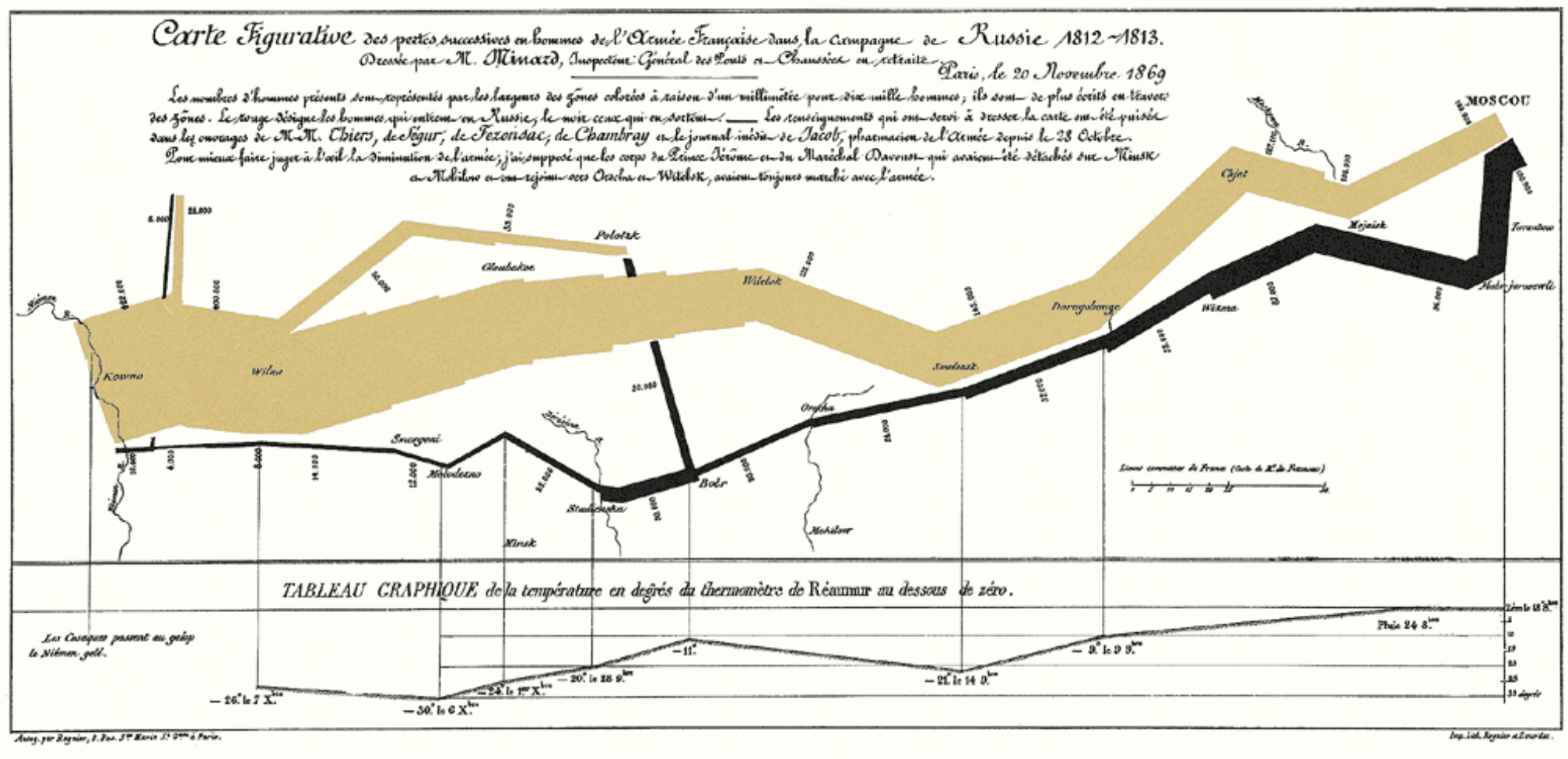
Variable Respuesta	Rango de Valores	Tipo de Variable	Escala de Medición
Estatura (cm)	$(150,\ 220)$	cuant/cont	razón
Afiliación política:	PRI, PAN, ...	cualitativo	nominal
Días en calendario: (gregoriano, maya)	$\{1, 2, 3, \dots\}$	cuant/disc	intervalo
Tipo de automóvil:	deportivo, de lujo, ...	cualitativa	nominal
Clasificación de película:	niños, adultos, ...	cualitativa	ordinal
Nivel de tonos bajos (dB)	$(-7,\ +7)$	cuant/cont	intervalo
Consumo eléctrico (kw/hr.)	$[0, \infty)$	cuant/cont	razón
Habilidad en Karate:	cinta amarilla, verde, ...	cualitativa	ordinal
Venta mensual	$(-\infty, \infty)$	cuant/cont	razón
Salida del sol (hrs.)	$[0, 24]$	cuant/cont	intervalo
Empleados enfermos	$\{0, 1, 2, \dots\}$	cuant/disc	razón

Clasificación Básica de la Estadística



* Uso extensivo de la *Probabilidad*.

Campaña napoleónica en Rusia



Charles Joseph Minard (1781-1870). Tomado de E. W. Tufte (1983) *The visual display of qualitative information*. Graphics Press.

Ejemplo: Encuesta de televisión por cable[a]

Una empresa de televisión por cable encargó a un bufete un estudio de mercado para conocer el perfil de los clientes potenciales de una zona residencial formada por dos colonias. Las colonias constan de 12 y 25 manzanas con un total de 236 y 605 hogares, respectivamente. Mediante muestreo probabilístico (no discutido aquí) se seleccionó una muestra de ocho manzanas y cinco hogares por manzana. En cada hogar seleccionado se recabaron varias respuestas de las que presentamos solamente algunas de éstas.

	Variable	Descripción
1	Colonia	Colonia a la que pertenece el hogar de la zona residencial
2	Manzana	Número de manzana a la que pertenece el hogar
3	Adultos	Número de adultos por hogar
4	Niños	Número de niños menores de 12 años por hogar
5	Teles	Número de televisores por hogar
6	Tipo	Tipo de televisor que posee: ByN, color, ambos
7	TVtot	Suma del número de horas frente al televisor en la semana de todos los miembros de la familia
8	Renta	Cantidad máxima de renta que el jefe del hogar estaría dispuesto a pagar al mes por servicio de TV por cable (múltiplos de $5)
9	Valor	Valor catastral del hogar (m$). La respuesta se usa para dar idea aproximada del ingreso familiar

[a] Aguirre et al. (2007)

Ejemplo: Encuesta de televisión por cable

Datos de la encuesta de televisión por cable

obs	colonia	manzana	adultos	ninos	teles	renta	tvtot	tipo	valor
1	2	20	3	2	2	50	68	B	79928
2	2	25	3	3	1	65	82	B	94415
3	2	20	1	2	1	45	40	A	120896
4	2	8	2	2	2	35	56	A	132867
5	2	25	1	2	0	0	0	N	141901
.	.	.	.	.	.	.	.	.	.
36	1	2	2	0	2	60	20	A	332699
37	1	2	3	0	3	70	28	C	336290
38	1	9	3	0	5	85	28	C	355641
39	1	9	2	0	3	70	20	C	357972
40	1	4	3	0	4	80	28	C	370325

Variables Cualitativas

Descripción Tabular

Tabla de Frecuencias para la variable *tipo* (tipo de televisión)

Una *tabla de frecuencias* nos muestra la frecuencia (absoluta o relativa) observada de cada una de las categorías de la variable.

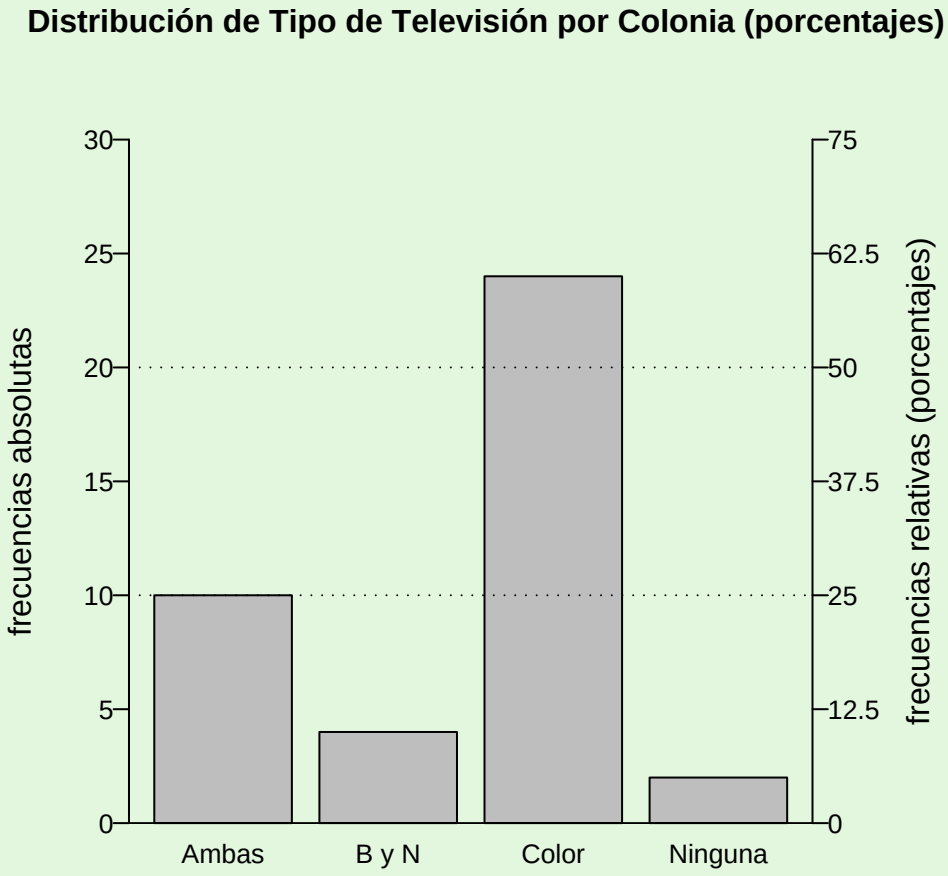
	total			Colonia 1			Colonia 2		
tipo	f_i	p_i	%	f_i	p_i	%	f_i	p_i	%
Ambos	10	0.25	25.0	2	0.133	13.3	8	0.320	32.0
Blanco y Negro	4	0.10	10.0	1	0.067	6.7	3	0.120	12.0
Color	24	0.60	60.0	12	0.800	80.0	12	0.480	48.0
Ninguno	2	0.05	5.0	0	0.000	0.0	2	0.080	8.0
Total (N)	40	1.00	100.0	15	1.000	100.0	25	1.000	100.0

Donde f_i son las frecuencias absolutas, $p_i = f_i/N$ las frecuencias relativas y % las frecuencias relativas expresadas en porcentajes.

Variables Cualitativas

Descripción Gráfica

a) Diagrama de Barras

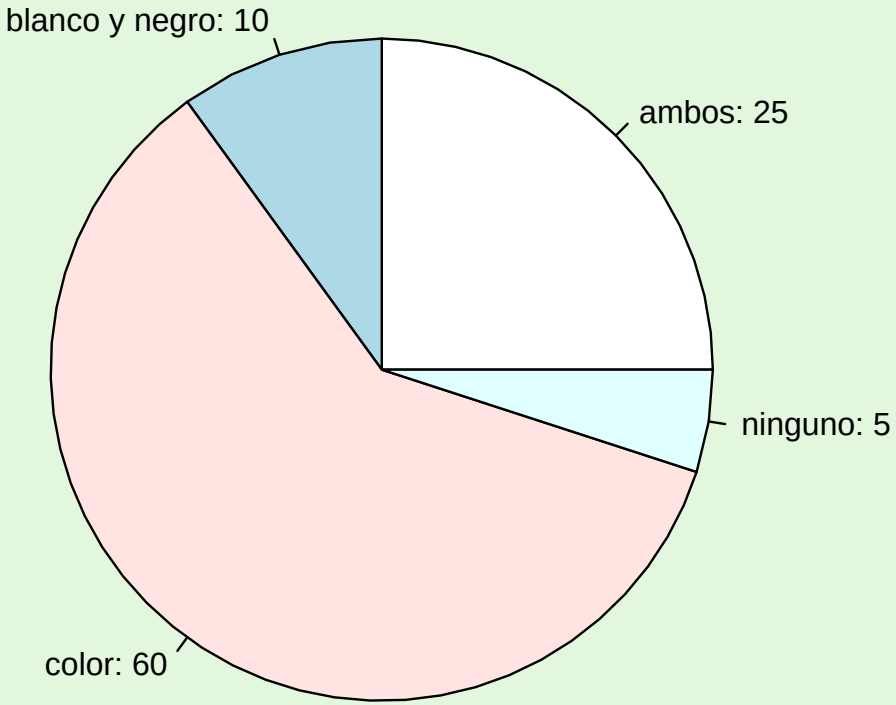


- Las alturas de las barras corresponden a las frecuencias absolutas o relativas.
- Hay una barra por cada una de las categorías.

Variables Cualitativas

Descripción Gráfica

b) Diagrama Circular o de Pastel



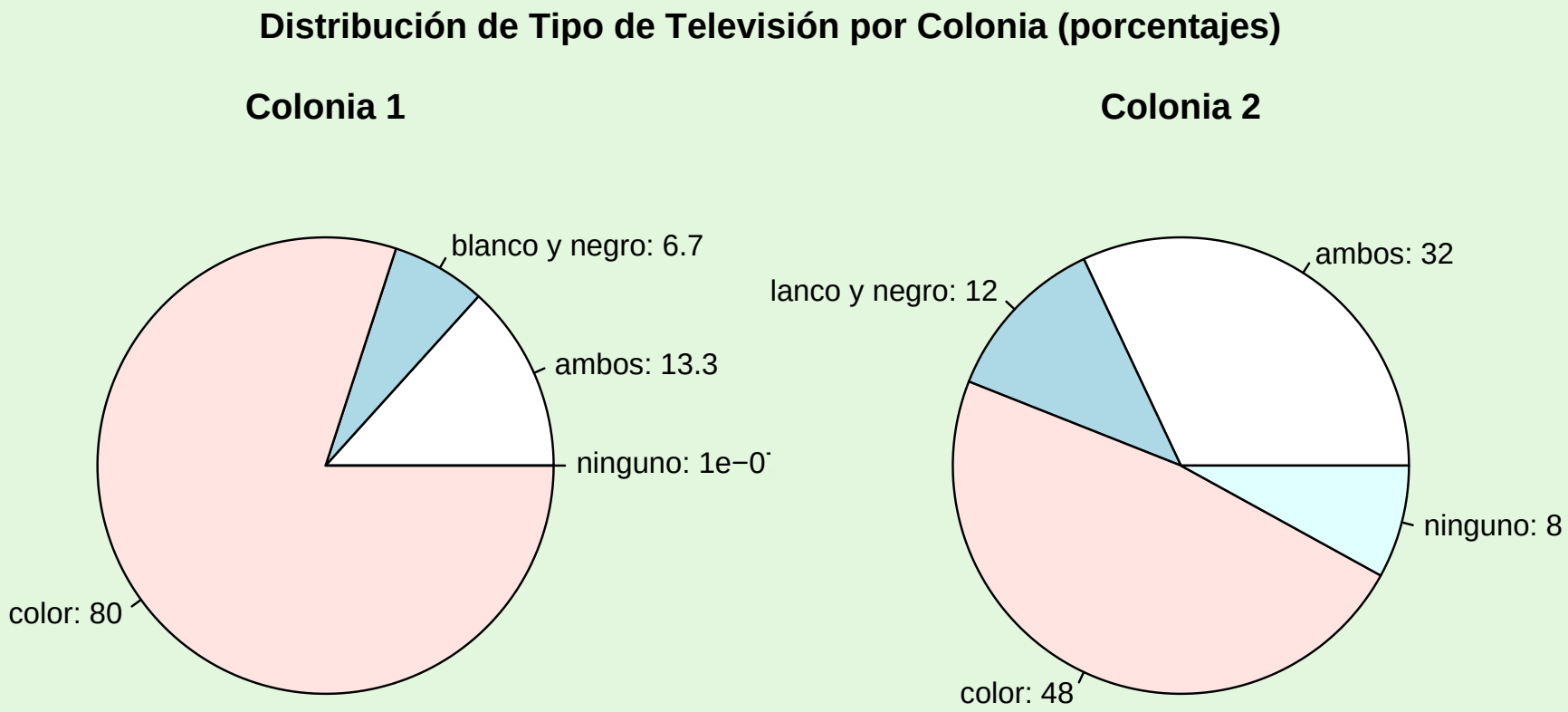
Los 360^o se dividen proporcionalmente de acuerdo a la frecuencia relativa p_i $(i = 1, \ldots, k)$.

Nota: Los diagramas de barras son preferibles sobre los de pastel. El ojo humano es bueno para juzgar medidas lineales pero malo en juzgar áreas relativas. Vea por ejemplo, la sección *Note* en:

`http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/pie.html`

Variables Cualitativas

Gráficas Circulares



• La presentación de gráficas de resultados para distintos grupos facilita el análisis.

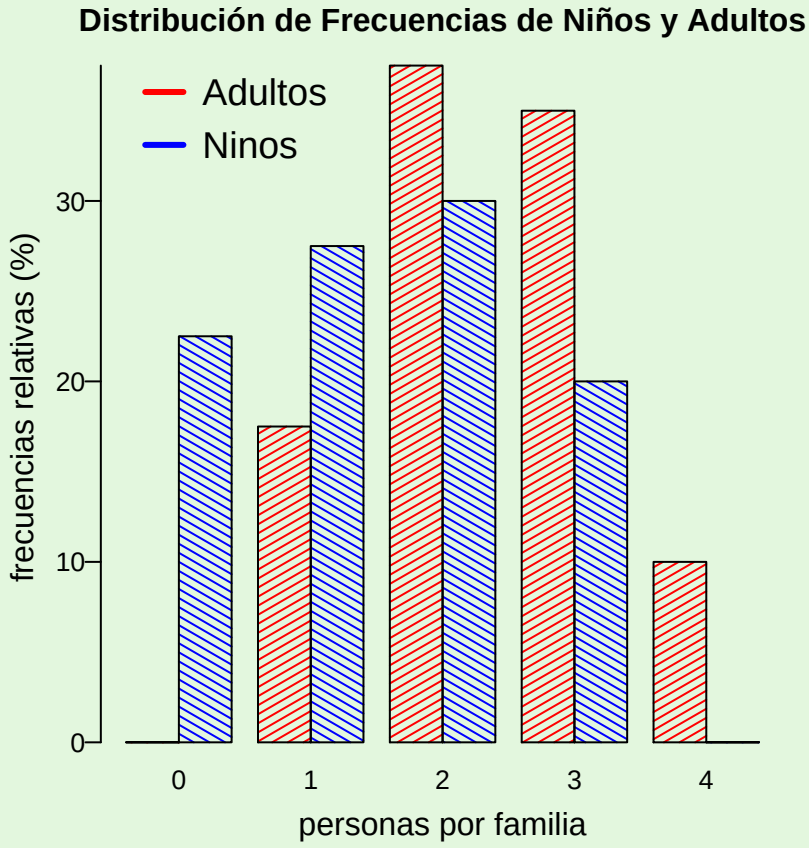
Variables Cuantitativas

Distribución de Frecuencias para Variables Discretas

Similar al diagrama de barras para variables cualitativas. Las categorías son los valores discretos.

Distribución de frecuencias para las variables *adultos* y *niños* (Encuesta de TV por cable)

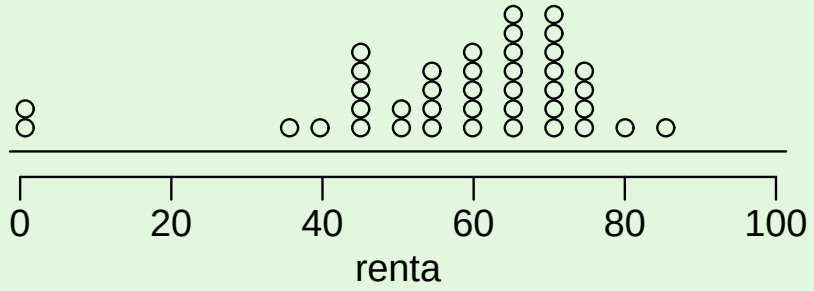
	adultos		niños	
valores	f_i	p_i	f_i	p_i
0	0	0.000	9	0.225
1	7	0.175	11	0.275
2	15	0.375	12	0.300
3	14	0.350	8	0.200
4	4	0.100	0	0.000
total	40	1.000	40	1.000



Variables Cuantitativas

Diagrama de Punto

Renta dispuesta a pagar por servicio de TV por cable



Variables Cuantitativas

Diagrama de Punto

Construcción:

- Eje horizontal representa el valor de las observaciones.
- Un punto (bolita) por cada observación. Para valores similares se coloca punto sobre punto.
- Fáciles de construir e interpretar cuando se tienen menos de 25 (50) datos y no hay tanta repetición de valores.

