Capítulo 2: Explorando datos

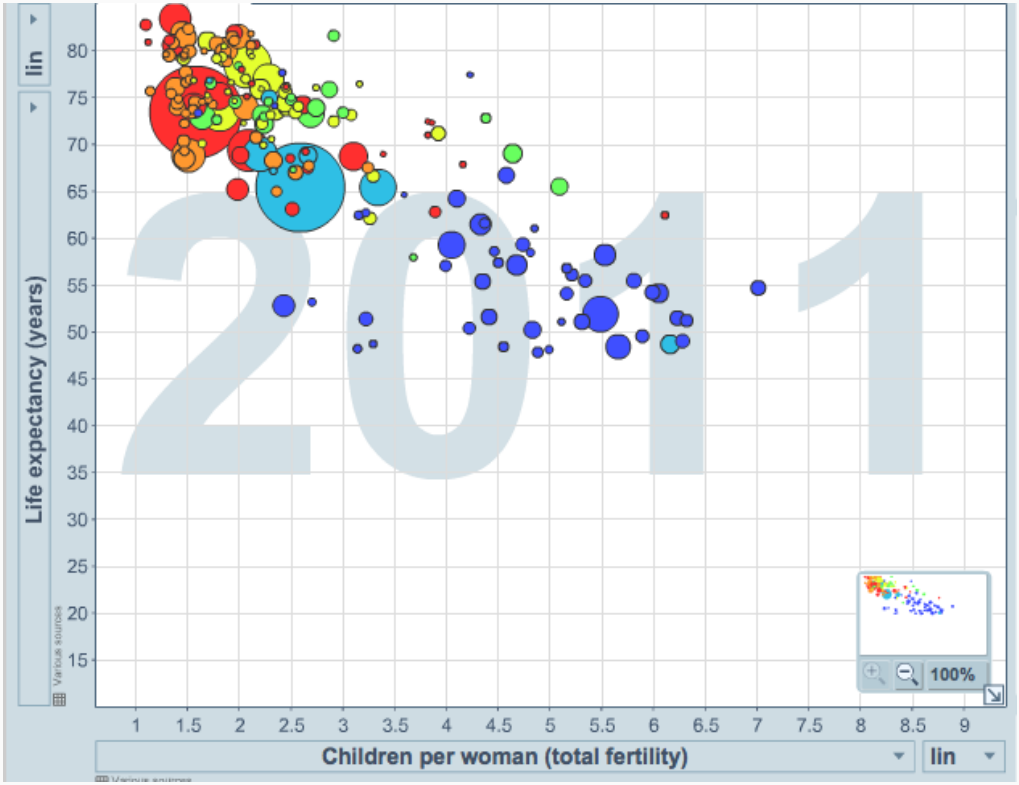
Metodología de análisis en Opinión Pública

Facultad de Ciencias Sociales
Universidad de Buenos Aires

Examinando datos numéricos

Diagrama de dispersión

Los diagramas de dispersión son útiles para visualizar la relación entre dos variables numéricas.



¿Parece que la esperanza de vida y la fertilidad total están asociadas o independientes?

Parecen estar asociados lineal y negativamente: a medida que aumenta la fertilidad, disminuye la esperanza de vida.

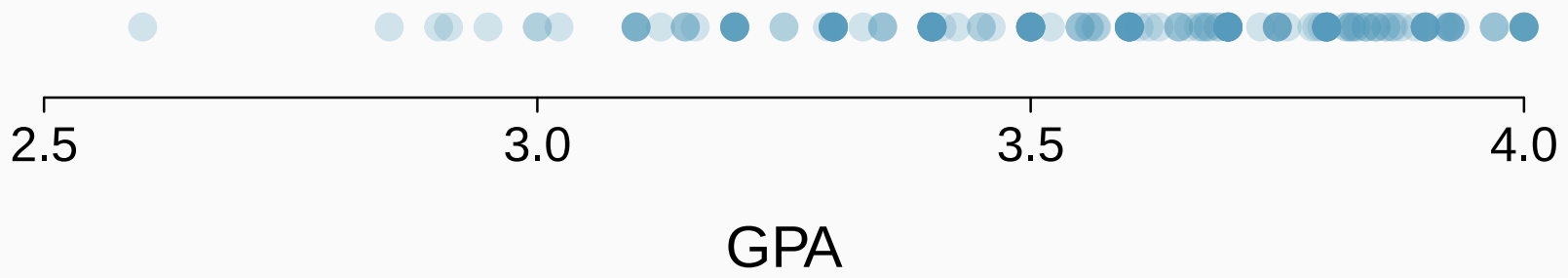
¿La relación fue la misma a lo largo de los años o cambió?

La relación cambió con los años.

http://www.gapminder.org/world

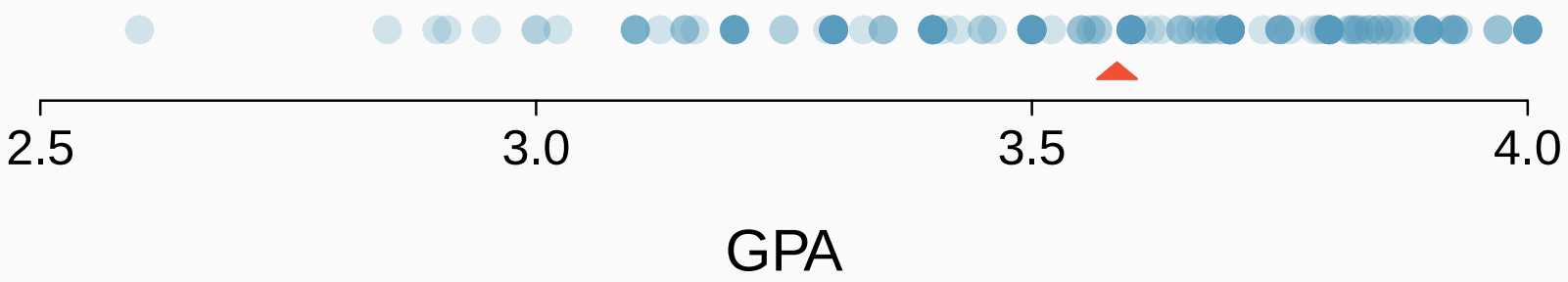
Gráficos de puntos

Útil para visualizar una variable numérica. Los colores más oscuros representan áreas donde hay más observaciones.



¿Cómo describiría la distribución de GPA en este conjunto de datos? Asegúrese de decir algo sobre el centro, la forma y la extensión de la distribución.

Diagramas de puntos & media



- La media, también llamada promedio (marcada con un triángulo en el gráfico anterior), es una forma de medir el centro de una distribución de datos.
- El GPA promedio es 3.59.

Media

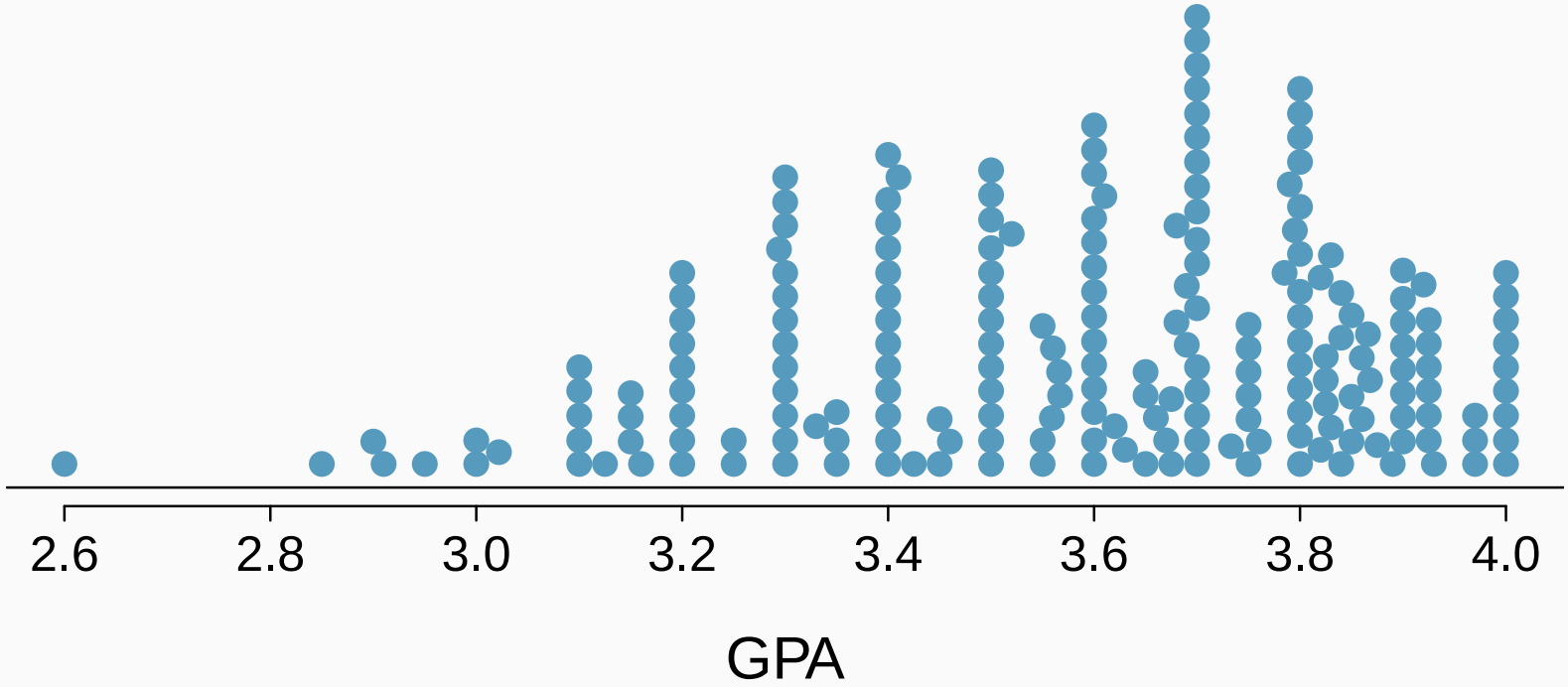
- La media muestral, denotada como $\bar{x}$, se puede calcular como

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

donde $x_1, x_2, \cdots, x_n$ representan los valores observados de n.
- La media de la población también se calcula de la misma manera pero se denota como μ. A menudo no es posible calcular μ ya que los datos de población rara vez están disponibles.
- La media muestral es una estadística muestral y sirve como estimación puntual de la media poblacional. Esta estimación puede no ser perfecta, pero si la muestra es buena (representativa de la población), por lo general es una estimación bastante buena.

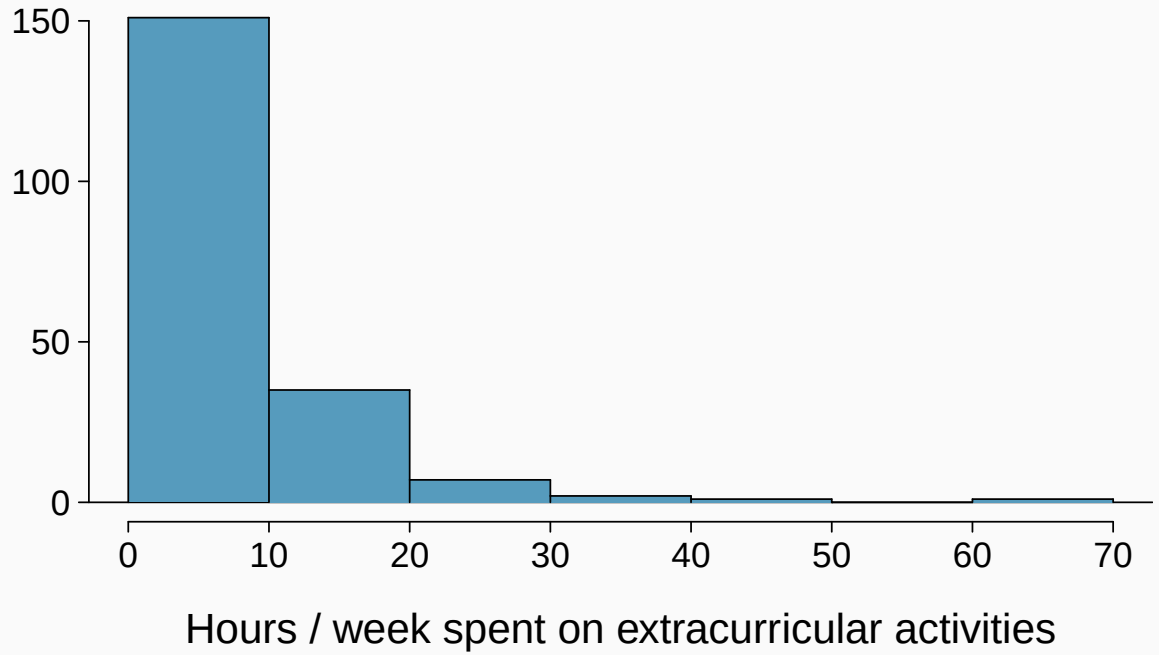
Diagrama de puntos apilados

Las barras más altas representan áreas donde hay más observaciones, lo que hace que sea un poco más fácil juzgar el centro y la forma de la distribución.