Características aparentes en los diagramas de punto:

- *Observaciones Atípicas* (“outliers”). Valores sustancialmente grandes o pequeños respecto al resto de los datos.
- *Huecos*. Espacios grandes entre grupo de puntos.
- *Perfil de la distribución*. Valores mas frecuentes.

Variables Cuantitativas

Diagrama de Tallo y Hojas

Construcción:

- Determinar el valor máximo y mínimo de las observaciones.
- Determinar una regla para separar los dígitos de cada observación en 2 partes (*tallos* y *hojas*). La regla se aplica a todos los datos.
- Para cada dato (observación) incluir una hoja en el tallo que corresponda.
- Una vez concluidos todos los datos se ordenan las hojas.

Tallos y Hojas para la variable *TVtot*

```
                  min=0, max=86
tallo | hojas                        tallo | hojas
    0 | 0 0                             0 | 0 0
    1 | 4 6                             1 | 4 6
    2 | 2 7 4 0 0 8 8 0 8               2 | 0 0 0 2 4 7 8 8 8
    3 | 0 4 8 1 5 5 2                   3 | 0 1 2 4 5 5 8
    4 | 0 2 0 2                         4 | 0 0 2 2
    5 | 6 4 2 4                         5 | 2 4 4 6
    6 | 8 2 0 9                         6 | 0 2 8 9
    7 | 6 4 0                           7 | 0 4 6
    8 | 2 2 4 6 4                       8 | 2 2 4 4 6
    NO ORDENADO                           ORDENADO
```

Variables Cuantitativas

Diagrama de Tallo y Hojas

Nota: En un diagrama de tallo y hojas se puede observar:

- Que tan alejados se encuentran los datos.
- Alrededor de que valor se concentran más los datos.
- Si hay muchos datos alejados del resto de las observaciones.
- Si hay simetría en los datos.
- Si hay grupos aislados.

Variables Cuantitativas

Diagrama de Tallo y Hojas

En ocasiones hay muchas hojas por tallo. En esos casos se pueden abrir los tallos para mayor detalle. E. g., Diagrama de tallo y hojas expandido para la variable *TVtot*.

```
0 | 0 0
  |
1 | 4
  | 6
2 | 0 0 0 2 4
  | 7 8 8 8
3 | 0 1 2 4
  | 5 5 8
4 | 0 0 2 2
  |
5 | 2 4 4
  | 6
6 | 0 2
  | 8 9
7 | 0 4
  | 6
8 | 2 2 4 4
  | 6
```

Variables Cuantitativas

Distribución de Frecuencias para Variables Continuas

En el caso de las variables continuas, puede suceder que no se repitan datos. Se construyen entonces intervalos para clasificar observaciones y se determinan las frecuencias de clase.

1. Determine máx, mín y rango $= \text{máx} - \text{mín} =$ amplitud. E. g., Variable *valor* en la encuesta de TV por cable.

   $$\text{máx} = 370325;\ \text{mín} = 79928;\ \text{rango} = \text{máx} - \text{mín} = 370325 - 79928 = 290379$$

2. Decidir cuántos intervalos de clase (k) usar, así como el ancho (c) de cada clase. (Recomendado $5 \leq k \leq 20$.) Elija el ancho del intervalo de modo que $k * c \geq$ rango (amplitud).

   Tomamos $k = 6, c = 50,000$.

Variables Cuantitativas

Distribución de Frecuencias para Variables Continuas

3. Elegir el valor inicial para el primer intervalo de clase. Este debe ser menor que el mínimo observado (mín).

   Tomamos 75, luego los intervalos de clase quedan:

clase	intervalos de clase	marca de clase	f_i	p_i (%)
1	(75, 125]	100	3	8
2	(125, 175]	150	8	20
3	(175, 225]	200	10	25
4	(225, 275]	250	8	20
5	(275, 325]	300	5	13
6	(325, 375]	350	6	15
			40	100

   Los datos agrupados pierden valores o magnitudes. Resulta conveniente definir el punto medio del intervalo como *representante o marca de clase* (m_i).

$$m_1 = \frac{75+125}{2} = 100, \quad m_2 = \frac{125+175}{2} = 150, \quad \ldots$$

Variables Cuantitativas

Distribución de Frecuencias para Variables Continuas

Otra característica de interés en datos cuantitativos es la *frecuencia acumulada*, absoluta (F_i), o relativa (P_i). Se obtiene sumando las frecuencias de todas las categorías menores incluyendo la clase en curso:

$$F_k = \sum_{i=1}^{k} f_i, \quad P_k = \sum_{i=1}^{k} p_i$$

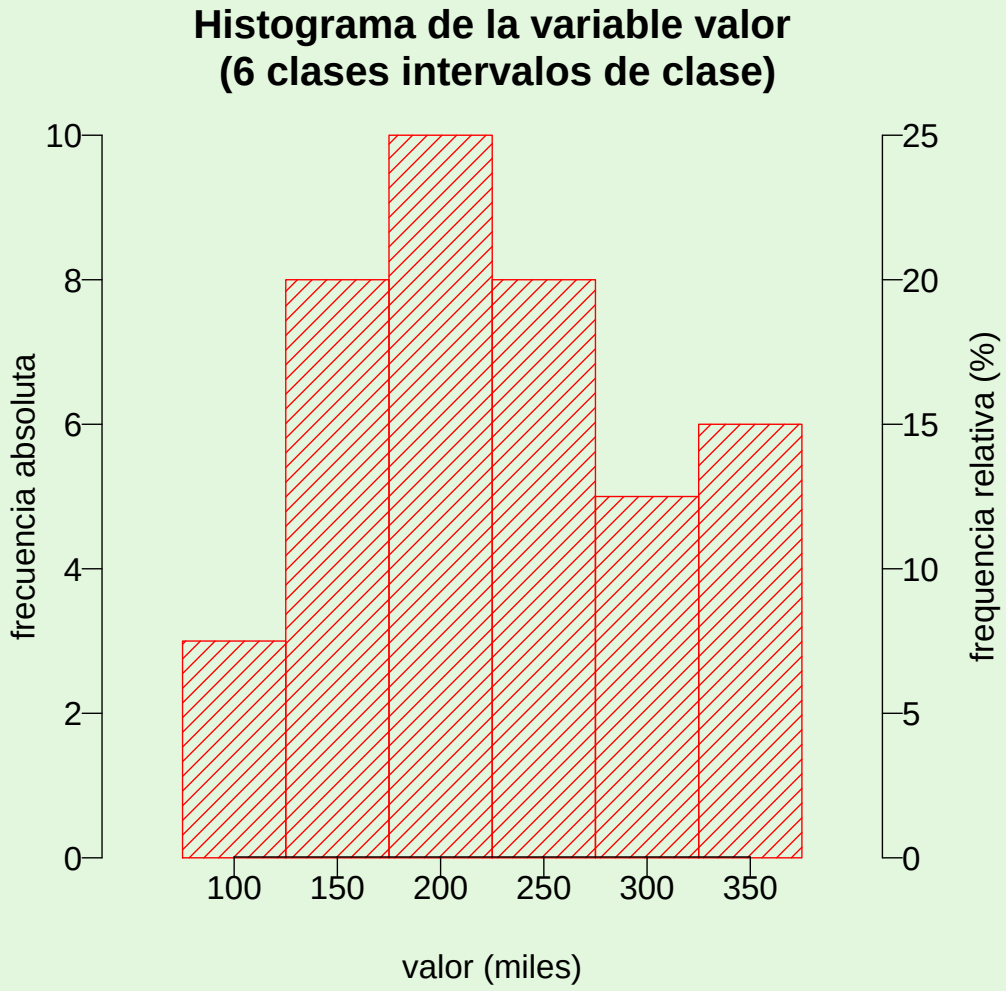
Tabla de frecuencias de la variable *valor*

	intervalos de	marca de	Por intervalo		Acumulada	
intervalo	clase	clase	absoluta	relativa	Absoluta	Relativa
i	I_i	m_i	f_i	p_i	F_i	P_i
1	(75, 125]	100	3	.08	3	.08
2	(125, 175]	150	8	.20	11	.28
3	(175, 225]	200	10	.25	21	.53
4	(225, 275]	250	8	.20	29	.73
5	(275, 325]	300	5	.13	34	.85
6	(325, 375]	350	6	.15	40	1.00

E. g., Podemos ver que, por ejemplo, 28 % de los hogares tienen valores catastrales menores a 175,000.

Variables Cuantitativas:

Distribución de Frecuencias para Variables Continuas: Histogramas

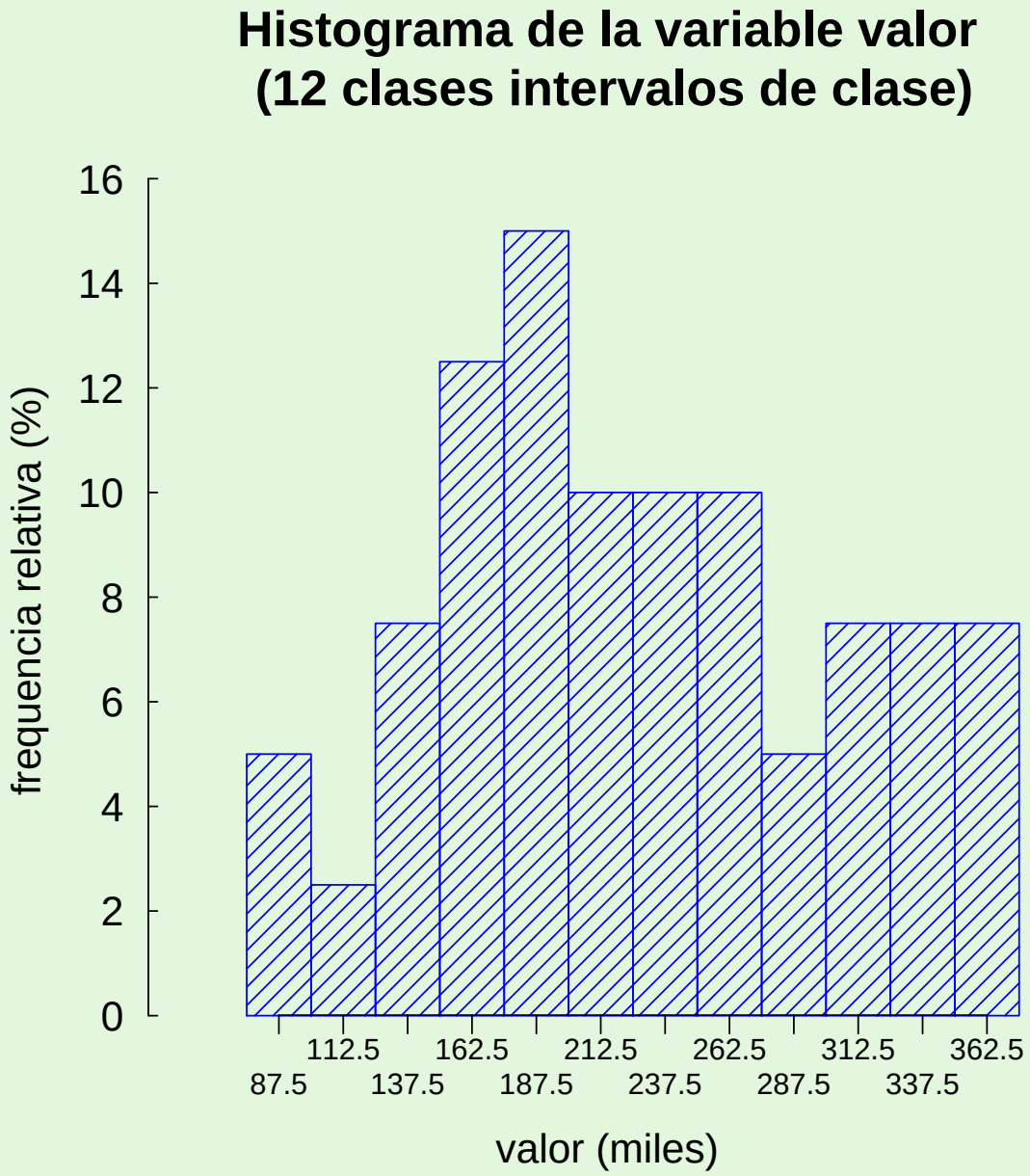


- Similar a los diagramas de barras para variables cualitativas. Las clases o categorías están formadas por los intervalos de clase.
- En un histograma las “barras” son adyacentes. Esto es por la *continuidad* de la variable graficada.

Variables Cuantitativas

Distribución de Frecuencias para Variables Continuas: Histogramas

- Si deseamos más detalle aumentamos el número de clases k.
- Si cambiamos el ancho de la clase (c) cambiarán las frecuencias.

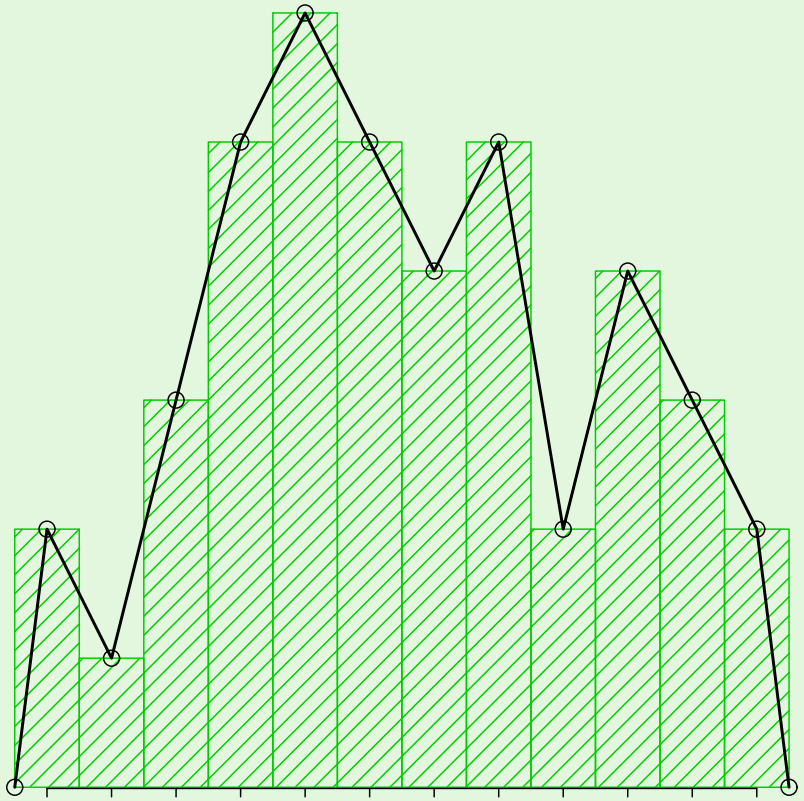


Relación entre histogramas y curvas poblacionales

Nota:

- Esperamos que la distribución de frecuencias nos sugiera un perfil similar al de la población.
- El perfil del histograma nos provee de una caracterización de la *variabilidad* y *distribución* de los valores de la población estadística.

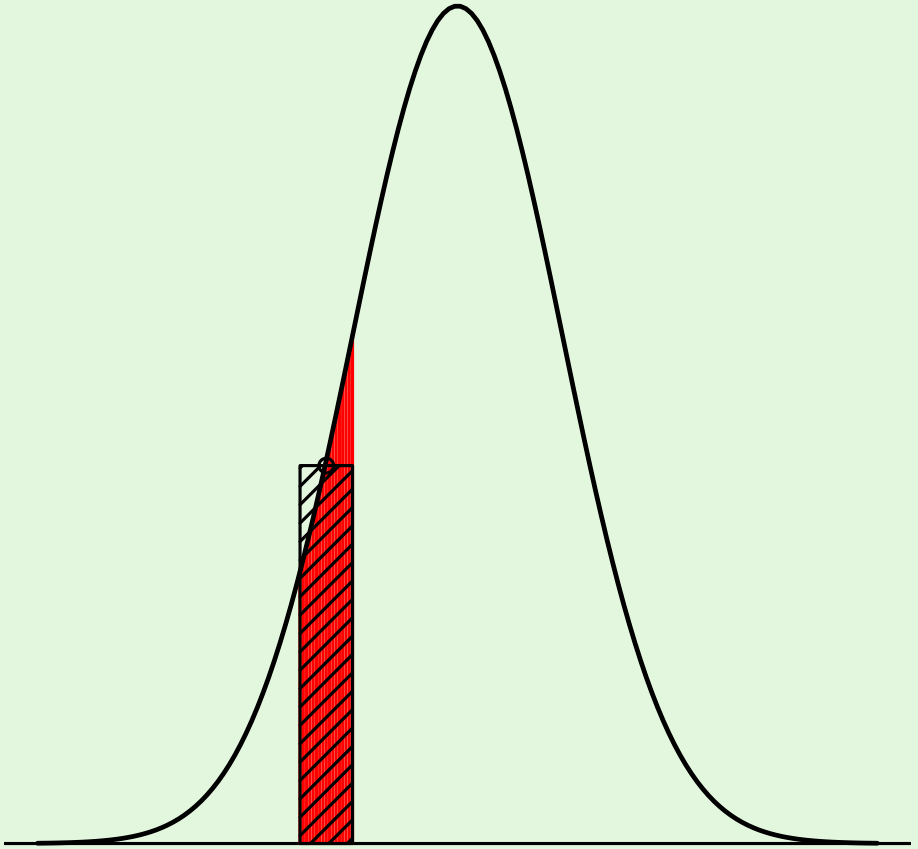
Histograma y polígono de frecuencias



Curvas poblacionales

El modelo matemático de la distribución de frecuencias poblacional de una variable continua se puede visualizar como la version *suavizada* de un histograma de *toda* la población.

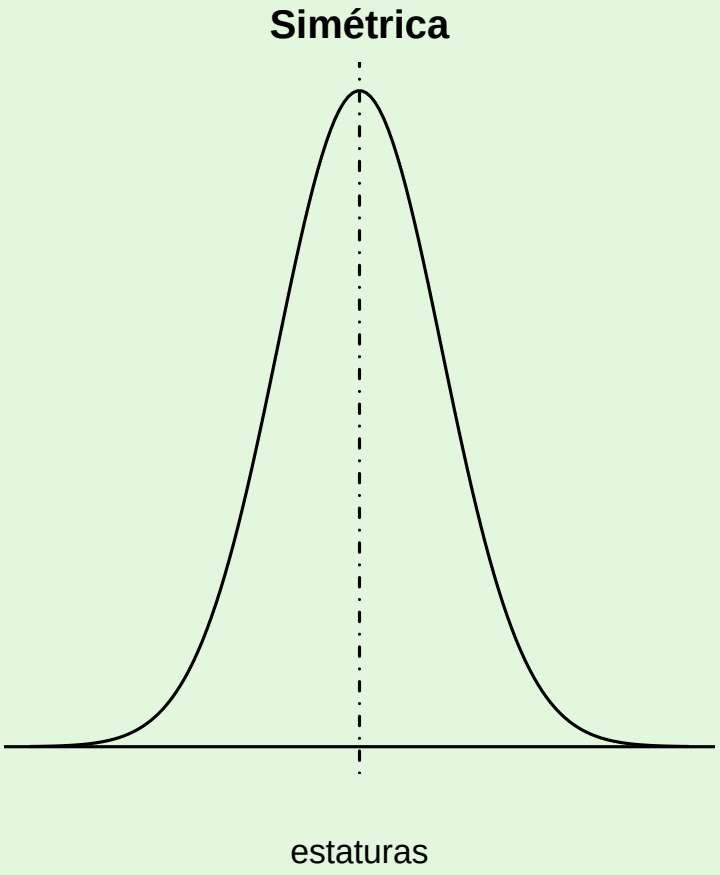
- Las frecuencias quedan por áreas bajo la curva.
- A la representación gráfica de las frecuencias poblacionales se le denomina *curva de distribución* de la frecuencia poblacional.
- Puede adquirir las siguientes formas:
  - Simétrica
  - Sesgada
  - Bimodal



Curvas poblacionales

Distribución Simétrica

- Se caracteriza por la existencia de un valor central alrededor del cual se distribuyen los valores probables de manera *simétrica*.
- E. g., la distribución de las estaturas de las estudiantes mujeres del ITAM.