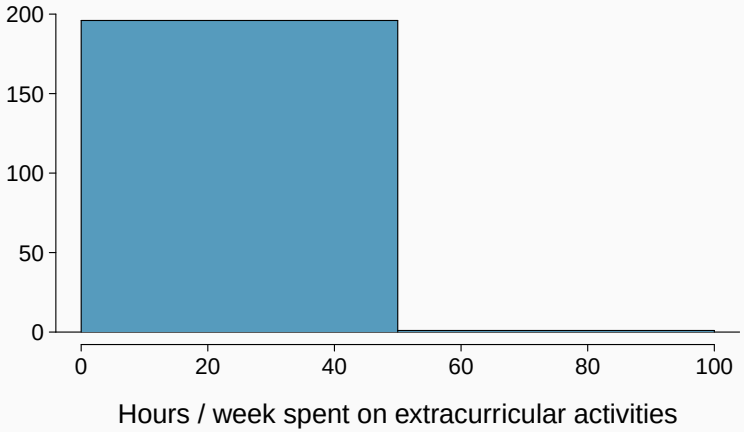
Histogramas - Horas extracurriculares

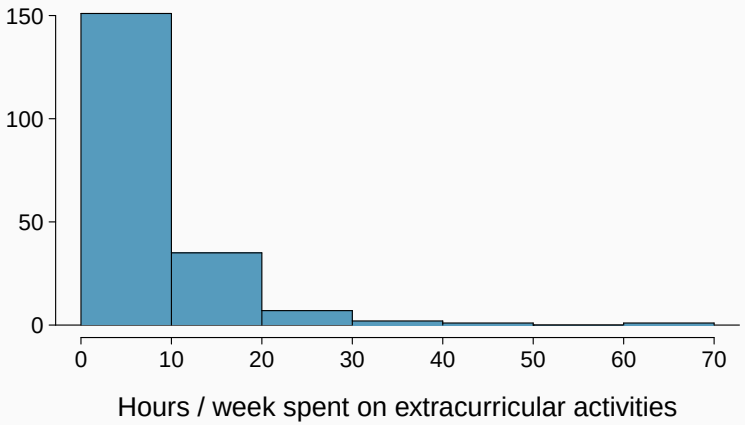
- Los histogramas proporcionan una vista de la densidad de datos. Las barras más altas representan dónde los datos son relativamente más comunes.
- Los histogramas son especialmente convenientes para describir la forma de la distribución de datos.
- El bin width elegido puede alterar la historia que cuenta el histograma.

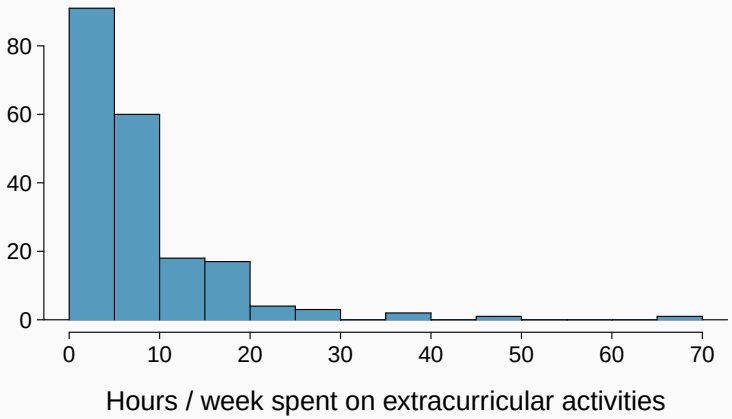


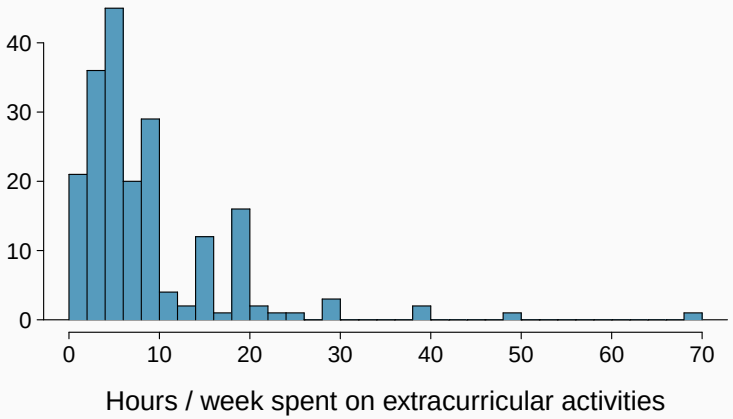
Ancho del contenedor

¿Cuál(es) de estos histogramas son útiles? ¿Cuáles revelan demasiado sobre los datos? ¿Cuáles esconden demasiado?



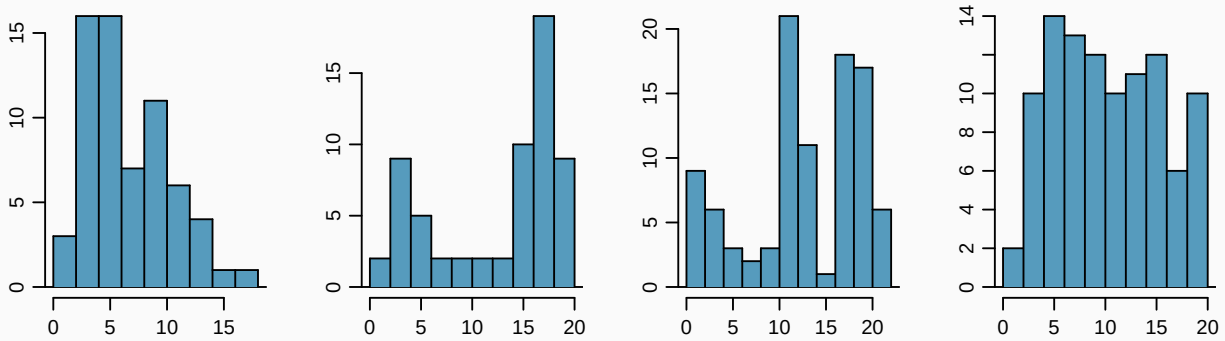






Forma de una distribución: modalidad

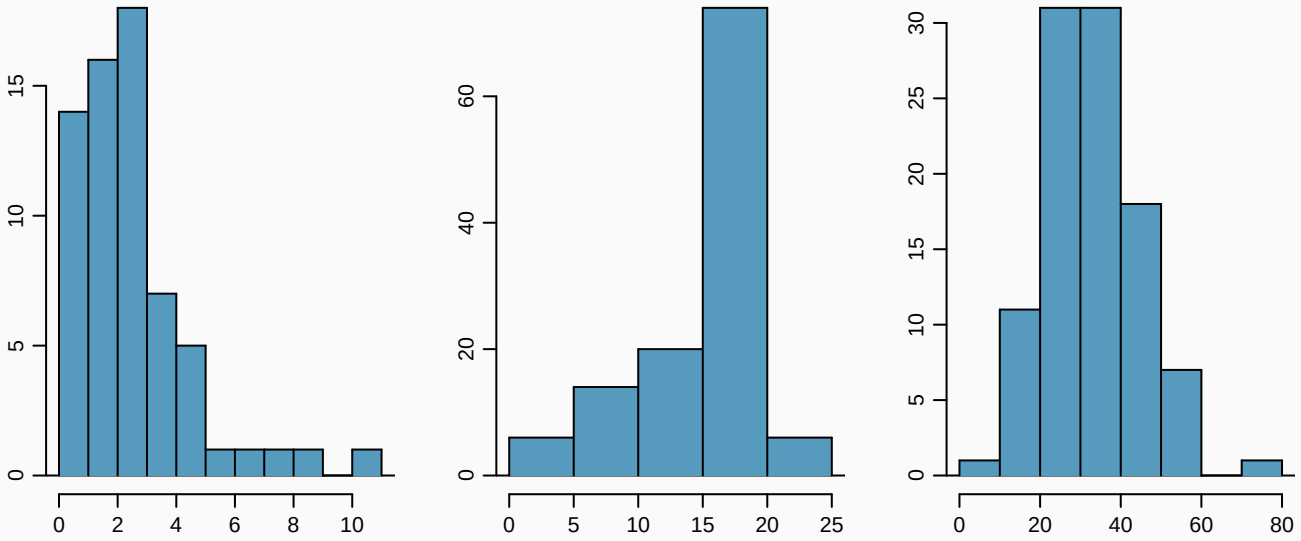
¿El histograma tiene un solo pico prominente (unimodal), varios picos prominentes (bimodal/multimodal) o ningún pico aparente (uniforme)?



Note: Para determinar la modalidad, dé un paso atrás e imagine una curva suave sobre el histograma; imagine que las barras son bloques de madera y deja caer un espagueti flácido sobre ellos, la forma que tomarían los espaguetis podría verse como una curva suave .

Forma de una distribución: asimetría

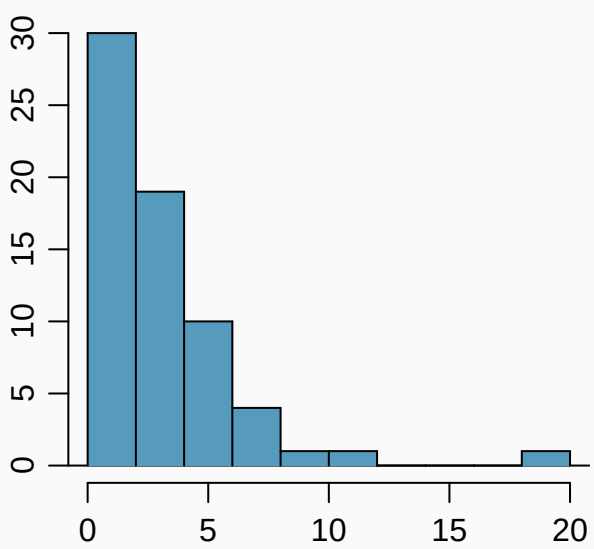
¿El histograma es sesgo a la derecha, sesgo a la izquierda o simétrico?

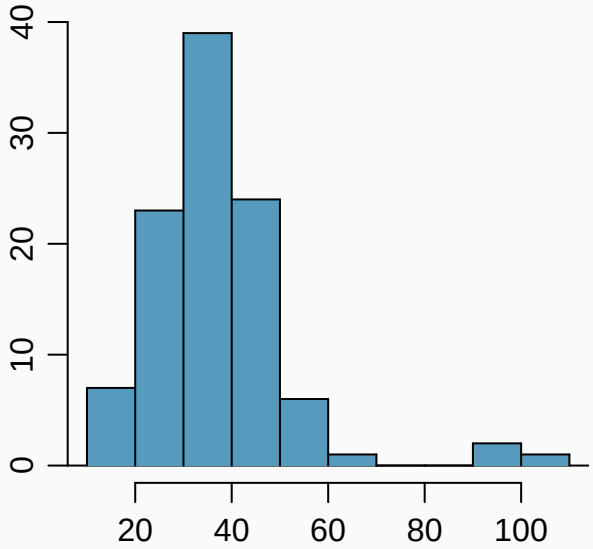


Note: Se dice que los histogramas están sesgados hacia el lado de la cola larga.