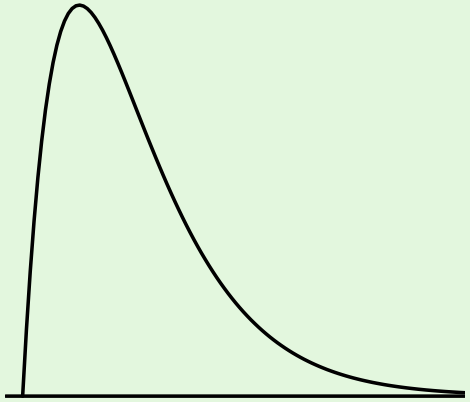


Curvas poblacionales

Distribución Sesgada

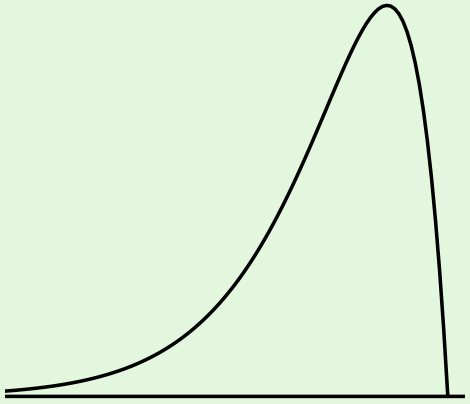
- Se caracteriza porque una de las extremidades (colas) está más extendida que la otra. La dirección del *sesgo* corresponde a la extremidad de mayor extensión.
- E. g., La distribución del ingreso es sesgada a la derecha. La tasa de retiros o jubilaciones es sesgada a la izquierda.

Sesgada a la derecha



pesos
ingreso

Sesgada a la izquierda

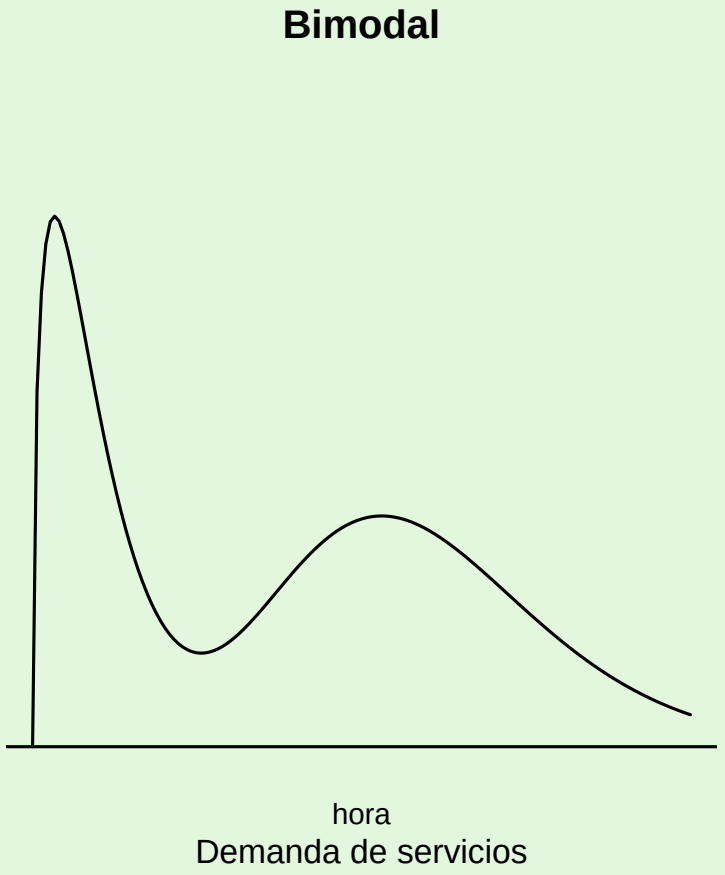


edad
tasa de retiro o jubilación

Curvas poblacionales

Distribución Bimodal

- Se caracteriza por tener dos cimas (*modas*) o "jorobas" separadas indicando la combinación de dos grupos con diferentes distribuciones.
- E. g., Distribución en el día del número de personas demandando un servicio. Las modas corresponderían poco después de abrir la oficina en la mañana y tarde.



Curvas poblacionales

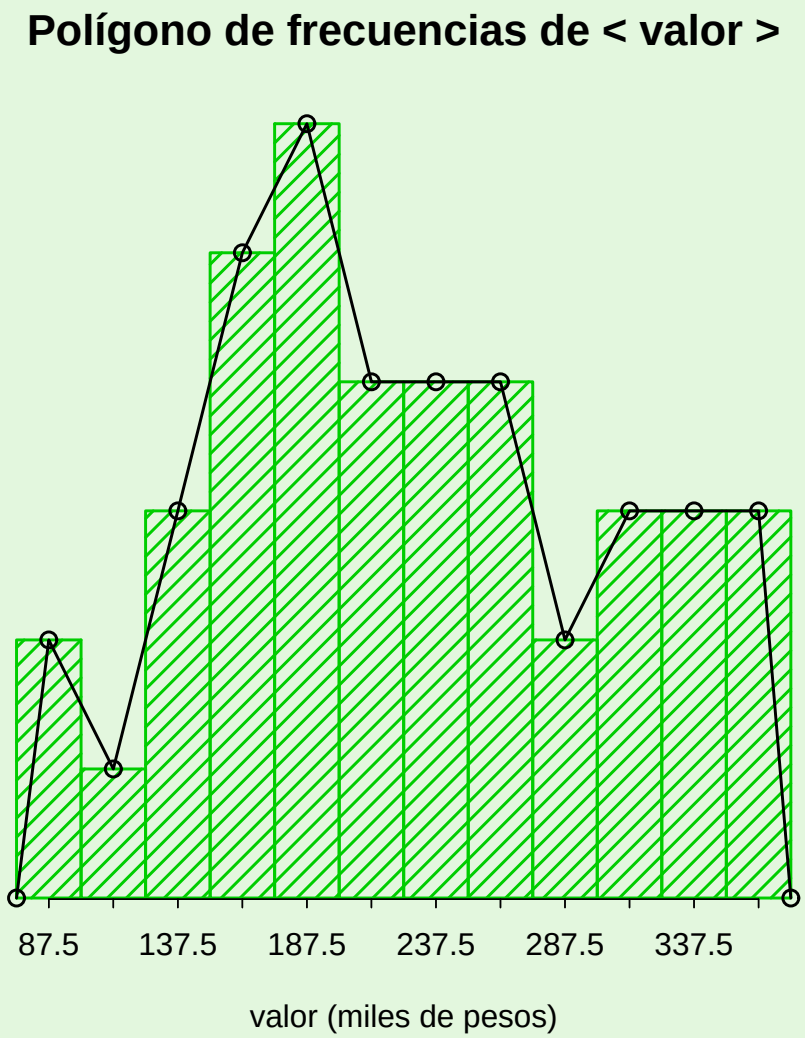
Polígono de Frecuencias

Construcción:

- Se unen los puntos medios de la parte superior de las barras del histograma y se cierran los extremos con el eje horizontal.

E. g.,

- La gráfica muestra el histograma y polígono de frecuencias de la variable *valor* en la encuesta para el estudio de TV por cable. Nótese la posible *bimodalidad* de la muestra.

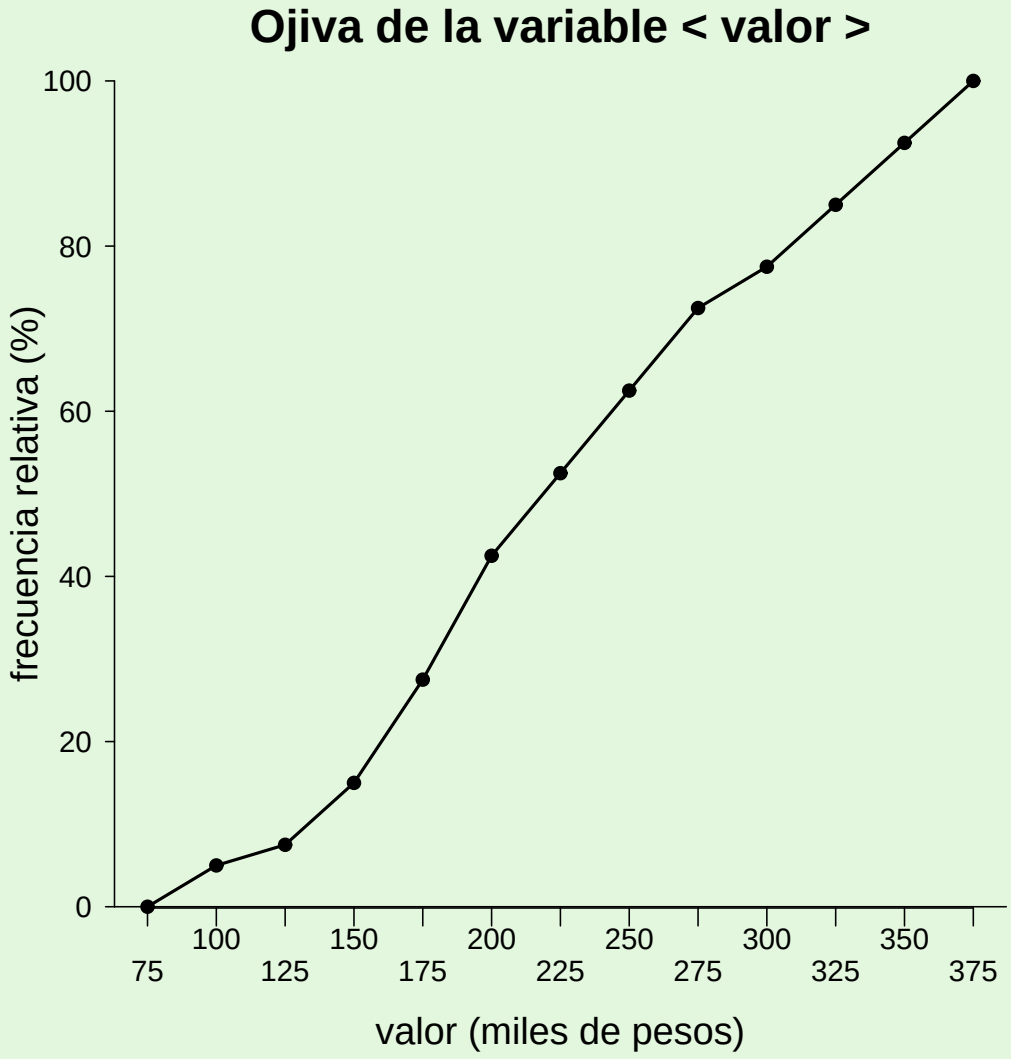


Curvas poblacionales

Ojiva

Construcción:

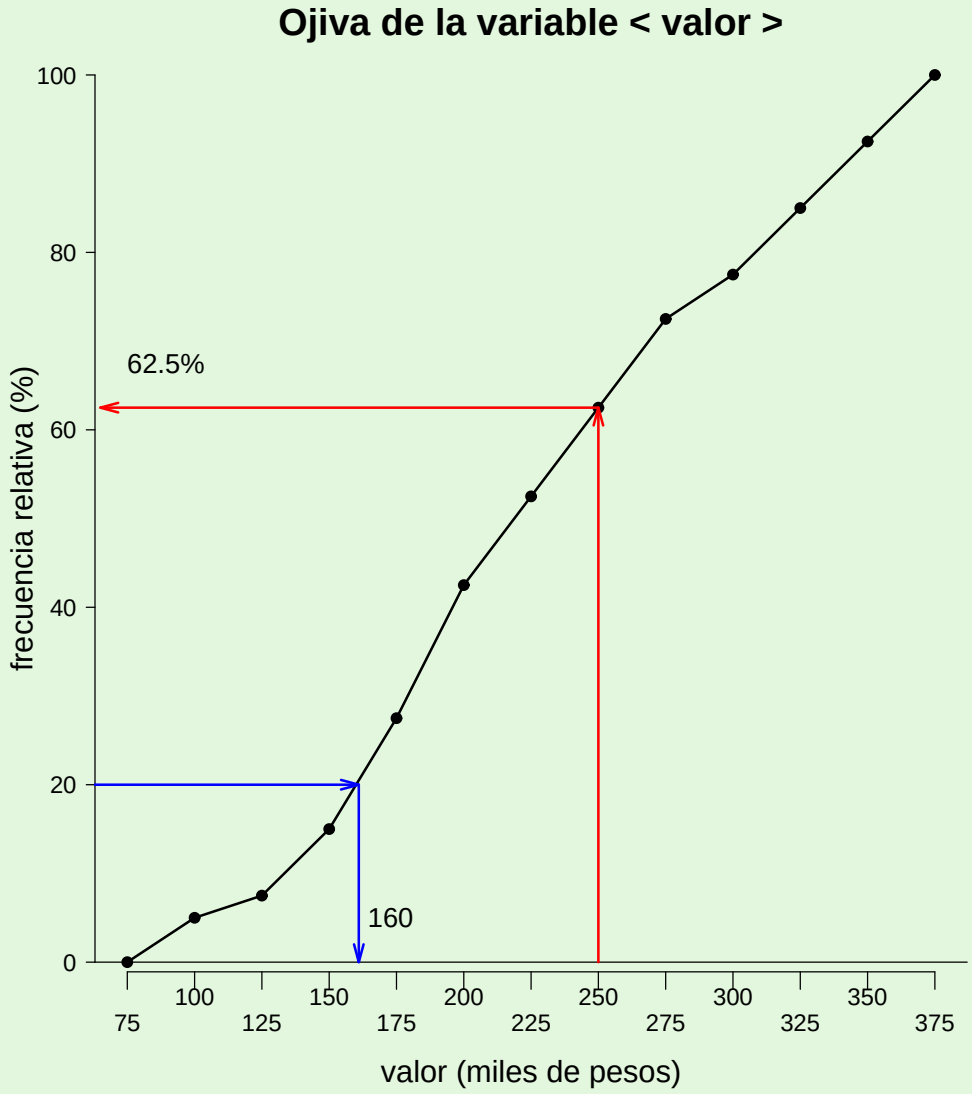
- La *ojiva* es la curva que resulta de graficar las frecuencias relativas acumuladas contra el límite superior de los intervalos de clase.



Curvas poblacionales

Ojiva

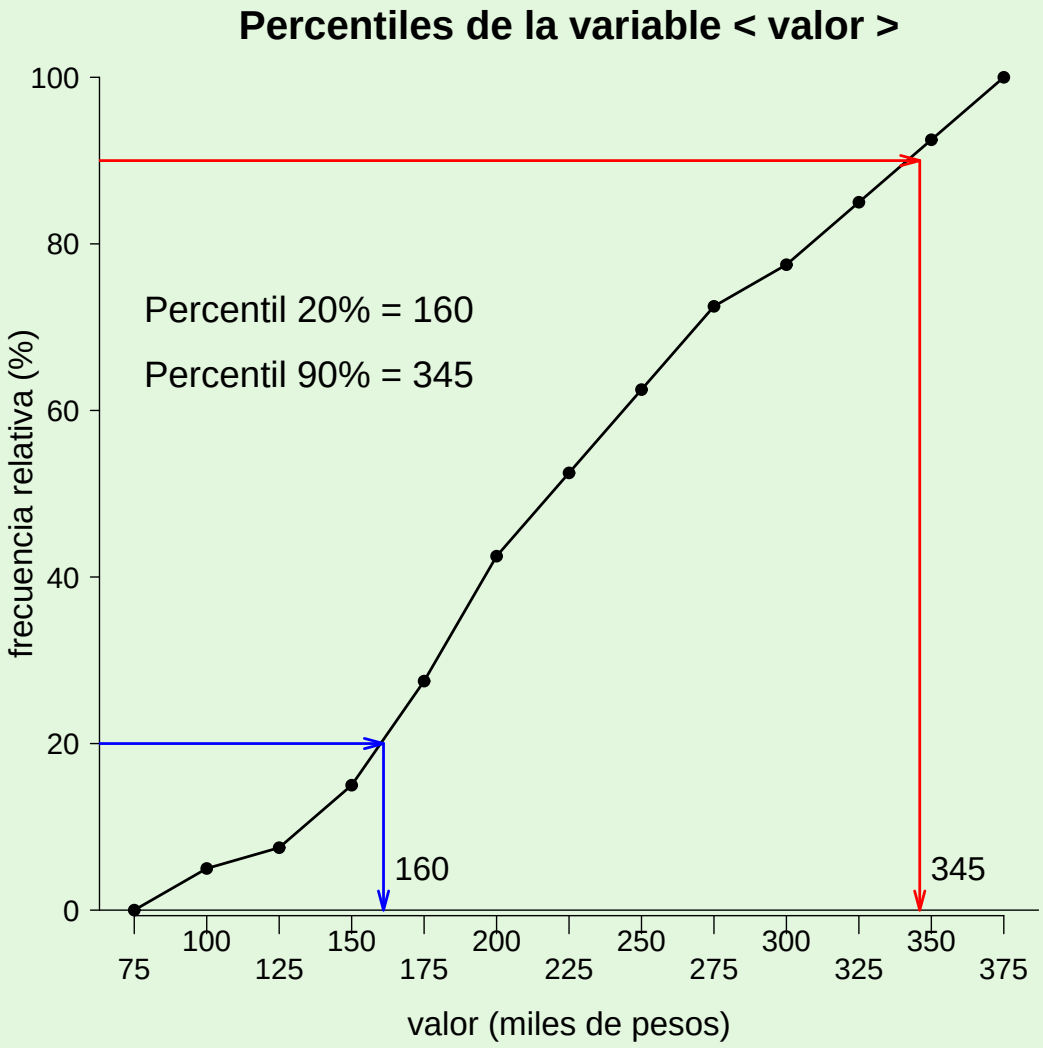
- Si deseamos obtener el porcentaje de casas cuyo valor catastral es menor de $250,000, localizamos en el eje horizontal el valor y lo subimos a que corte la ojiva y leemos el valor en el eje vertical.
- Si deseamos saber (estimar) cuál es el valor catastral del primer 20 % de la población, trazamos una recta horizontal a la altura de 20 % hasta cortar la ojiva, después proyectamos verticalmente y leemos la cantidad en el eje vertical.



Curvas poblacionales

Percentiles

- Los *percentiles* dan información acerca de cómo se distribuyen los valores de la variable en estudio.
- Si p está entre 0 % y 100 %, el *p-ésimo percentil* es el valor de la abscisa (eje horizontal) tal que al menos $100p\,\%$ tienen un valor menor o igual a él.
- En el ejemplo de TV por cable, el percentil del 20 % es aproximadamente \$160,000 y el 90-percentil es de \$345,000. Luego, solamente 10 % de casas tiene un valor catastral mayor a \$345,000.



Distribución de Frecuencias

Agrupación de Variables

La agrupación de variables consiste en formar una variable cualitativa o categórica combinando los valores de otra variable.

• De esta manera se puede convertir una variable categórica en otra pero con menos clases. E. g., tipo de televisión en la encuesta de TV por cable.

tipo	p_i	televisión	$p_i{}'$
Ninguna	0.05	sin	0.15
Blanco y negro	0.10		
Color	0.60	con	0.85
Ambos	0.25		

• De igual forma se pueden agrupar variables cuantitativas en categóricas. En este caso se definen las categorías en términos del valor de la variable en estudio. E. g., para la variable *valor* (en miles de pesos) definimos las clases *bajo* (< 200), *medio* ($200 \leq$ valor ≤ 300), y *alto* (> 300).

clase	intervalo	f_i	p_i	F_i	P_i
bajo	(75,200)	17	.425	17	0.425
medio	[200,300]	14	.350	31	0.775
alto	(300,400)	9	.225	40	1.000

Medidas Descriptivas

Además de la descripción gráfica de la variación de los valores de una muestra o de la población total, existen algunos *números* que ayudan a mostrar aspectos relevantes de la distribución de frecuencias.

Estas pueden describir alrededor de qué valor los datos se distribuyen, así como conocer qué tanto varían los datos alrededor de estos valores.

A estos valores se les denomina *medidas de tendencia central* y *medidas de variabilidad*, respectivamente. En conjunto se les conoce como *medidas descriptivas*.

Medidas de Tendencia Central

Las medidas de tendencia central son valores numéricos que intentan, en cierto sentido, localizar la parte central de la distribución de frecuencias.

Mediana

Es el percentil del 50 %. Es el valor que ocupa la posición central de los datos después que han sido ordenados de manera ascendente. Luego, 50 % de los datos es menor o igual que la mediana, y el otro 50 % es mayor o igual que la mediana.

Denotamos M a la mediana de la distribución de valores poblacionales y m ($\tilde{x}$) a la mediana de la distribución de una muestra.

La mediana es una *medida de tendencia central* útil en casos de distribuciones sesgadas.

Medidas de Tendencia Central

Mediana

Calculo:

- Uso de la ojiva de la distribución. Recuerde que la mediana es el percentil del 50 %.
- Uso del diagrama de tallos y hojas *ordenado*. Localice el valor central. Si el numero de hojas n es impar la mediana corresponde al valor en la posición $(n+1)/2$. Si el numero de hojas n es par, la mediana será el promedio de los valores centrales (en las posiciones $n/2$ y $(n+2)/2$).
- Sea $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, la muestra o población ordenada de menor a mayor. Considere el índice

$$\ell = 0.5n + 0.5 = k + \delta$$

donde k es la parte entera de ℓ y $0 \leq \delta < 1$ la parte decimal.

$$m = x_{([\![\ell]\!])} \begin{cases} x_{(k)} & \text{si } \delta < 0.5 \\ x_{(k+1)} & \text{si } \delta > 0.5 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & \text{si } \delta = 0.5 \end{cases}$$

Medidas de Tendencia Central

Ejemplo 1.4.1

Una empresa fabricante de productos cosméticos y de limpieza maneja ventas de alrededor de 400 productos distintos a través de once centros de acopio en toda la república. Dado el gran volumen de producto que se maneja es importante que haya un buen control de inventarios, ya que si se tienen mucho inventario ocioso significa dinero que no se está empleando en producir, mientras que un inventario escaso significa tener una demanda no satisfecha.

La empresa contrató los servicios de un bufete y recibió la siguiente fórmula para el control de inventarios:

$$\text{nivel de reabastecimiento} = 1.3 * \text{ días en tránsito * venta máxima}$$

donde *días en tránsito* significa el número de días que tarda en llegar un pedido al centro de acopio, y el factor 1.3, el bufete lo llamó el factor de "paranoia".

Medidas de Tendencia Central

Ejemplo 1.4.1 (cont.)

Ventas diarias de suavizante para ropa. Número de cajas vendidas. Centro de acopio Guadalajara.

semana 1	semana 2	semana 3	semana 4	semana 5	semana 6	semana 7
0	2838	413	5592	0	465	2199
515	590	47	673	80	703	
746	331	340	561	159	462	
1237	450	265	548	183	175	
879	570	1083	216	113	422	

La tabla presenta la venta en cajas de un suavizante para ropa en el centro de acopio de Guadalajara. La planta de manufactura está en México, por lo que los días de tránsito son 3. Aplicando la fórmula anterior se obtiene que para Guadalajara el nivel de reabastecimiento es de 21,808. Claramente, ésta es una recomendación exagerada de inventario. La empresa decidió hacer un estudio estadístico. Para comenzar se analizaron los pedidos diarios de los últimos meses. La tabla presenta las ventas del último mes y medio ya que las conclusiones son similares.

Medidas de Tendencia Central

Ejemplo 1.4.1 (cont.)

NO ORDENADO	ORDENADO
0 | 00, 47, 00, 80	0 | 00, 00, 47, 80
1 | 59, 83, 13, 75	1 | 13, 59, 75, 83
2 | 65, 16	2 | 16, 65
3 | 31, 40	3 | 31, 40
4 | 50, 13, 65, 62, 22	4 | 13, 22, 50, 62, 65
5 | 15, 90, 70, 61, 48	5 | 15, 48, 61, 70, 90
6 | 73	6 | 73
7 | 46, 03	7 | 03, 46
8 | 79	8 | 79
9 |	9 |
10 | 83	10 | 83
11 |	11 |
12 | 37	12 | 37
21 | 19	21 | 19
28 | 38	28 | 38
55 | 92	55 | 92

El número de datos es 31, luego la mediana corresponde al dato $(31+1)/2 = 16$, es decir, $m = 462$.

Es claro que resulta cuestionable un inventario de cerca de 20,000 cajas cuando el 50 % de las ventas es menor que 462 cajas.

Medidas de Tendencia Central

Media

De las medidas de tendencia central la *media* es la más común. Ésta es el promedio aritmético de los datos. Conceptualmente, el promedio de todas las mediciones de la población estadística es la *media poblacional* y se denota por μ (letra griega llamada "mu").

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

donde N es el total de la población. La *media muestral* se denota por $\bar{x}$ y está dada por el promedio de los valores de la muestra. Esto es,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

donde n es el tamaño de la muestra.

Medidas de Tendencia Central

Media

En el ejemplo 1.4.1, $\bar{x} = 734.68$. La media y la mediana están alejadas pues la distribución muestral es sesgada a la derecha. En este caso la media no es un buen indicador de la tendencia central.

$\star$ La media es una buena medida de tendencia central cuando la muestra no es muy sesgada y no hay valores atípicos.

Medidas de Tendencia Central

Media

Cálculo de la media usando la distribución de frecuencia - Datos Agrupados

El cálculo de la media a partir de una tabla de frecuencias es aproximado pues no se cuenta con el detalles de los datos. En este caso,

$$\bar{x}_g = \frac{1}{n}\sum_{i=1}^{k} f_i \cdot m_i$$

donde f_i y m_i son la frecuencia absoluta y la marca de clase del i-ésimo intervalo de la distribución de frecuencias y k es el número de intervalos o clases. Luego,

$$\bar{x}_g = \frac{1}{n}\sum_{i=1}^{k} f_i m_i = \sum_{i=1}^{k} \frac{f_i m_i}{n} = \sum_{i=1}^{k} \frac{f_i}{n} m_i = \sum_{i=1}^{k} p_i m_i$$

Medidas de Tendencia Central

Media

Cálculo de la media usando la distribución de frecuencia - Datos Agrupados

En el estudio de TV por cable, la media de la variable *valor* está dado por:

$$\begin{aligned} \bar{v}_g &= 0.075(100000) + 0.200(150000) + \cdots + 0.150(350000) \\ &= 227500 \end{aligned}$$

El promedio aritmético de las observaciones es realmente 227966, bastante cercano al calculado de la tabla de frecuencias. En este caso la mediana de la muestra es $m = 216393$, cercano también a $\bar{v}$, pues la distribución de *valor* no es muy sesgada.

Medidas de Tendencia Central

Mediana

Cálculo de la mediana usando la distribución de frecuencia - Datos Agrupados

$$\tilde{x}_g = x_{p_1} + (x_{p_2} - x_{p_1}) \cdot \frac{0.5 - p_1}{p_2 - p_1}$$

donde, x_{p_1} es el límite inferior del intervalo que contiene la mediana con una frecuencia acumulada de p_1; x_{p_2} es el límite superior del intervalo que contiene la mediana con una frecuencia acumulada de p_2.

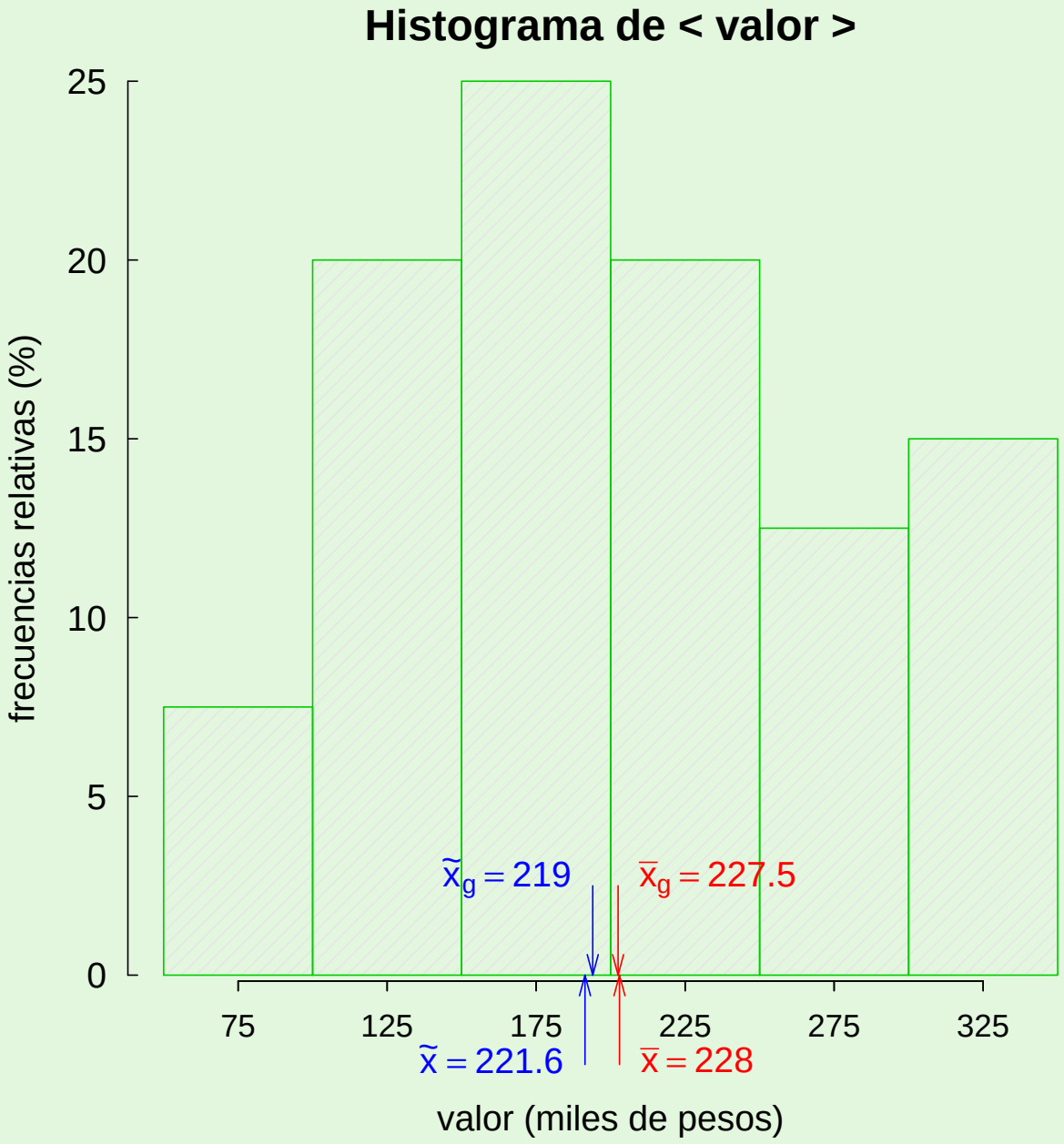
En el estudio de TV por cable, la mediana de la variable *valor*, calculada en base a los datos agrupados es (en miles de pesos):

$$\tilde{v}_g = 175 + (225 - 175)\frac{0.50 - 0.28}{0.53 - 0.28} = 219$$

cuando en realidad, a partir de los datos individuales es 216.393.

Medidas de Tendencia Central

Media y Mediana



Medidas de Tendencia Central

Moda

Para un conjunto de datos discretos la *moda* se define como aquel valor que ocurre con mayor frecuencia. Si el valor es único, decimos que la distribución de frecuencias es *unimodal*. Para ver si hay mas de una moda, es conveniente observar la gráfica de barras de la distribución de frecuencias y buscar cimas (picos). Los valores debajo de las cimas serán los candidatos a modas.

En el caso de variables continuas, a partir de los polígonos de frecuencias, aquellos picos o cimas aparentes corresponderán a valores de posibles modas.

Percentiles o Medidas de Posición

Los *percentiles* o medidas de posición son otras medidas descriptivas de gran utilidad.

Recuerde que la *mediana* corresponde al percentil del 50 %, esto es, aquel valor que divide los datos en dos partes iguales (50 % cada una).

Los *cuartiles* son los valores que dividen al conjunto de datos ordenados en cuatro partes. Es decir, aquellos valores que tienen 25 %, 50 % y 75 % de los valores de la distribución de frecuencias por debajo de ellos.

Percentiles o Medidas de Posición

Mediana

• Sea $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, la muestra o población ordenada de menor a mayor. Considere el índice

$$\ell = 0.5n + 0.5 = k + \delta$$

donde k es la parte entera de ℓ y $0 \leq \delta < 1$ la parte decimal.

$$m = x_{([\![\ell]\!])} \begin{cases} x_{(k)} & \text{si } \delta < 0.5 \\ x_{(k+1)} & \text{si } \delta > 0.5 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & \text{si } \delta = 0.5 \end{cases}$$

Percentiles o Medidas de Posición

Cuartil inferior o primer cuartil

Tiene por debajo a 25 % de los valores de la distribución de frecuencias. El primer cuartil *poblacional* se denota por Q_1 mientras que el primer cuartil *muestral* por q_1.

En datos ordenados el primer cuartil se localiza en

$$\ell(q_1) = 0.25n + 0.5$$

y el *primer cuartil* es:

$$q_1 = x_{([\![\ell(q_1)]\!])}$$

En el ejemplo de los inventarios de productos cosméticos,

$$\ell(q_1) = 0.25(31) + 0.5 = 8.25$$

y del diagrama de tallos y hojas,

$$q_1 = x_{(8)} = 183$$

Percentiles o Medidas de Posición

Cuartil superior o tercer cuartil

Tiene por debajo a 75 % de los valores de la distribución de frecuencias. El tercer cuartil *poblacional* se denota por Q_3 mientras que el tercer cuartil *muestral* por q_3.

El tercer cuartil se localiza en

$$\ell(q_3) = 0.75n + 0.5$$

y el *tercer cuartil* es:

$$q_3 = x_{([\![\ell(q_3)]\!])}$$

En el ejemplo de inventarios de productos cosméticos,

$$\ell(q_3) = 0.75(31) + 0.5 = 23.75$$

y del diagrama de tallos y hojas,

$$q_3 = x_{(24)} = 703$$

Percentiles o Medidas de Posición

Percentiles

El p-*ésimo percentil* es aquel valor que deja $p\,\%$ de los datos ordenados (de menor a mayor) por debajo ($0 \leq p \leq 100$)

Cálculo del p-ésimo percentil:

- Ordenar los datos de manera ascendente.
- Calcular el índice

$$\ell_p = \frac{p}{100}\, n + 0.5$$

donde p es el percentil de interés y n es el tamaño de muestra. El p-*ésimo percentil* será entonces

$$x_p = x_{([\![\ell_p]\!])}$$

Deciles

Los *deciles* son los percentiles del 10 %, 20 %, . . . , 90 %.

Percentiles o Medidas de Posición

Valores observados (en miles de pesos) de la variable *valor* en la encuesta:

```
  79.928  94.415 120.896 132.867 141.901 147.997 156.410 156.841 157.041 161.222
 162.509 180.124 180.437 190.314 192.265 192.816 193.279 205.656 216.190 216.321
 216.465 225.694 237.752 241.531 249.098 252.221 261.763 269.898 271.556 279.163
 299.558 311.195 318.551 322.652 329.198 332.699 336.290 355.641 357.972 370.325
122.0745
```

Para calcular el *primer decil*, i. e., el percentil del 10 %:

$$\ell_{10} = \frac{10}{100}\,40 + 0.5 = 4.5$$

Luego,

$$p_{10} = (132.867 + 141.901)/2 = 137.384$$

Similarmente, para el cuarto decil (40 %):

$$\ell_{40} = \frac{40}{100}\,40 + 0.5 = 16.5$$

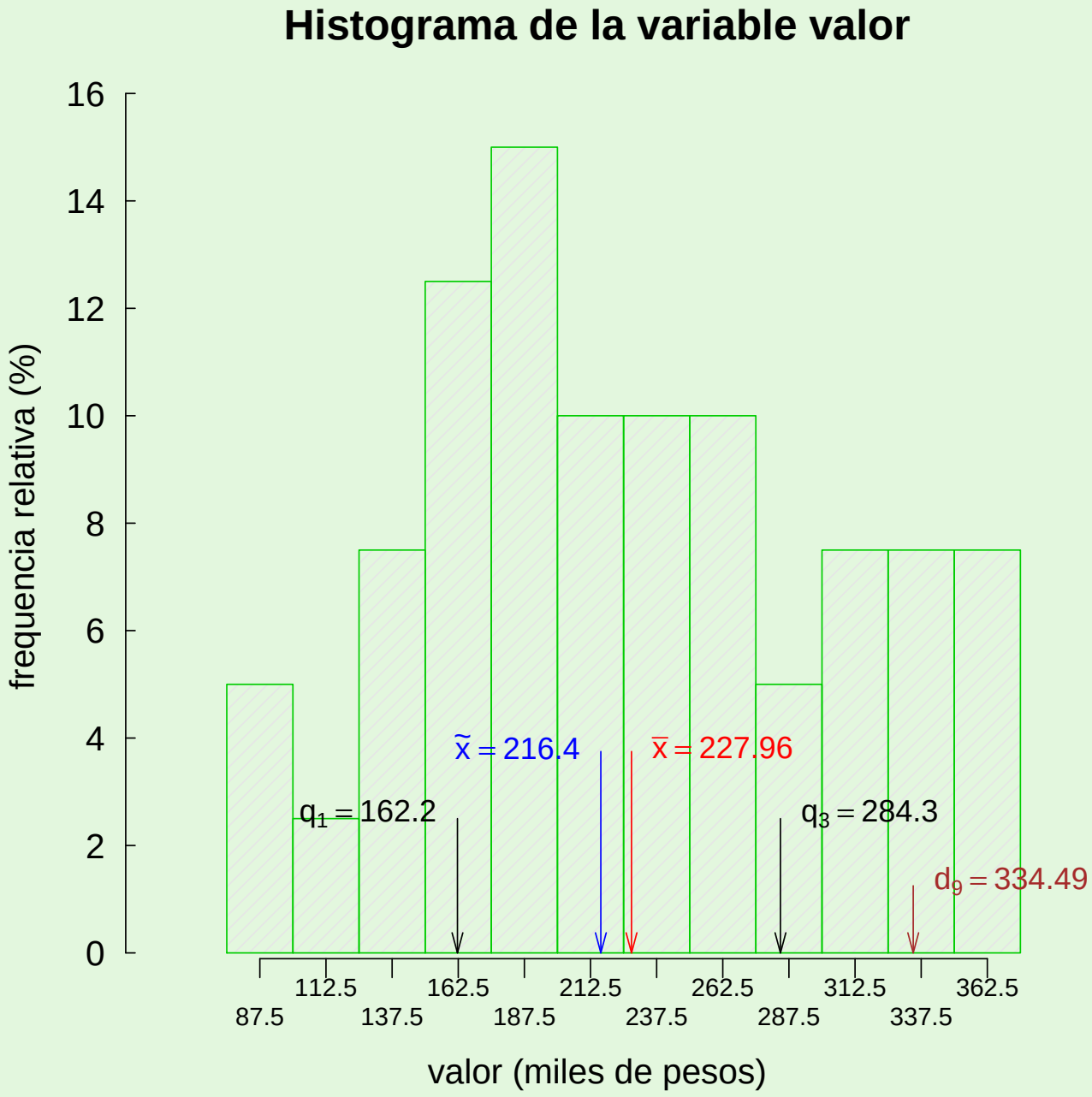
$$p_{40} = (192.816 + 193.279)/2 = 193.0475$$

Deciles para la variable *valor* en miles de pesos

porcentaje (%)	ℓ_p	decil
10	4.5	**137.3840**
20	8.5	156.9410
30	12.5	180.2805
40	16.5	**193.0475**
50	20.5	216.3930
60	24.5	245.3145
70	28.5	270.7270
80	32.5	314.8730
90	36.5	334.4950

Medidas de Tendencia Central y de Posición

Media, Mediana y Cuartiles



Medidas de Dispersión

Las medidas de tendencia central nos sirven para saber alrededor de qué valores se distribuyen las observaciones pero no qué tanto estos datos varían.

Las *medidas de dispersión* nos dicen que tanto varían los datos. Estas medidas serán pequeñas si no hay mucha diferencia entre las observaciones y serán grandes en caso contrario.