Forma de una distribución: observaciones inusuales

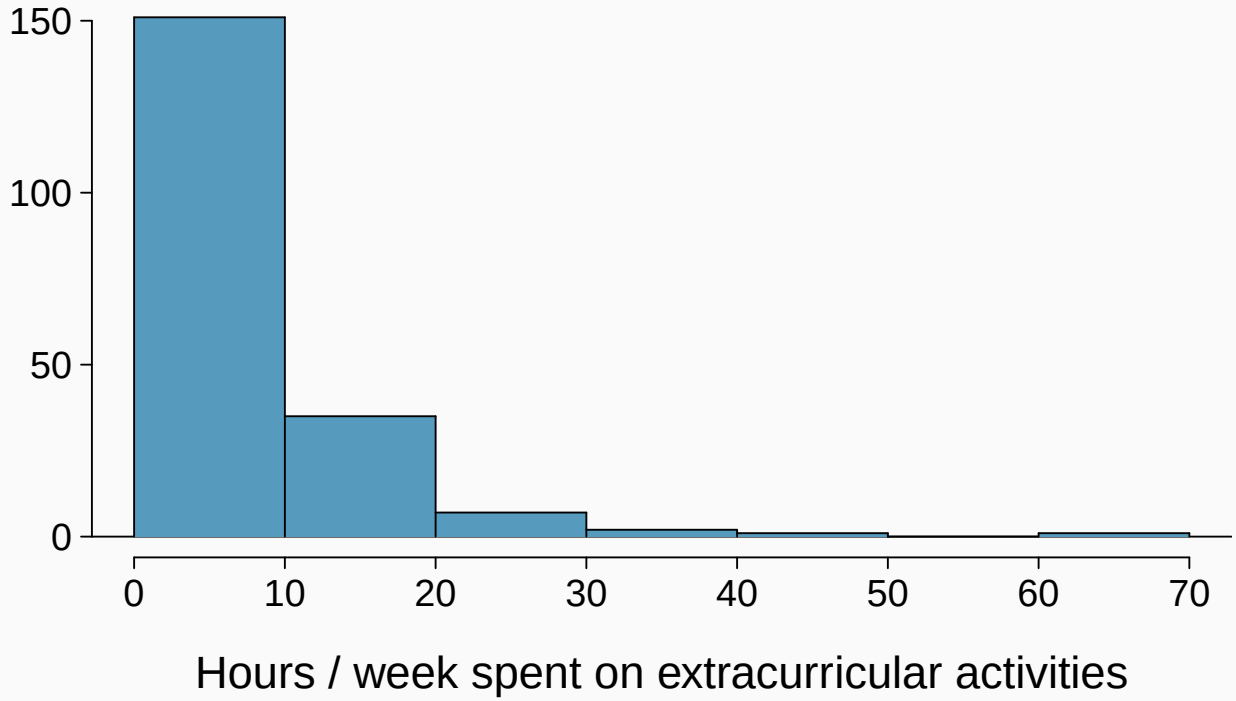
¿Existen observaciones inusuales o posibles valores atípicos?





Actividades extracurriculares

¿Cómo describiría la forma de la distribución de las horas por semana que los estudiantes dedican a actividades extracurriculares?

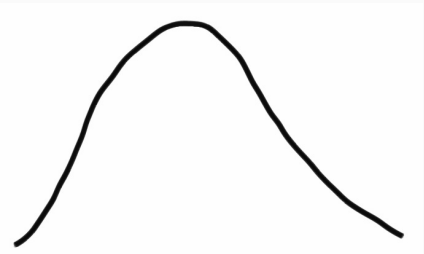


Unimodal y sesgada a la derecha, con una observación potencialmente inusual a las 60 horas/semana.

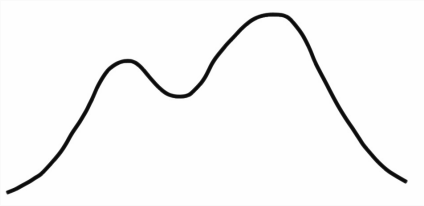
Formas de distribuciones comúnmente observadas

- modalidad

unimodal

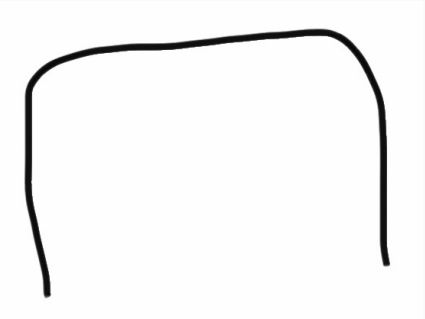


bimodal



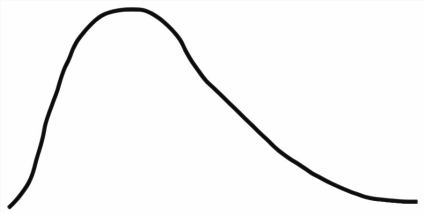
multimodal

uniforme

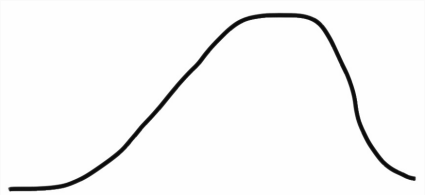


- sesgo

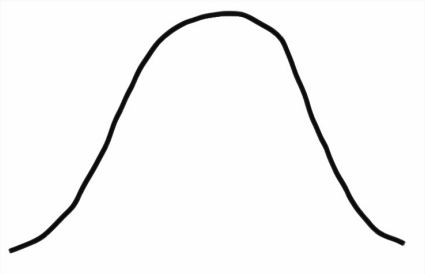
sesgo a la derecha



sesgo a la izquierda



simétrico



Varianza

Varianza es aproximadamente la desviación cuadrada promedio de la media.

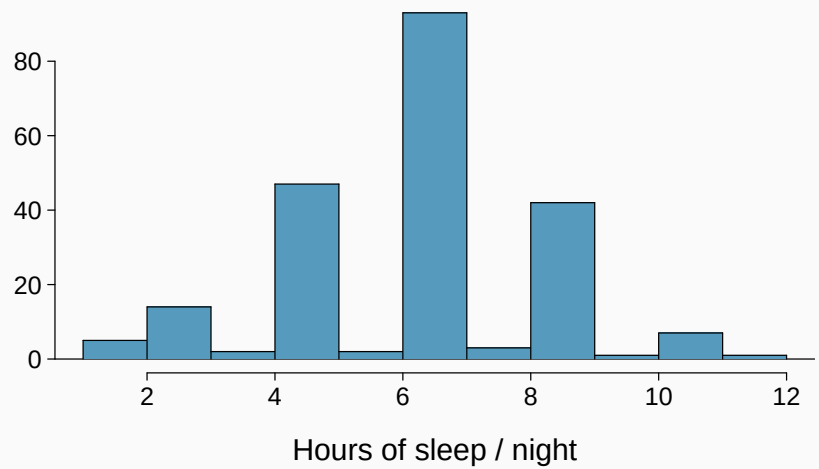
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Varianza

Varianza es aproximadamente la desviación cuadrada promedio de la media.

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

- La media de la muestra es $\bar{x} = 6.71$, y el tamaño de la muestra es $n = 217$.
- La variación de la cantidad de sueño que los estudiantes obtienen por noche se puede calcular como:



$$s^2 = \frac{(5-6,71)^2 + (9-6,71)^2 + \cdots + (7-6,71)^2}{217-1} = 4,11 \text{ horas}^2$$

Varianza (cont.)

¿Por qué usamos la desviación al cuadrado en el cálculo de la varianza?

- Para deshacerse de los negativos para que las observaciones igualmente distantes de la media se pesen igualmente.
- Para pesar más las desviaciones más grandes.

Desviación estándar

La desviación estándar es la raíz cuadrada de la varianza y tiene las mismas unidades que los datos

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

Desviación estándar

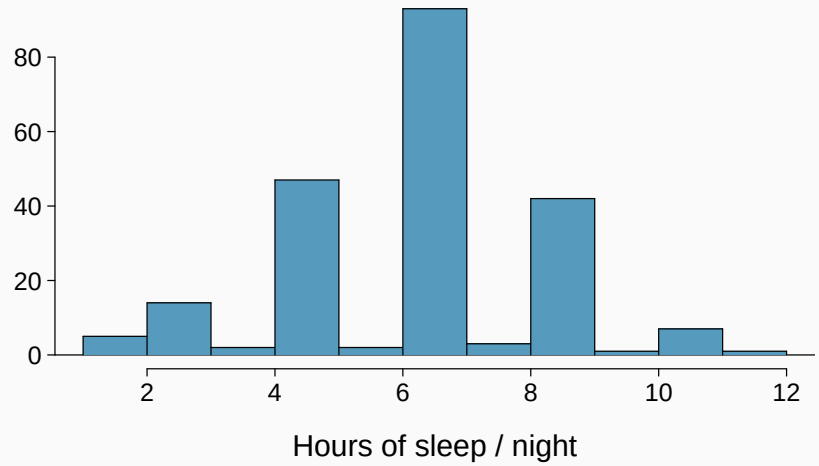
La desviación estándar es la raíz cuadrada de la varianza y tiene las mismas unidades que los datos

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

- La desviación estándar de la cantidad de horas de sueño de los estudiantes por noche se puede calcular como:

$$s = \sqrt{4.11} = 2.03 \text{ horas}$$

- Podemos ver que todos los datos están dentro de 3 desviaciones estándar de la media.



Mediana

- La mediana es el valor que divide los datos por la mitad cuando se ordenan en orden ascendente.

$$0, 1, 2, 3, 4$$

- Si hay un número par de observaciones, entonces la mediana es el promedio de los dos valores en el medio.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Dado que la mediana es el punto medio de los datos, el 50% de los valores están por debajo de ella. Por lo tanto, también es el 50^{th} percentil.

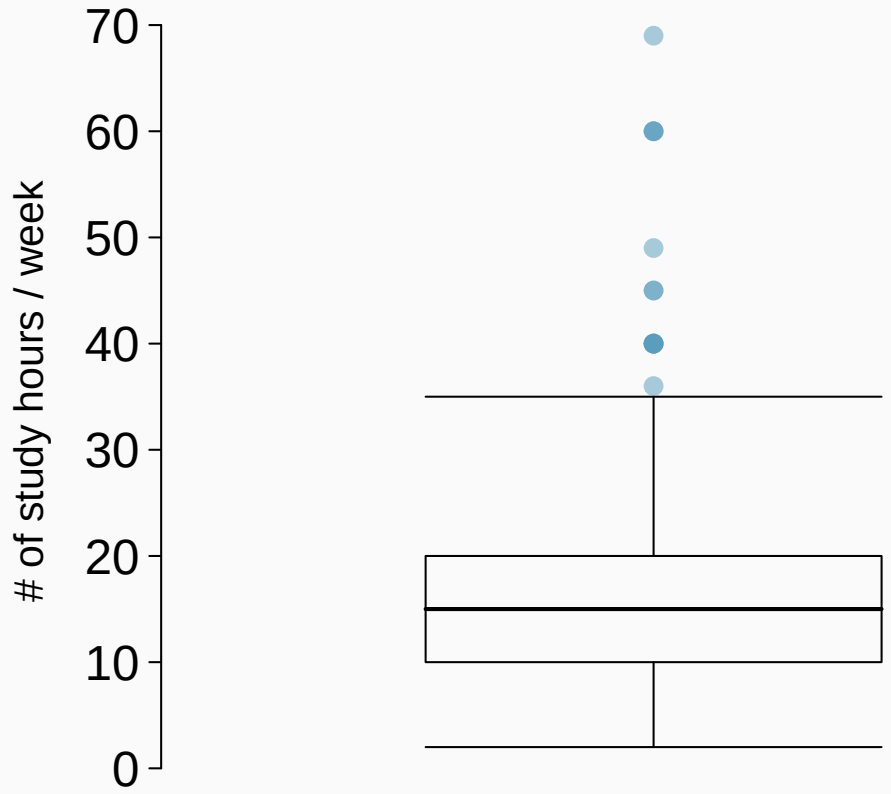
Q1, Q3 y IQR

- El percentil 25^{th} también se denomina primer cuartil, Q1.
- El percentil de 50^{th} también se llama mediana.
- El percentil 75^{th} también se denomina tercer cuartil, Q3.
- Entre Q1 y Q3 está el 50% medio de los datos. El rango que abarcan estos datos se llama rango intercuartílico o IQR.