Medidas de Dispersión

Amplitud o Rango (R)

Amplitud o Rango (R). Mide la distancia entre el mayor ($\mathrm{máx}$) y el menor ($\mathrm{mín}$) de los datos (x's).

$$R = \text{amplitud} = x_{\mathrm{máx}} - x_{\mathrm{mín}}$$

E. g., en el caso de la variable *valor*:

$$R = 370.30 - 79.93 = 290.47$$

Amplitud Intercuartílica (A.I.) o Rango Intercuartiles (R.I.). También se basa en la distancia entre cuartiles. Es la diferencia entre el *cuartil superior* (q_3) y el *cuartil inferior* (q_1). E. g., para la variable *valor*

$$\text{A. I.} = q_3 - q_1 = 284.30 - 162.20 = 122.1$$

Medidas de Dispersión

Varianza (σ^2)

La suma de desviaciones de la media muestral (promedio aritmético $\bar{x}$) es cero. Esto es, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Luego, aunque los datos varíen mucho, la suma de desviaciones es nula. Para evitar la cancelación de valores mayores de la media con los valores menores, se suman las desviaciones al cuadrado.

De manera similar como lo hicimos con la media poblacional μ, se define la *varianza poblacional* como el valor medio de las desviaciones al cuadrado. Esto es,

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

La *varianza muestral* se define como la suma de desviaciones respecto a la media $\bar{x}$ al cuadrado y se denota por s^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

En la práctica, la varianza muestral se calcula aprovechando la igualdad:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Medidas de Dispersión

Varianza (σ^2)

De la misma manera que esperamos que media muestral $\bar{x}$ esté cercana de la poblacional μ, también esperamos que la varianza muestral s^2 esté cerca de σ^2.

Note que la varianza se expresa en las unidades originales al cuadrado. Entonces, la raíz cuadrada de la varianza está en las unidades originales y se conoce como la *desviación estándar*: $\sigma = \sqrt{\sigma^2}$, y $s = \sqrt{s^2}$.

característica	poblacional	muestral
media	μ	$\bar{x}$
mediana	M	m
varianza	σ^2	s^2
desviación estándar	σ	s

Nota: Tanto la varianza como la desviación estándar son medidas *no* resistentes o robustas, en el sentido de que son sensibles a datos extremos.

Medidas de Dispersión

Varianza (σ^2)

Cálculo de la varianza usando la distribución de frecuencia - Datos Agrupados

El cálculo de la varianza en una distribución de frecuencias es similar a como se hizo con la media. Las marcas de clase m_i $(1 \leq i \leq k)$ son representantes de las observaciones en el correspondiente intervalo de clase.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} f_i (m_i - \bar{x}_g)^2$$

donde f_i es la frecuencia absoluta correspondiente al i-ésimo intervalo de clase. Alternativamente, s^2 también se puede calcular como

$$s^2 = \frac{\sum_{i=1}^{k} f_i m_i^2 - n\bar{x}_g^2}{n-1}$$

Medidas de Dispersión

Varianza (σ^2)

Cálculo de la varianza usando la distribución de frecuencia - Datos Agrupados

Para la variable *valor*, $\bar{x} = 227,966$

$$\begin{aligned} s^2 &= \tfrac{1}{40-1}\left[(79,928 - 227,966)^2 + \cdots + (370,325 - 227,966)^2\right] \\ &= 5,973.786,750 \end{aligned}$$

o bien, utilizando la distribución de frecuencias con 6 intervalos de clase ($\tilde{x}_g = 227,500$),

$$\begin{aligned} s_g^2 &= \tfrac{1}{40-1}\left[3(100,000)^2 + 8(150,000)^2 + \cdots + 6(350,000)^2 - 40(227,500)^2\right] \\ &= 5,762.820,513 \end{aligned}$$

Por lo que la desviación estándar estaría dada por

$$s = \sqrt{5,762.820,513} = 75,913.24$$

Medidas de Dispersión

Coeficiente de Variación (C.V.)

El coeficiente de variación mide la dispersión relativa de un conjunto de valores al dividir la desviación estándar entre la media.

El *coeficiente de variación poblacional* (C.V.) y el *coeficiente de variación muestral* (c.v.) están dados respectivamente por:

$$\text{C.V.} = \frac{\sigma}{\mu}, \qquad \text{c.v.} = \frac{s}{\bar{x}}$$

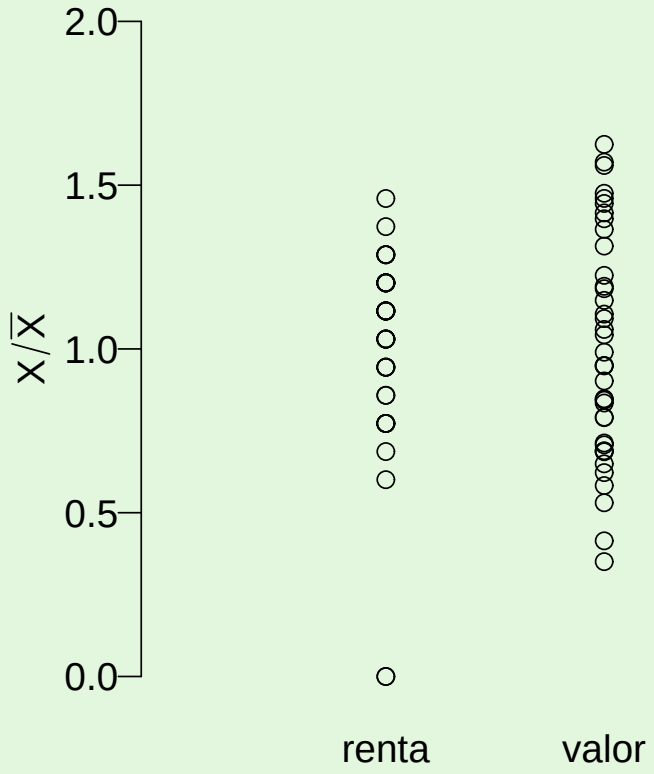
Nota: El coeficiente de variación permite expresar la desviación estándar como proporción de la media y es independiente de las unidades. Esto permite comparar la variabilidad de dos conjuntos de datos.

Medidas de Dispersión

Coeficiente de Variación (C.V.)

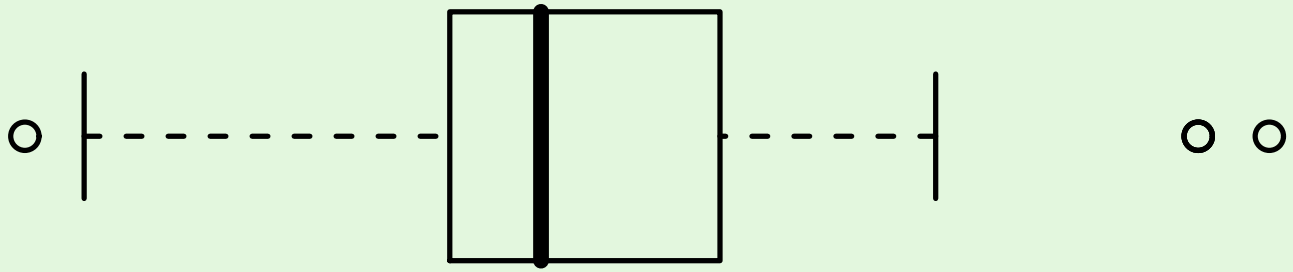
Estadísticos Descriptivos

	renta	valor
Min.	0.00	79.93
1st Qu.	50.00	162.20
Median	62.50	216.40
Mean	58.25	228.00
3rd Qu.	70.00	284.30
Max.	85.00	370.30
var	316.09	5973.79
sdev	17.78	77.29
cv	0.31	0.34

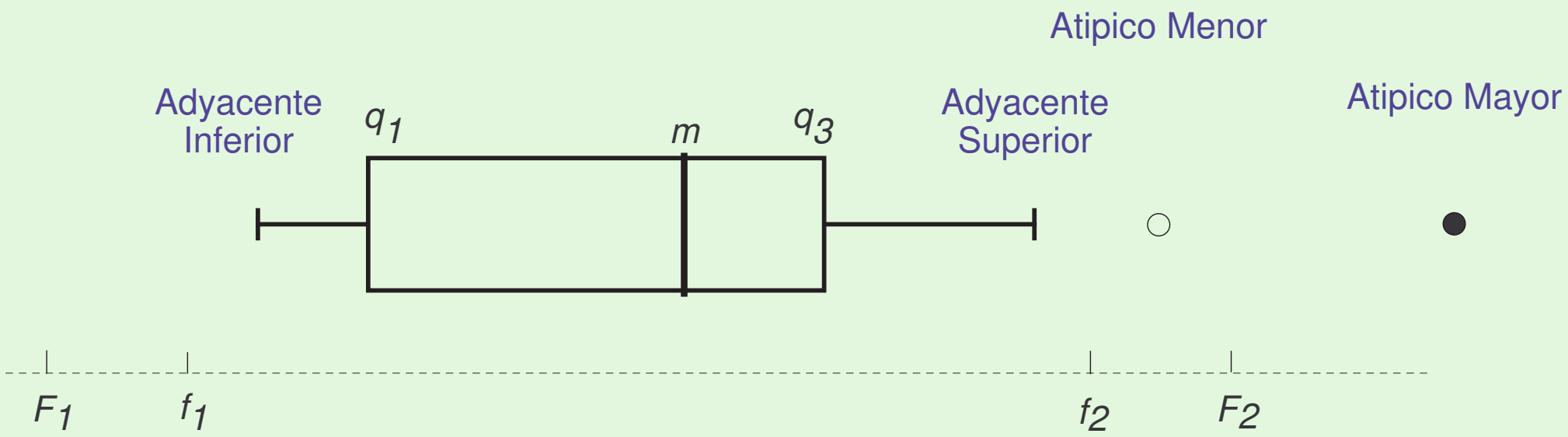


Diagramas de Caja y Brazos

Los *diagramas de cajas y brazos* se emplean para analizar y presentar las características más importantes de un conjunto de observaciones como son localización, dispersión, simetría y observaciones atípicas. Además resultan útiles en la comparación de dos o más conjunto de datos.



Diagramas de Caja y Brazos



Factor de Escala: *fes* $= 1.5 *$ A.I.

Barreras Interiores:	$f_1 = q_1 - fes$	$f_2 = q_3 + fes$
Barreras Exteriores:	$F_1 = f_1 - fes$	$F_2 = f_2 + fes$
	$= q_1 - 2fes$	$= q_2 + 2fes$

Adyacente Inferior: Observación más pequeña superior a f_1 y menor a q_1

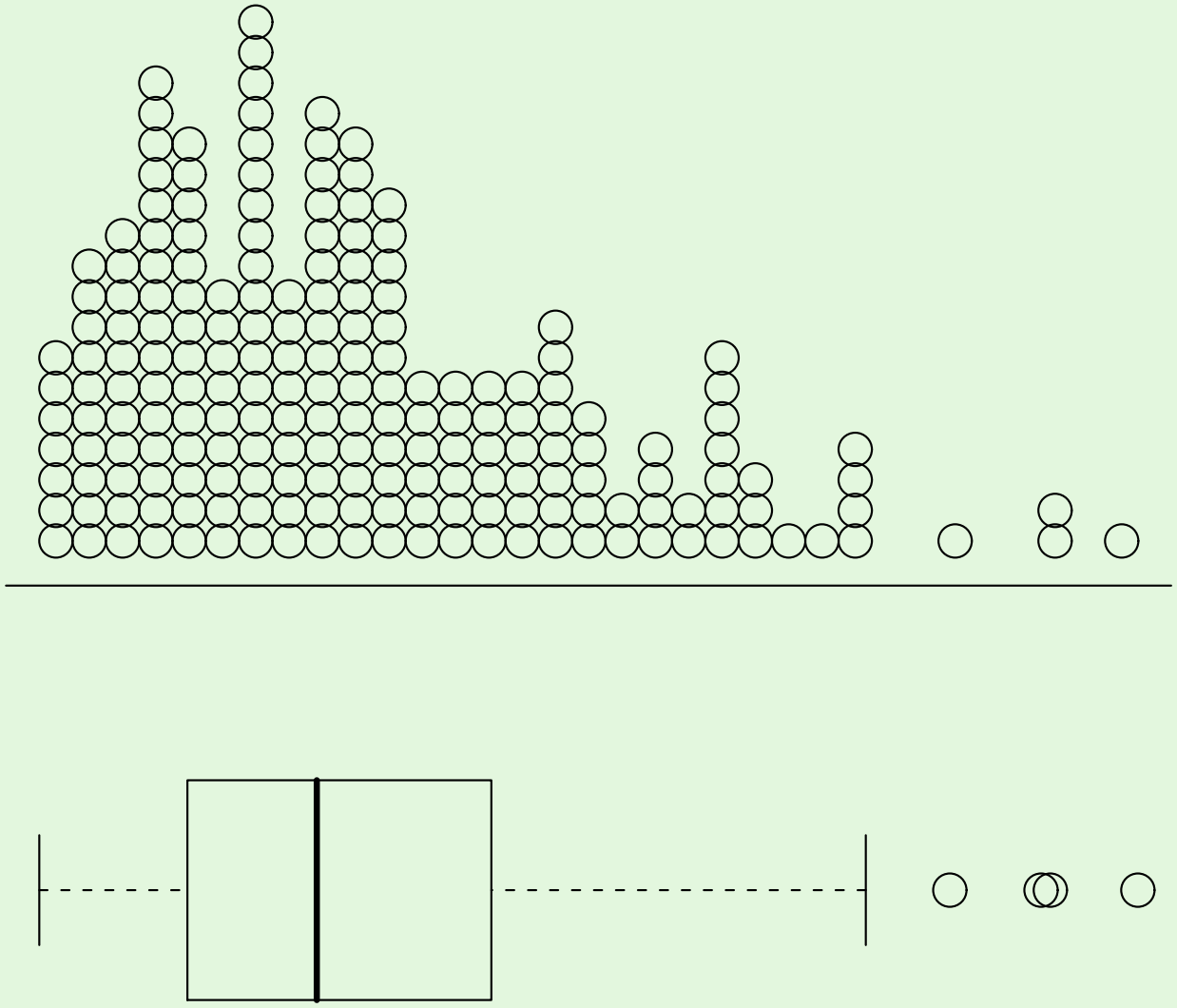
Adyacente Superior: Observación más grande inferior a f_2 y mayor a q_3

Atípicos Menores: Aquellos datos entre f y F

Atípicos Mayores: Aquellos datos más allá de F

Diagramas de Caja y Brazos

Ejemplo



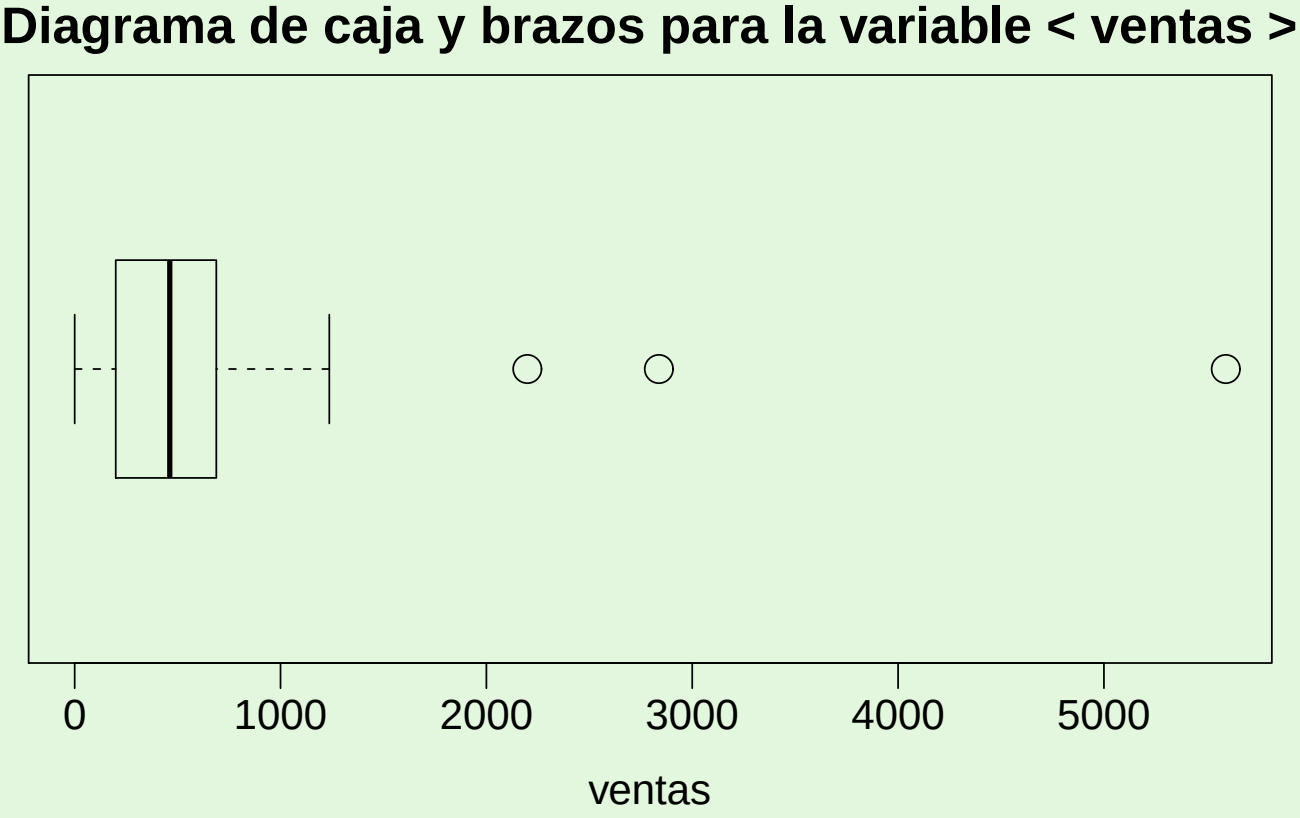
Diagramas de Caja y Brazos

Para el ejemplo de la venta de suavizantes:

Diagrama de Tallo y Hojas

```
ORDENADO

 0 | 00, 00, 47, 80
 1 | 13, 59, 75, 83
 2 | 16, 65
 3 | 31, 40
 4 | 13, 22, 50, 62, 65
 5 | 15, 48, 61, 70, 90
 6 | 73
 7 | 03, 46
 8 | 79
 9 |
10 | 83
11 |
12 | 37
   |
21 | 19
   |
28 | 38
   |
55 | 92
```

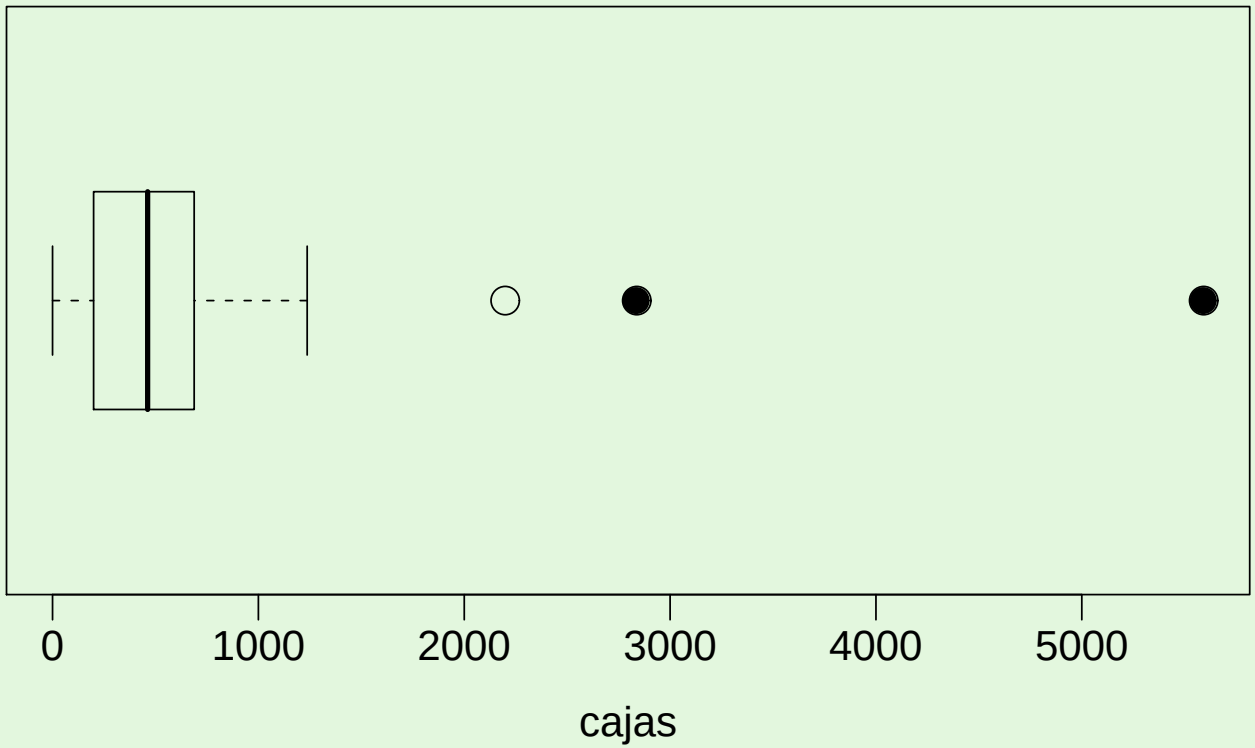


Diagramas de Caja y Brazos

Para el ejemplo de la venta de suavizantes:

$$
\begin{aligned}
\ell_1 &= (.25)31 + .5 = 7.75 + .5 = 8.25 \\
[\![\ell_1]\!] &= 8 \\
q_1 &= x_{(8)} = 183 \\
\ell_2 &= (.5)31 + .5 = 15.5 + .5 = 16 \\
[\![\ell_2]\!] &= 16 \\
\tilde{x} &= x_{(16)} = 462 \\
\ell_3 &= (.75)31 + .5 = 23.25 + .5 = 23.75 \\
[\![\ell_3]\!] &= 24 \\
q_3 &= x_{(24)} = 703
\end{aligned}
$$

$$
\begin{aligned}
R.I. &= 703 - 183 = 520 \\
\text{fes} &= 1.5(520) = 780 \\
f_1 &= 183 - 780 \to 0 \\
f_2 &= 703 + 780 = 1483 \\
F_1 &= 183 - 2(780) \to 0 \\
F_2 &= 703 + 2(780) = 2263
\end{aligned}
$$



Diagramas de Caja y Brazos

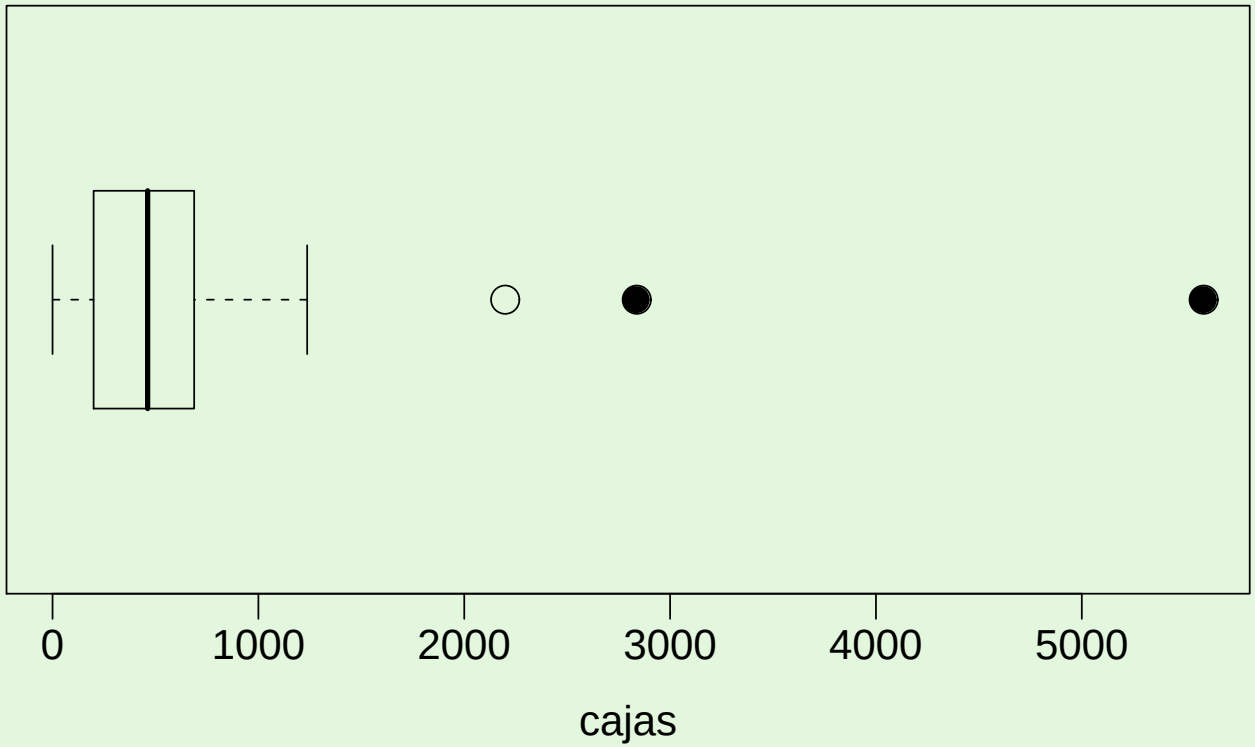
Para el ejemplo de la venta de suavizantes[1]:

$$q_1 = 199.5, \quad q_3 = 688.0, \quad \text{A.I.} = 688.0 - 199.5 = 488.5$$

$$m = 462, \quad \textit{fes} = 1.5(488.5) = 732.75$$

$$f_1 = 199.5 - 732.75 = -533.25 \qquad f_2 = 688 + 732.75 = 1420.75$$

$$F_1 = -533.25 - 732.75 = -1266.0 \quad F_2 = 1420.75 + 723.75 = 2153.50$$

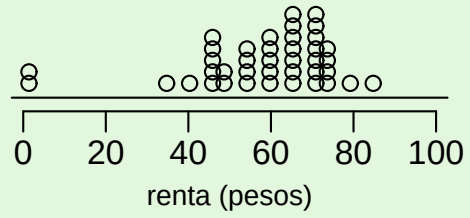


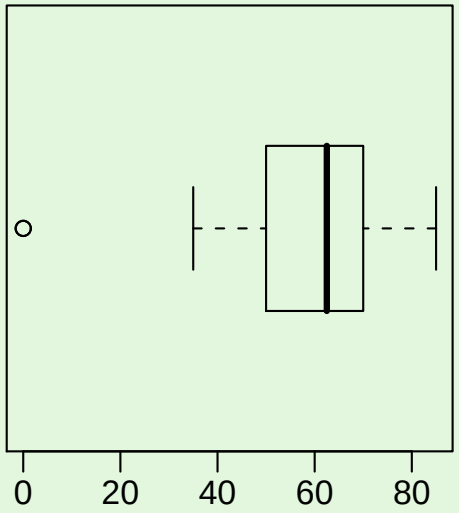
[1] Cálculos obtenidos y gráfica generada con R.

Diagramas de Caja y Brazos

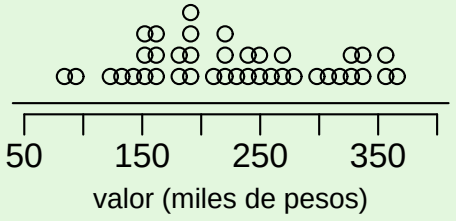
TV por Cable

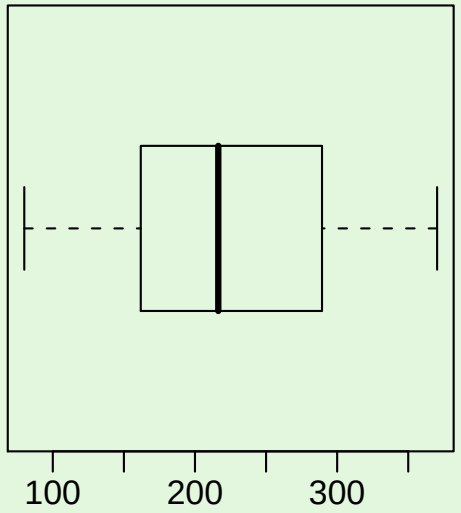
Renta





Valor





Problema de Comparación

Entre los temas más importantes de la Estadística están los *problemas de comparación* y los *problemas de asociación*.

El *problema de comparación* consiste en contrastar las distribuciones de frecuencia entre dos o mas subpoblaciones (grupos) basándose en los datos de muestras.

Por ejemplo, estudiando el problema del tabaquismo, definimos la variable cualitativa *habito del fumar* con las siguientes clases: *nunca ha fumado*, *dejó de fumar* y *fuma actualmente*. Deseamos comparar los grupos (subpoblaciones) *hombres* y *mujeres*.

Problema de Comparación

Subpoblaciones

Una manera de generar subpoblaciones es usando una variable *cualitativa nominal* para definirlas, e. g., *género*. Si la variable cualitativa empleada para la definición es *ordinal* entonces el problema puede verse como uno de *asociación*.

Otro ejemplo, de la industria de manufactura, sería comparar la dureza del acero entre proveedores nacionales y extranjeros. En este caso, la variable de interés es la *dureza* y las subpoblaciones serían LSA, USSTEEL, ACERIE-FRANÇAISE.

En ambos ejemplos se requiere responder las siguientes preguntas:

1. ¿Hay alguna diferencia en las distribuciones poblacionales?
2. ¿Cuál es la naturaleza de esas diferencias?
3. ¿Qué tan grande son esas diferencias?

Problema de Comparación

Subpoblaciones

Nótese que si bien las preguntas son planteadas en términos de las distribuciones de frecuencia poblacionales, en la práctica éstas se responden con base a *muestras* de dichas poblaciones.

Emplearemos elementos de la *Estadística Descriptiva* para responder estas preguntas. Para un análisis confirmatorio mas formal necesitamos de la *Estadística Inferencial*.

Problema de Comparación

Variable Cualitativa

Tablas de Contingencia

Cuando la variable es *cualitativa* es posible la comparación de distribuciones de frecuencia entre subpoblaciones empleando arreglos tabulares bidimensionales, llamados *tablas de contingencia* o *tabulación cruzada*.

La tabla muestra frecuencias absolutas por grupo y subpoblación. Por ejemplo,

Tabla de contingencia, encuesta estudiantil (frecuencias absolutas).

	Hábito de Fumar			
Género	Nunca ha fumado	Dejó de fumar	Fuma actualmente	Total
Masculino	154	25	185	364
Femenino	127	11	38	176
Total	281	36	223	540

Se puede ver en la tabla anterior que entre los hombres el grupo más numeroso es el de aquellos que fuman actualmente, siendo pocos los exfumadores. Esta distribución es distinta a la de las mujeres donde la mayoría de las encuestadas nunca han fumado. Este análisis puede hacerse más fácilmente si en la tabla presentamos frecuencias relativas.

Problema de Comparación

Variable Cualitativa

Frecuencias marginales

Dividiendo las celdas de la tabla anterior entre el total de la muestra (540):

Tabla de contingencia, encuesta estudiantil (frecuencias relativas %).

	Hábito de Fumar			
Género	Nunca ha fumado	Dejó de fumar	Fuma actualmente	Total
Masculino	28.5	4.6	34.3	67.4
Femenino	23.5	2.1	7.0	32.6
Frecuencias	52.0	6.7	41.3	100.0

De la tabla se puede ver que los hombres que fuman son el grupo más frecuente mientras que los casos de las mujeres han dejado de fumar son los menos frecuentes. Las *frecuencias marginales* (están en los márgenes de la tabla) nos muestran la frecuencia del atributo en la población en general.

Problema de Comparación

Variable Cualitativa

Frecuencias condicionales

Tabla de contingencia, encuesta estudiantil
Frecuencias relativas (%) condicionales por género.

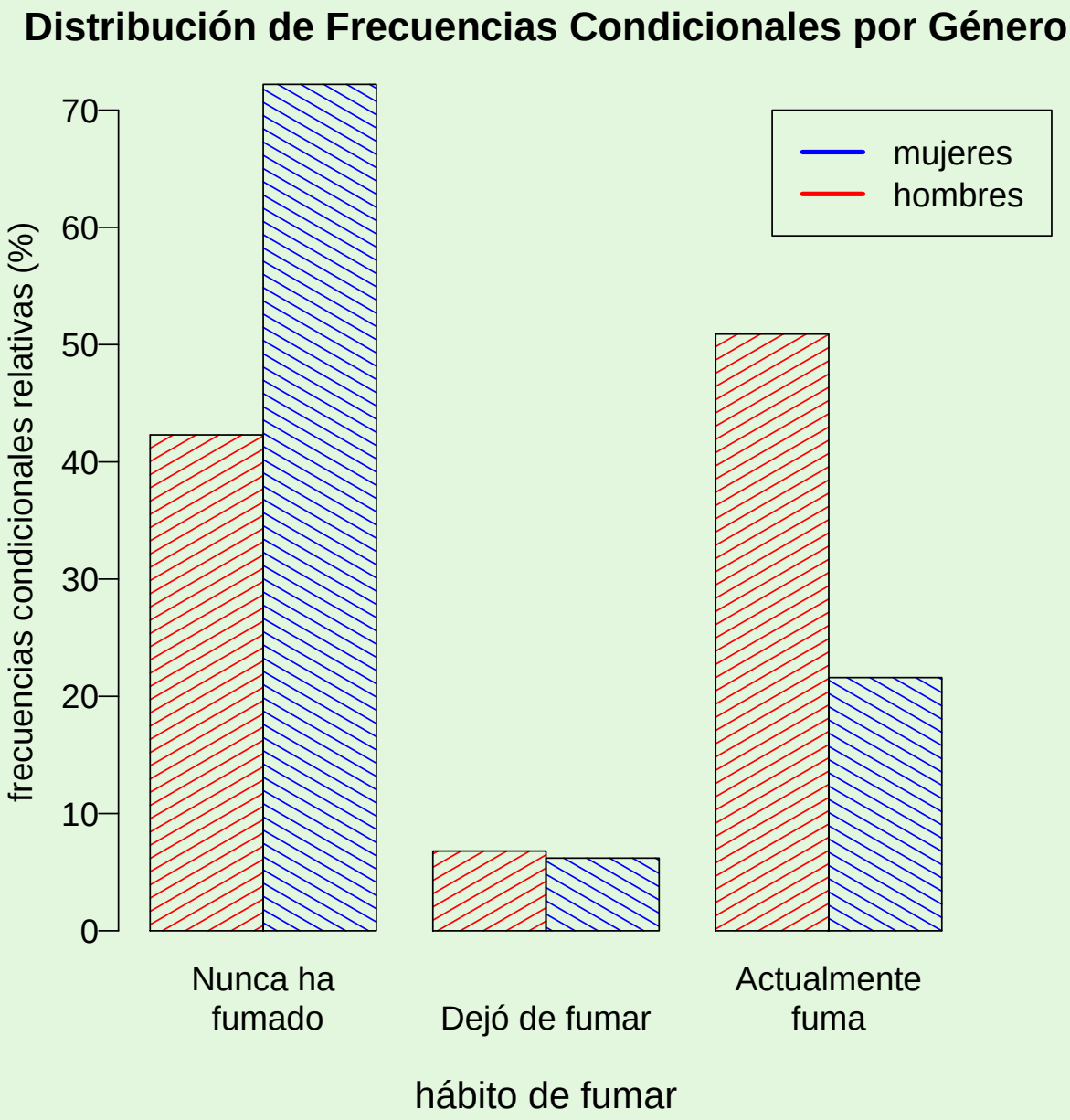
	Hábito de Fumar			
Género	Nunca ha fumado	Dejó de fumar	Fuma actualmente	Total
Masculino	42.3	6.8	50.9	100.0
Femenino	72.2	6.2	21.6	100.0
Frecuencias	52.0	6.7	41.3	100.0

De la tabla anterior se puede ver que aproximadamente 72 % de la población femenina nunca ha fumado; que la proporción de los que han dejado de fumar es más o menos la misma entre hombres y mujeres; y finalmente que más de la mitad de los estudiantes varones fuman actualmente.

Cuando hay muchas categorías presentes, una manera rápida de comparación es contrastar las frecuencias condicionales contra las frecuencias marginales correspondientes.

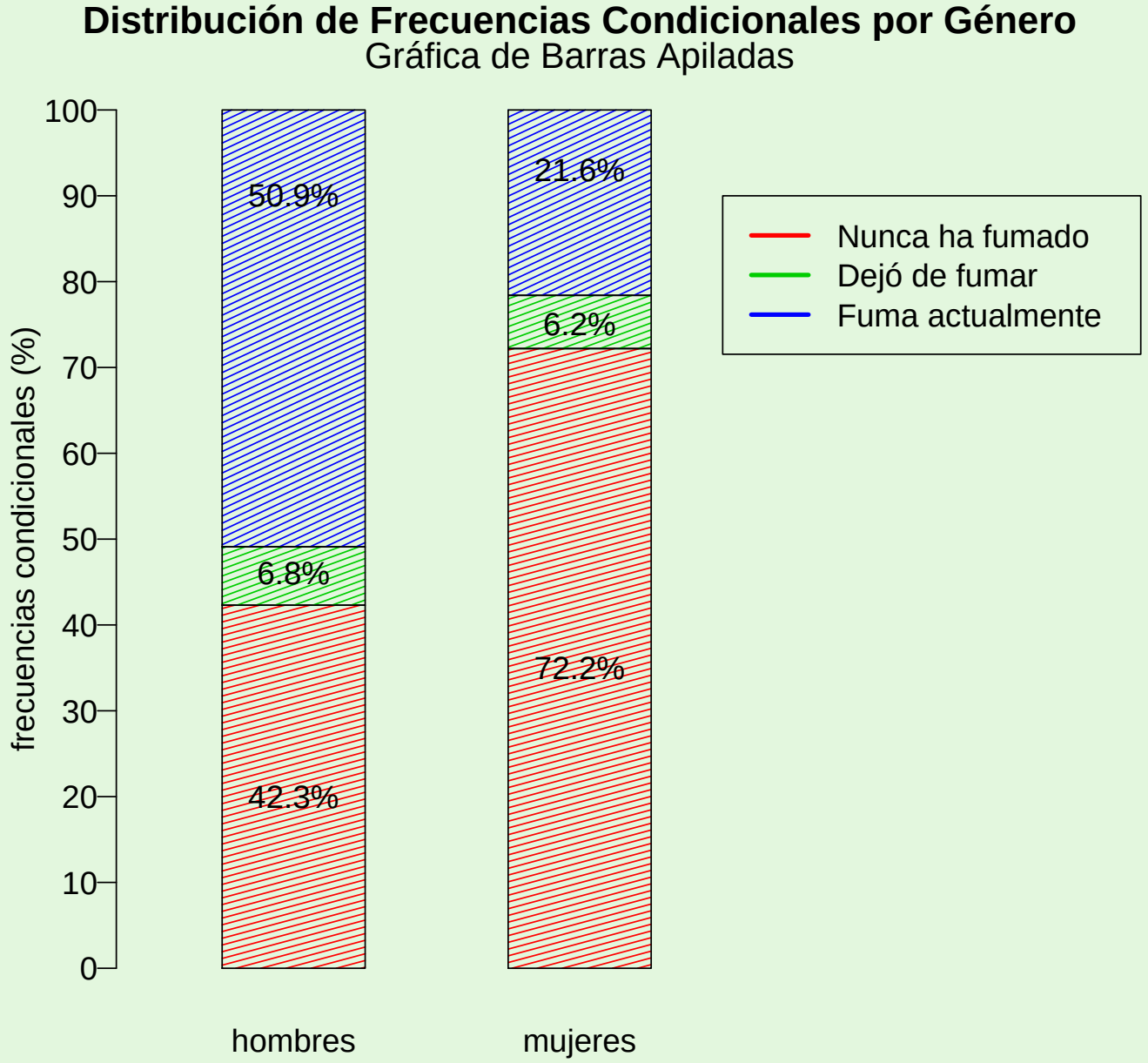
Problema de Comparación

Variable Cualitativa



Problema de Comparación

Variable Cualitativa



Problema de Comparación

Variable Cuantitativa Discreta

En este caso la comparación puede hacerse de la misma forma que se hizo con las variables cualitativas. E. g., encuesta de TV por cable:

Número de televisores por casa.