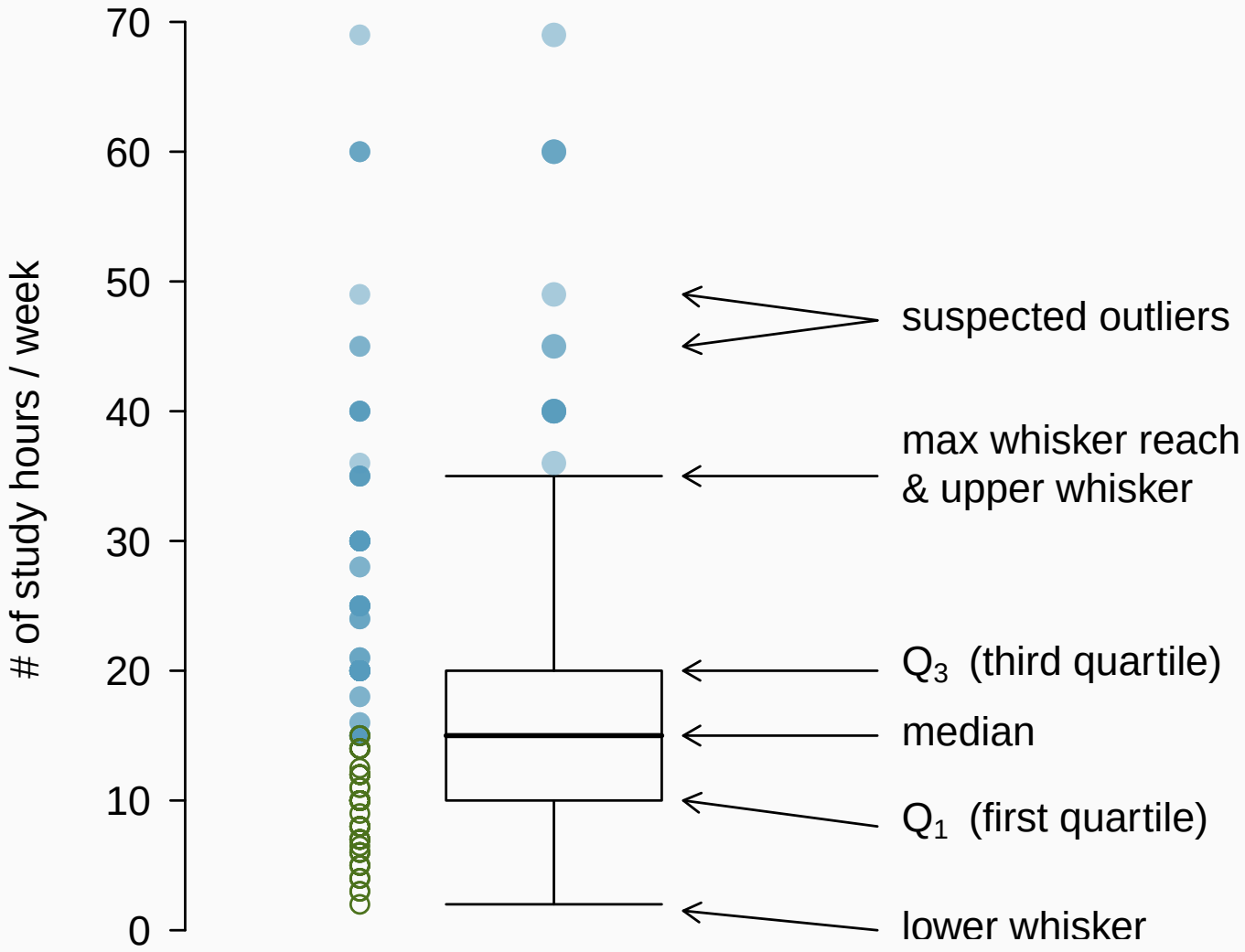
$$\text{IQR} = \text{Q3} - \text{Q1}$$

Diagrama de caja

La caja en un diagrama de caja representa el 50% medio de los datos, y la línea gruesa en la caja es la mediana.



Anatomía de un diagrama de caja



Bigotes y valores atípicos

- Bigotes de un diagrama de caja puede extenderse hasta $1.5 \times \text{IQR}$ lejos de los cuartiles.

$$\text{alcance máximo del bigot superior} = \text{Q3} + 1.5 \times \text{IQR}$$

$$\text{alcance máximo del bigot inferior} = \text{Q1} - 1.5 \times \text{IQR}$$

$$\text{IQR} : 20 - 10 = 10$$

$$\text{alcance máximo del bigot superior} = 20 + 1.5 \times 10 = 35$$

$$\text{alcance máximo del bigot inferior} = 10 - 1.5 \times 10 = -5$$

- Un potencial outlier se define como una observación más allá del alcance máximo de los bigotes. Es una observación que parece extrema en relación con el resto de los datos.

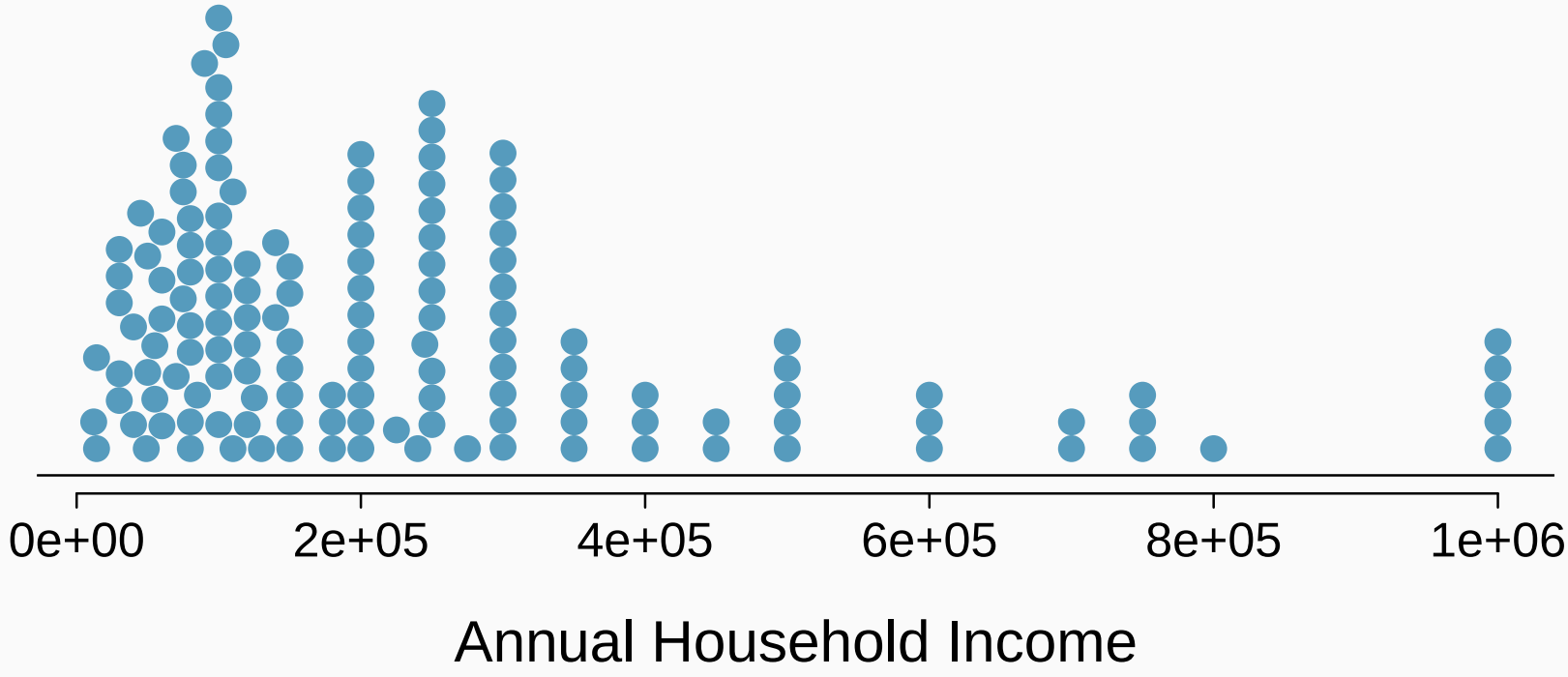
Valores atípicos (continuación)

¿Por qué es importante buscar valores atípicos?