Colonia 1		Colonia 1	
Manzana	Televisores	Manzana	Televisores
9	4,3,4,3,5	14	0,1,1,4
2	3,3,2,4,3	22	1,3,4,3,2
4	2,3,3,3,2	8	2,2,2,3,1
		20	2,3,3,1,3
		25	2,0,3,1,1

Problema de Comparación

Variable Cuantitativa Discreta

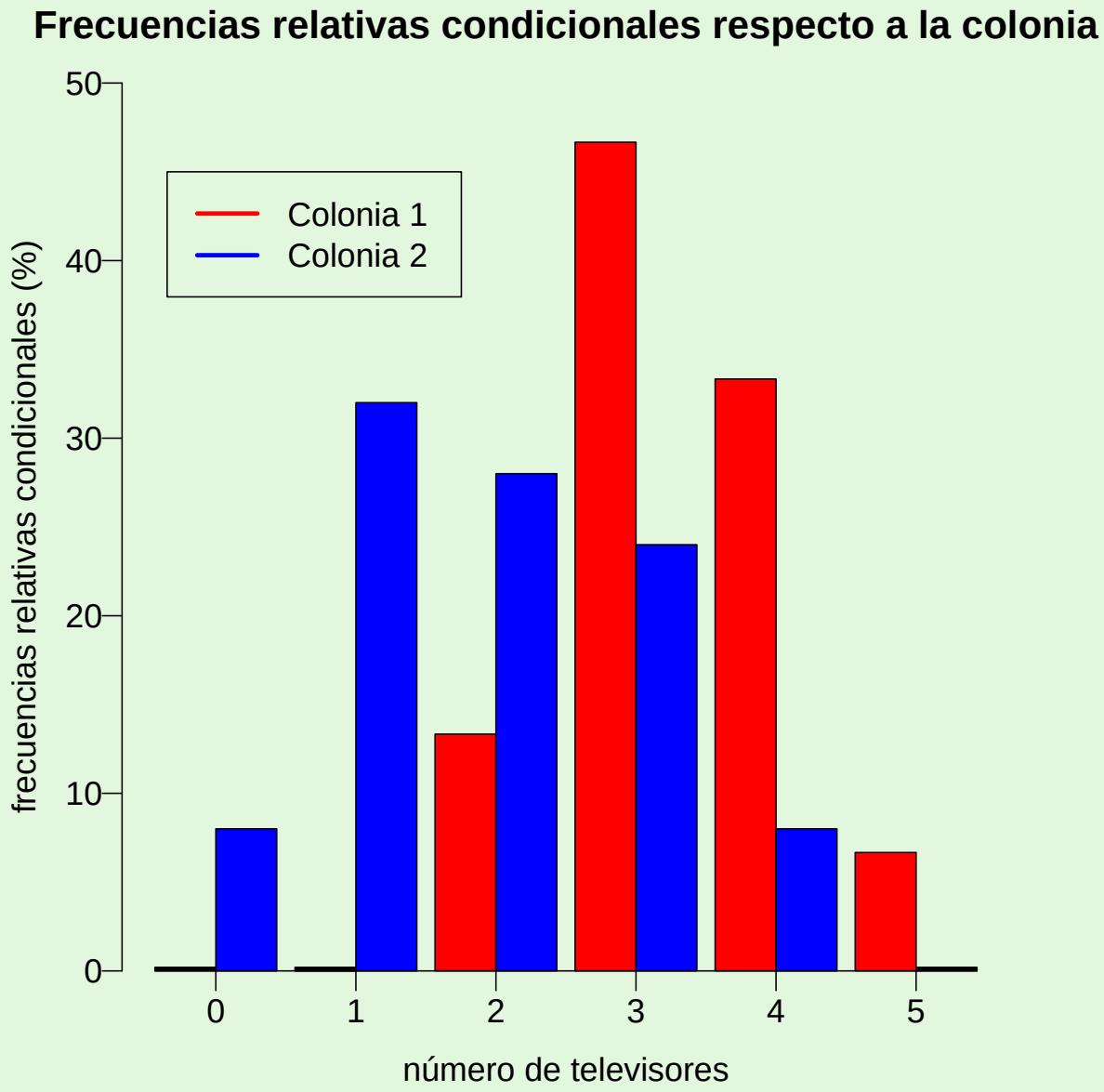
Comparación de la distribución del número de televisores por casa por entre colonias.

Tabulación cruzada	Número de televisores por casa						
	0	1	2	3	4	5	
Colonia 1	0	0	2	7	5	1	15
Colonia 2	2	8	7	6	2	0	25
Total	2	8	9	13	7	1	40
Frecuencias relativas condicionales	Número de televisores por casa (%)						
Colonia	0	1	2	3	4	5	
1	0	0	20	53	20	7	100
2	8	32	24	28	8	0	100

Al igual que con las variables cualitativas la información puede presentarse de manera gráfica.

Problema de Comparación

Variable Cuantitativa Discreta



Problema de Comparación

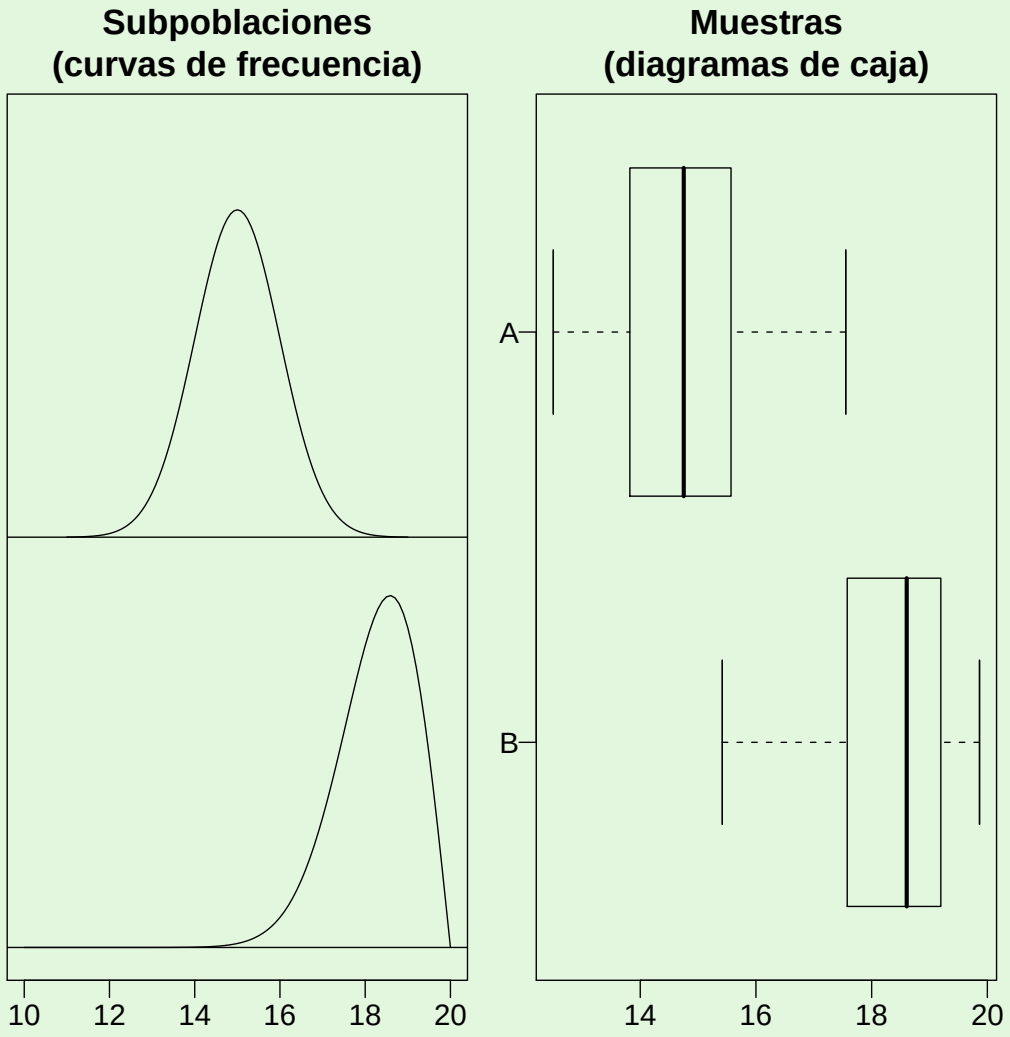
Variable Continua

En este caso estamos interesados en comparar tanto la tendencia central como la dispersión de las poblaciones.

El lado derecho muestra los diagramas de caja de muestras tomadas de las poblaciones correspondientes.

Las conclusiones obtenidas de las muestras se aplican también a las poblaciones:

- La población A es simétrica alrededor de 15 y a la izquierda de la población B.
- La población B es sesgada a la izquierda con mediana (centro) poco mayor que 18.



Problema de Comparación

Variable Continua

Ejemplo: Productores de acero

Los datos de la tabla provienen de ensayos de dureza de lámina de acero de tres proveedores de una empresa que produce manufacturas troqueladas. Una *característica de calidad* importante es la dureza de la materia prima. Los datos corresponden al primer semestre y las unidades son kg/cm^2.

LSA		USSTEEL		ACERIE FRANÇAISE	
52.4	47.9	54.4	48.8	48.8	42.7
50.8	50.1	50.2	47.9	49.8	52.7
45.5	52.2	49.4	47.5	43.2	51.6
44.4	41.2	57.0	49.2	45.7	51.2
45.2	51.9	55.5	49.0	48.1	39.8
46.2	50.8	54.9	47.6	48.9	39.1
46.2	45.4	49.9	47.9	49.1	51.1
46.2	47.9	48.7	51.7	46.7	41.1
52.5		53.0	47.3	38.9	51.3
46.7		50.9	50.7	43.2	

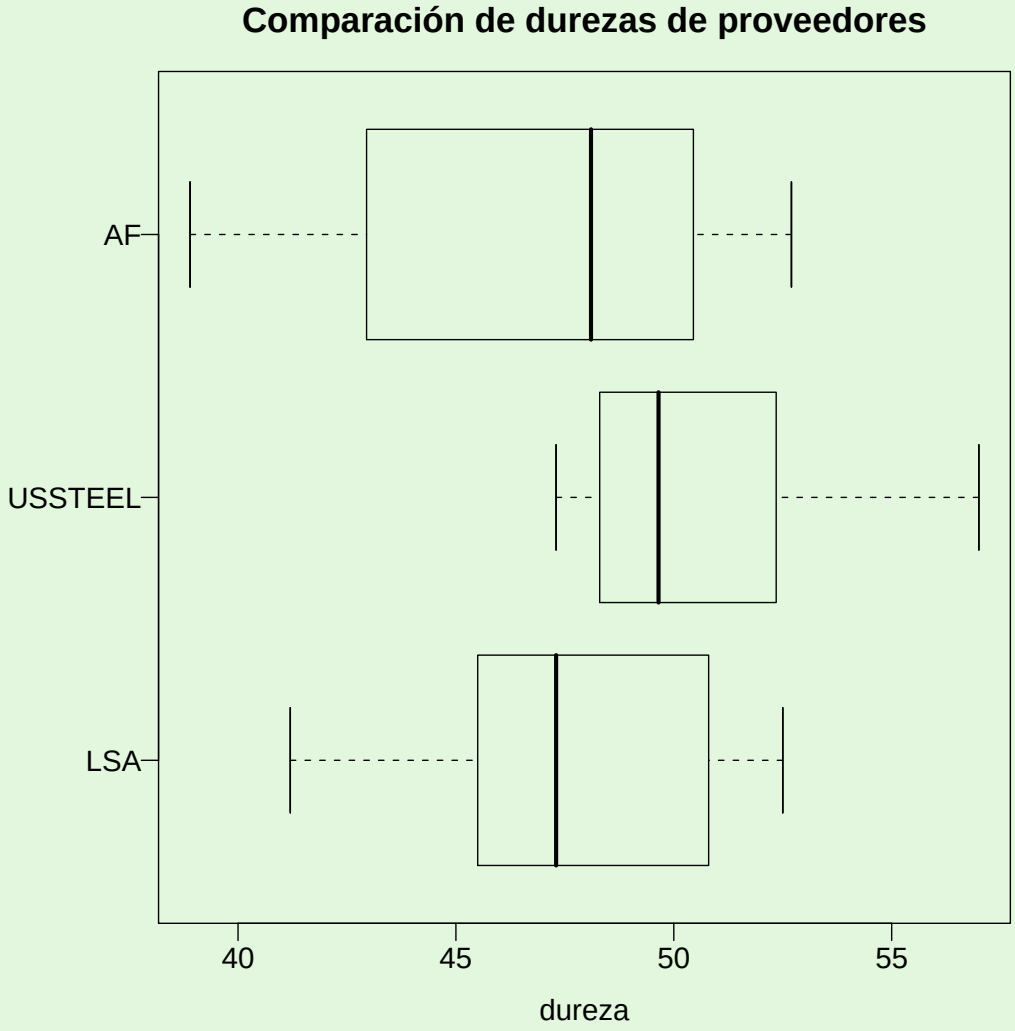
Problema de Comparación

Variable Continua

Ejemplo: Productores de acero

Del diagrama de caja se sigue lo siguiente:

- El productor USSTEEL provee de lámina de mayor dureza y más consistentemente (menor variabilidad/dispersión) que LSA y ACERIE-FRANÇAISE.
- La producción de USSTEEL parece estar sesgada a la derecha mientras que las otras dos compañías parecen más bien sesgadas a las izquierda.



Problema de Asociación

En ocasiones es importante conocer si una variable influye en el comportamiento (modo de variación) de otra variable. E. g., una cadena de establecimientos comerciales desea saber si el tamaño del establecimiento influye en el volumen de ventas.

Otro ejemplo sería aquel al estudiar el sector agrícola y qué tanto influye los insumos de trabajo o capital en la producción del ramo.

Ambos casos pueden caracterizarse como un problema de *asociación* en el cual nos interesa conocer si el incremento o decremento de una variable, X, tiene efecto o influye en otra variable, digamos Y. Note que por la naturaleza del problema, las variables consideradas X y Y, deben ser al menos de escala ordinal.

Problema de Asociación

Ambas Variables son Ordinales

Una manera de analizar el problema de asociación cuando ambas variables son ordinales es mediante el uso de *tablas de contingencia* y los correspondientes diagramas de barras.

E. g., consideremos una encuesta sobre el *horario de verano*, en el cual interesa relacionar la posición respecto al cambio de horario (Y) con el nivel socioeconómico del encuestado (X). Los valores (*niveles*) de Y son: *desacuerdo*, *indiferente* y *de acuerdo*, mientras que los de X: *bajo*, *medio* y *alto*.

Tabla de contingencia (frecuencias absolutas)

		Posición respecto al horario de verano			
		Desacuerdo	Indiferente	Acuerdo	Total
Nivel socio-económico	Bajo	98	201	111	410
	Medio	134	91	60	285
	Alto	12	21	25	58
	Total	**244**	**313**	**196**	**753**

Tabla de contingencia (frecuencias relativas condicionales)

		Posición respecto al horario de verano (%)			
		Desacuerdo	Indiferente	Acuerdo	Total
Nivel socio-económico	Bajo	24	49	27	100
	Medio	47	32	21	100
	Alto	21	36	43	100

Problema de Asociación

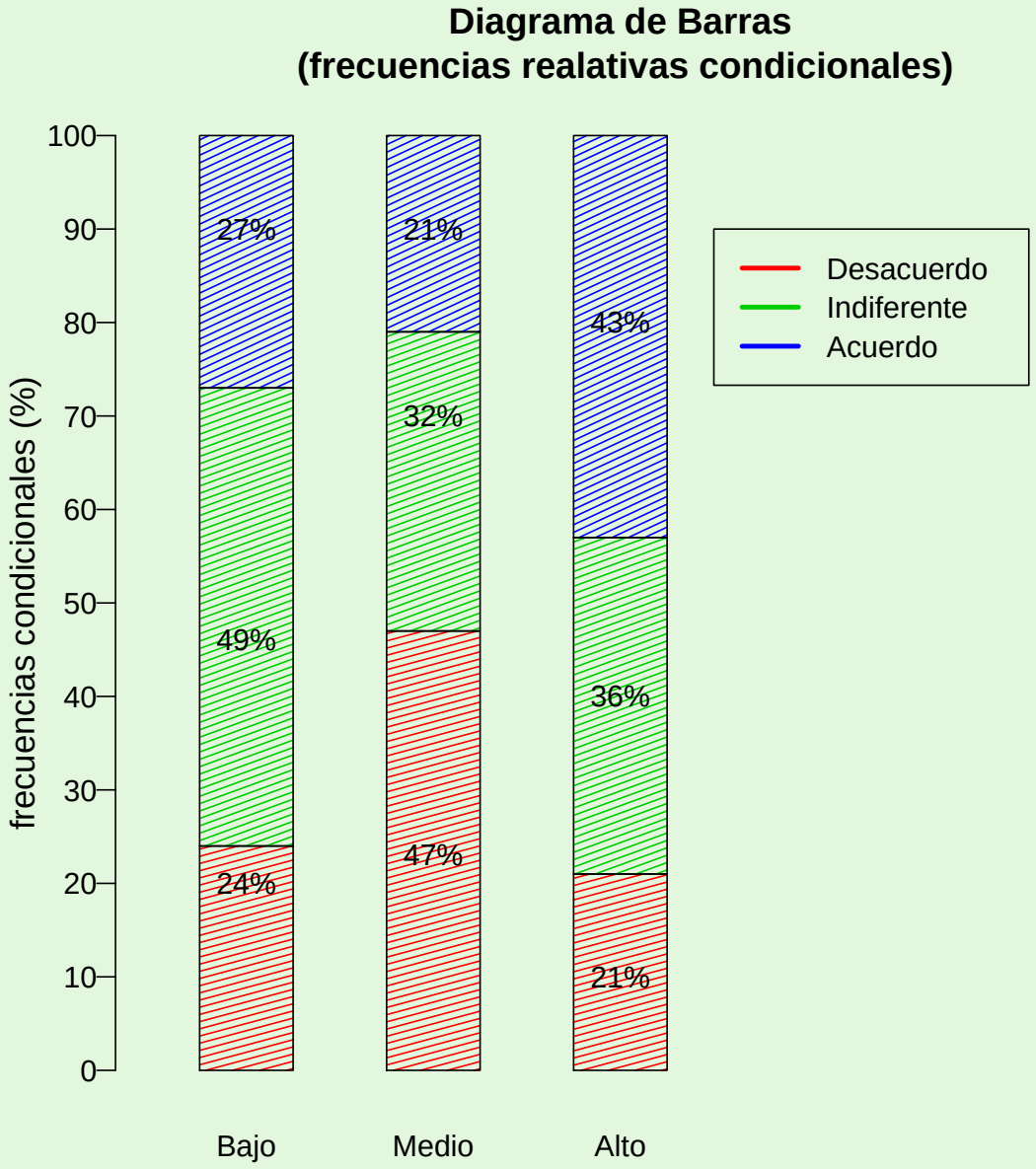
Ambas Variables son Ordinales

Encuesta sobre Horario de Verano

Tabla de contingencia
(frecuencias relativas condicionales)

Posición respecto al horario de verano (%)

	Desacuerdo	Indiferente	Acuerdo	Total
Bajo	24	49	27	100
Medio	47	32	21	100
Alto	21	36	43	100



Problema de Asociación

Una variable ordinal y la otra cuantitativa

En este caso es posible visualizar ambos, *localización* (tendencia central) y la *variación* (dispersión) de la variable cuantitativa de acuerdo a los distintos niveles de la variable ordinal.

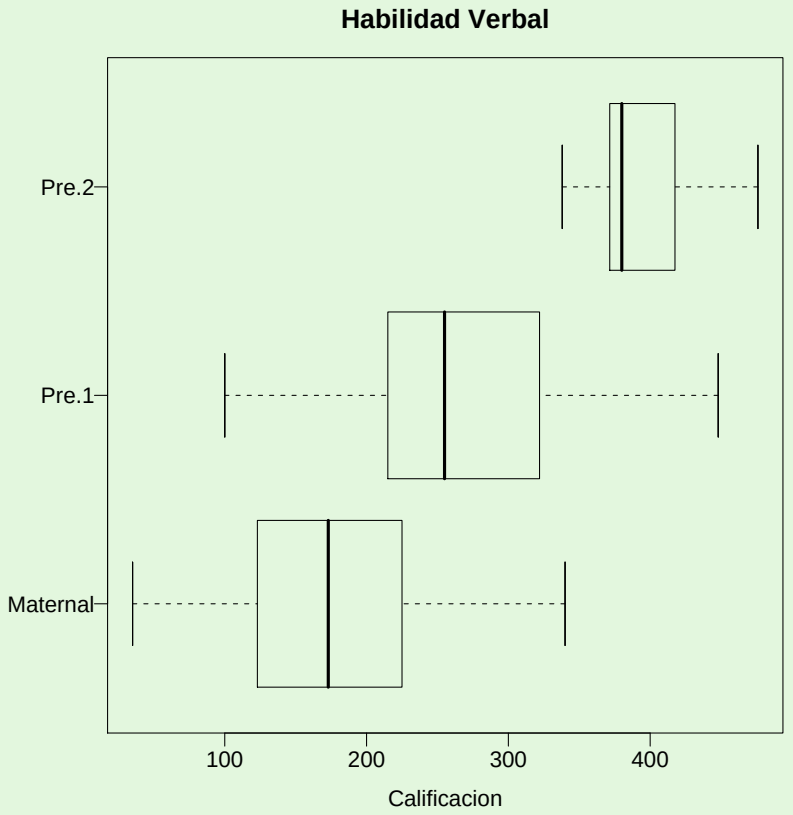
Problema de Asociación

Una variable ordinal y la otra cuantitativa

Habilidad Verbal en Pre-escolar

La siguiente tabla corresponde a una prueba de habilidad verbal para una muestra de un jardín de niños. La variable Y es la evaluación de desarrollo de la habilidad verbal, y la variable X, es el grado escolar del niño.

Grado Escolar		
Maternal	Pre-escolar 1	Pre-escolar 2
68	255	425
35	202	370
145	317	380
173	327	476
190	247	410
225	100	358
340	448	338
123	412	373
228	228	377
NA	192	467
NA	297	388



Problema de Asociación

Ambas Variables Cuantitativas

En esta situación el *diagrama de dispersión* es una herramienta gráfica de gran utilidad. Consiste en representar cada pareja de la muestra $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ sobre el plano cartesiano $X-Y$.

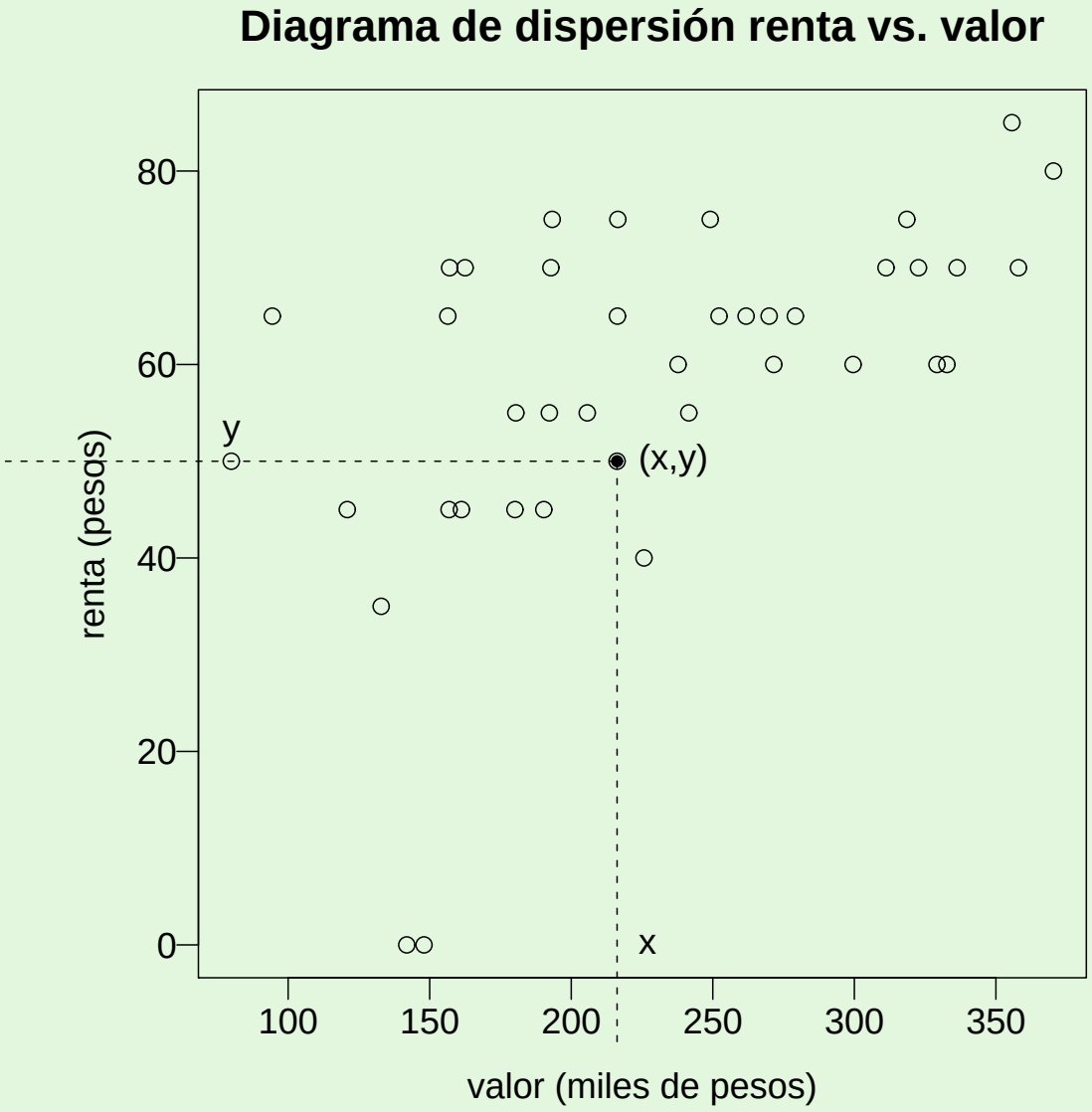
Construcción:

1. Sobre un par de ejes cartesianos seleccionar una escala en el eje X (correspondiente a una de las variables) y otra en el eje Y (para la otra variable) de modo de que quepan todos los valores observados.
2. Graficar cada pareja (x_i, y_i) en el punto del plano que le corresponda. Si hay puntos repetidos, represéntelos como puntos concéntricos.

Problema de Asociación

Ambas Variables Cuantitativas

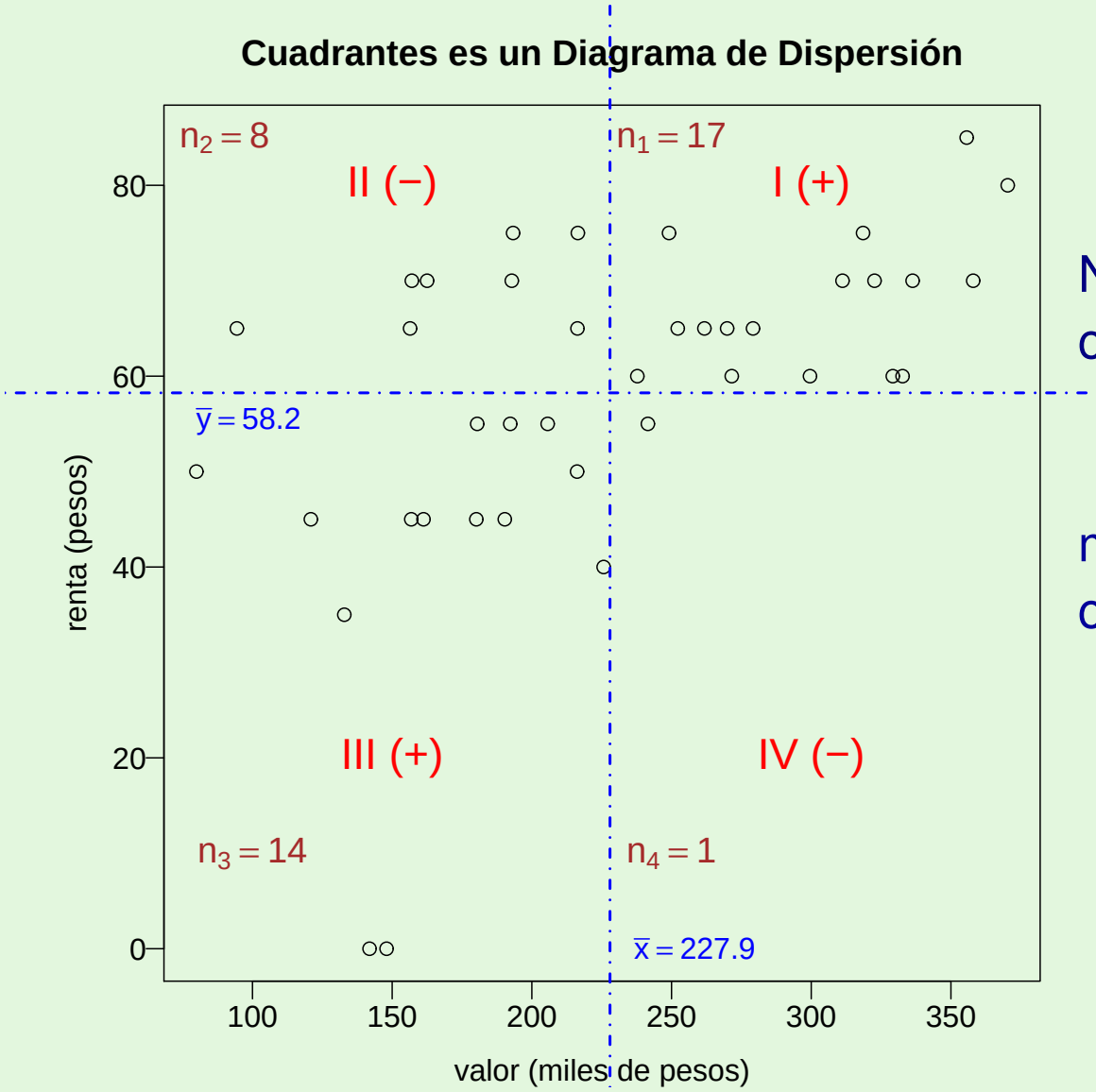
Encuesta de TV por Cable



Problema de Asociación

Ambas Variables Cuantitativas

Encuesta de TV por Cable



Note que para el primer y tercer cuadrante,

$$n_1 + n_3 = 17 + 14 = 31$$

mientras que para el segundo y cuarto cuadrante

$$n_2 + n_4 = 8 + 1 = 9$$

Problema de Asociación

Ambas Variables Cuantitativas

Además del análisis gráfico es interesante tener una medida de la asociación entre las dos variables.

Covarianza de dos variables cuantitativas X y Y:

$$\text{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}\right]$$

Note que las unidades de la *covarianza* son las unidades originales al cuadrado. Igualmente, si cambia de escala una o ambas variables la covarianza cambiará.

Problema de Asociación

Ambas Variables Cuantitativas

Para expresar la asociación de X y Y, independiente de las escalas se utiliza el *coeficiente de correlación*.

Correlación de dos variables cuantitativas X y Y:

$$r = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}} = \frac{\text{cov}(X, Y)}{S_X \cdot S_Y}$$

donde

$$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \text{y} \quad S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

son las desviaciones estándar de X y Y respectivamente.

Problema de Asociación

Ambas Variables Cuantitativas

Nota:

- Existen las correspondientes definiciones poblacionales:

  Covarianza

$$\text{cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

  Coeficiente de Correlación

$$\text{corr}(X,Y) = \rho = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}}$$

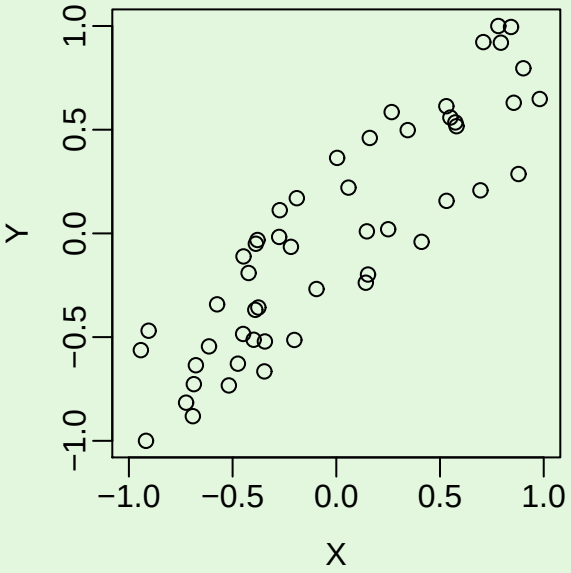
- El coeficiente de correlación no tiene unidades (adimensional).
- El coeficiente de correlación (poblacional o muestral) es siempre mayor o igual que -1 y menor o igual que $+1$. Esto es,

$$-1 \leq \rho \leq +1, \quad -1 \leq r \leq +1$$

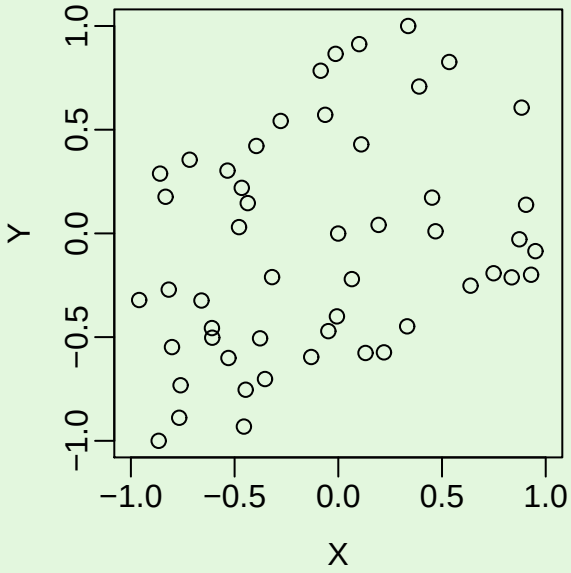
Problema de Asociación

Coeficiente de Correlación

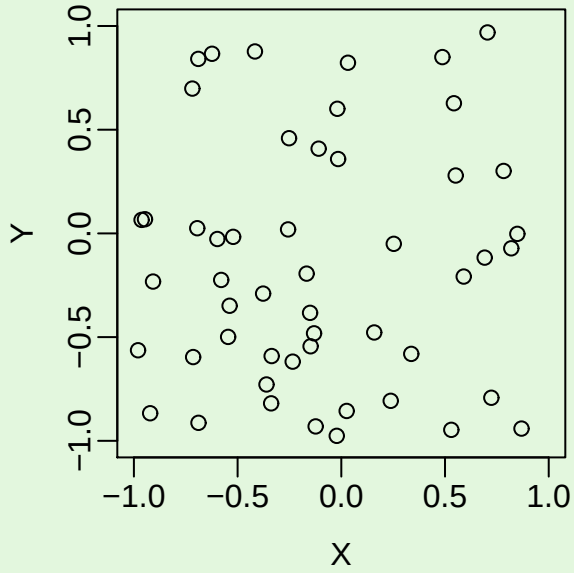
a) Fuerte asociacion positiva: r=0.89



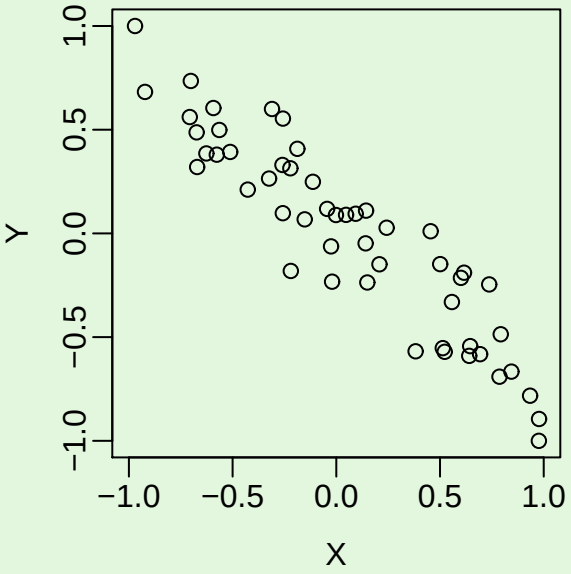
b) Asociacion positiva: r=0.31



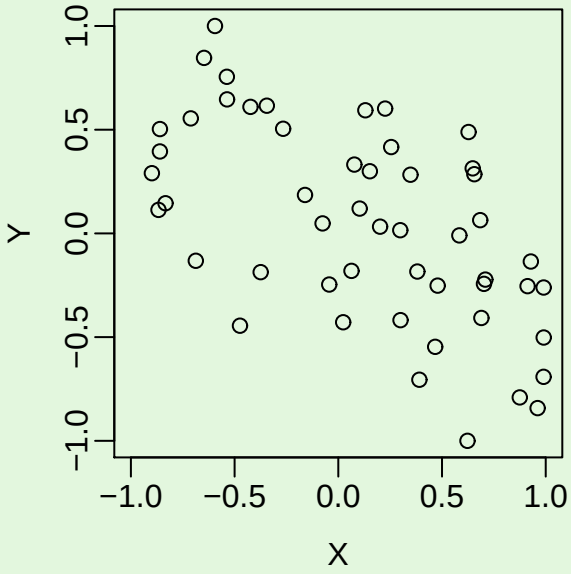
c) Sin asociacion: r=0.04



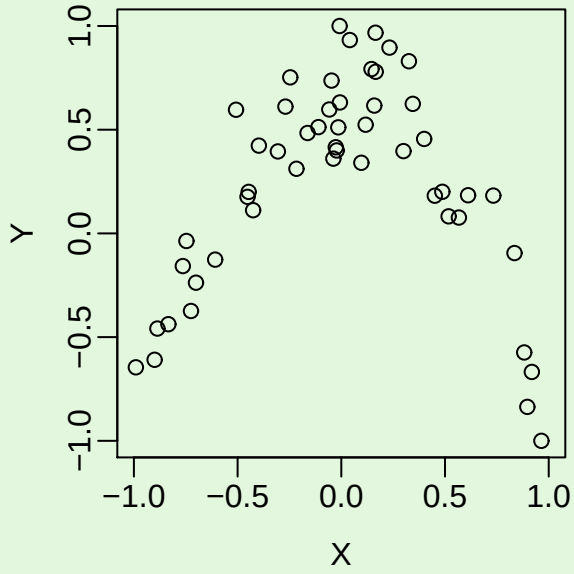
d) Fuerte asociacion negativa: r=−0.93



e) Asociacion negativa: r=−0.58



f) Asociacion no lineal: r=0.04

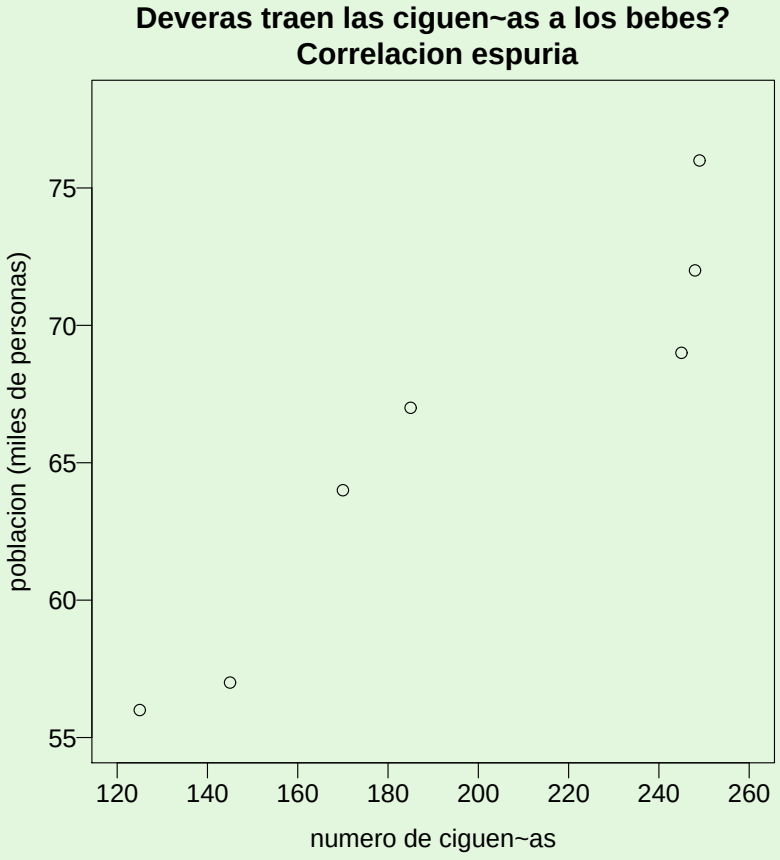


Problema de Asociación

Ambas Variables Cuantitativas

Notas:

- El valor de $|r|$ será más cercano a 1 conforme la nube de datos se acerque más a un linea recta. Por lo mismo, r es conocido también como *coeficiente de correlación lineal*.
- Correlación no implica *causalidad*. Véase la gráfica de la derecha. Datos de la población anual de una población inglesa (1930–1936) y el número de avistamientos de cigüeñas al año.
- El tipo de correlación mostrado entre población y cigüeñas se conoce como *correlación espuria*.
- Puede haber correlación de variables pero no necesariamente lineal. Véase, e. g., el diagrama de dispersión f) de la pasada lámina.



Referencias

1. Victor Aguirre y Begoña Artolia (2007). *Análisis Exploratorio de Datos*. Capítulo 1 de Aguirre et al. *Fundamentos de Probabilidad y Estadística*. Editorial Jit Press. Segunda Edición.

2. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

3. LaTeX Users Group. *http://www.tug.org*. TeX Users Group.