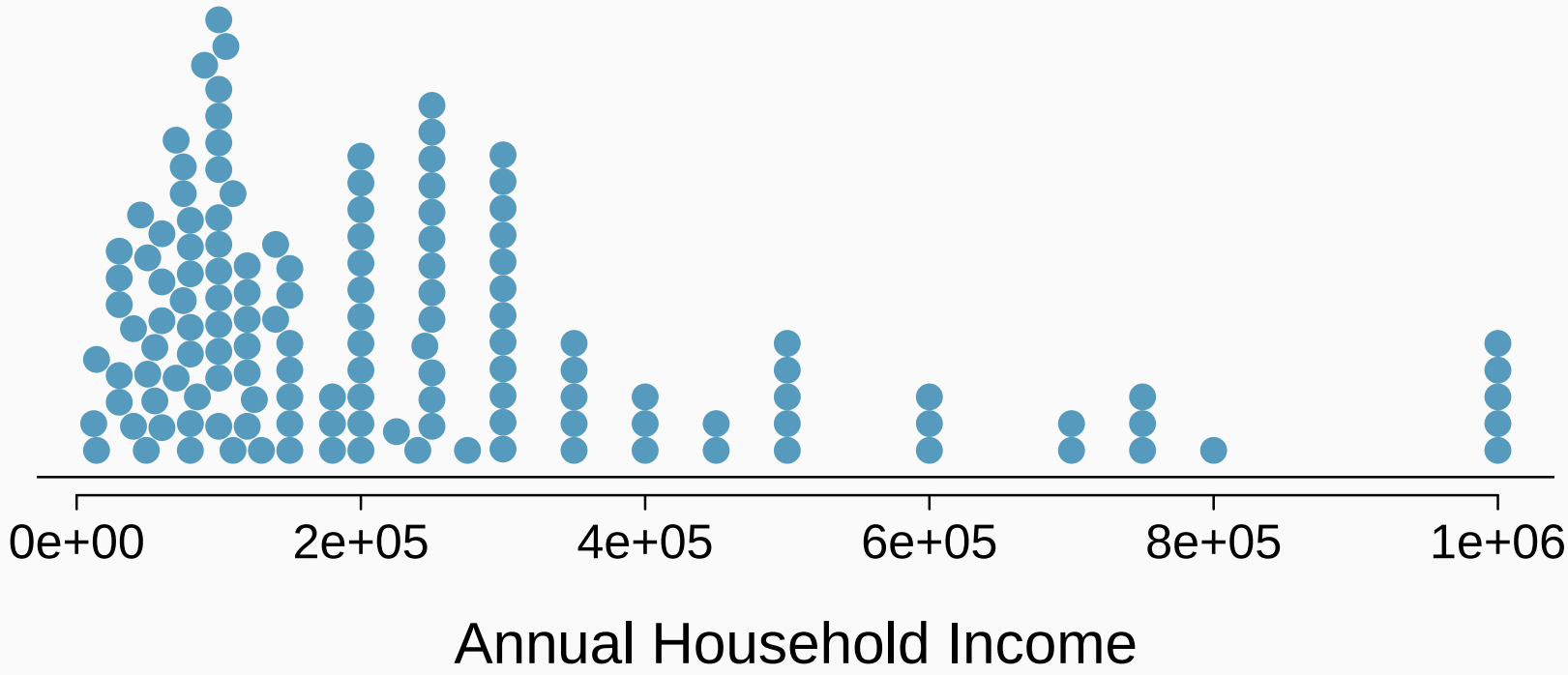
- Sesgos extremos en la distribución.
- Recopilación de datos y los errores de entrada.
- Proporcionar información sobre características interesantes de los datos.

Observaciones extremas

¿Cómo se verían afectadas las estadísticas muestrales como la media, la mediana, la SD y el IQR del ingreso familiar si el valor más grande se reemplazara con $10 millones? ¿Qué pasaría si el valor más pequeño se reemplazara con $10 millones?



Estadísticas sólidas



	robusto		no robusto	
escenario	mediana	IQR	$\bar{x}$	s
datos originales	190K	200K	245K	226K
mueve más grande a $10 millones	190K	200K	309K	853K
mueve el más pequeño a $10 millones	200K	200K	316K	854K

Estadísticas sólidas

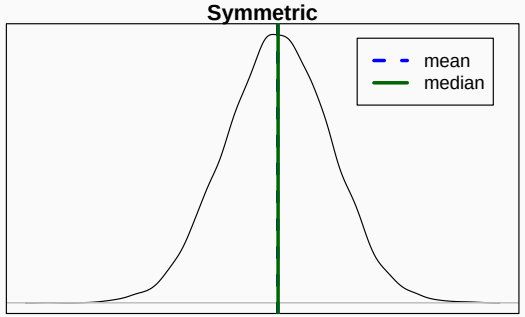
La mediana y el IQR son más resistentes a la asimetría y los valores atípicos que la media y la SD. Por lo tanto,

- para distribuciones sesgadas, a menudo es más útil usar la mediana y el IQR para describir el centro y la dispersión
- para distribuciones simétricas, a menudo es más útil usar la media y la SD para describir el centro y la dispersión

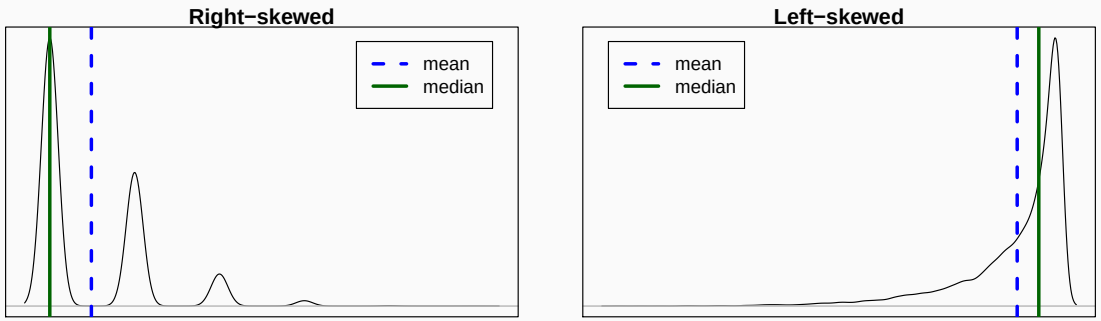
Si quisiera estimar el ingreso familiar típico de un estudiante, ¿estaría más interesado en el ingreso medio o mediano?

Media frente a mediana

- Si la distribución es simétrica, el centro a menudo se define como la media: media $\approx$ mediana

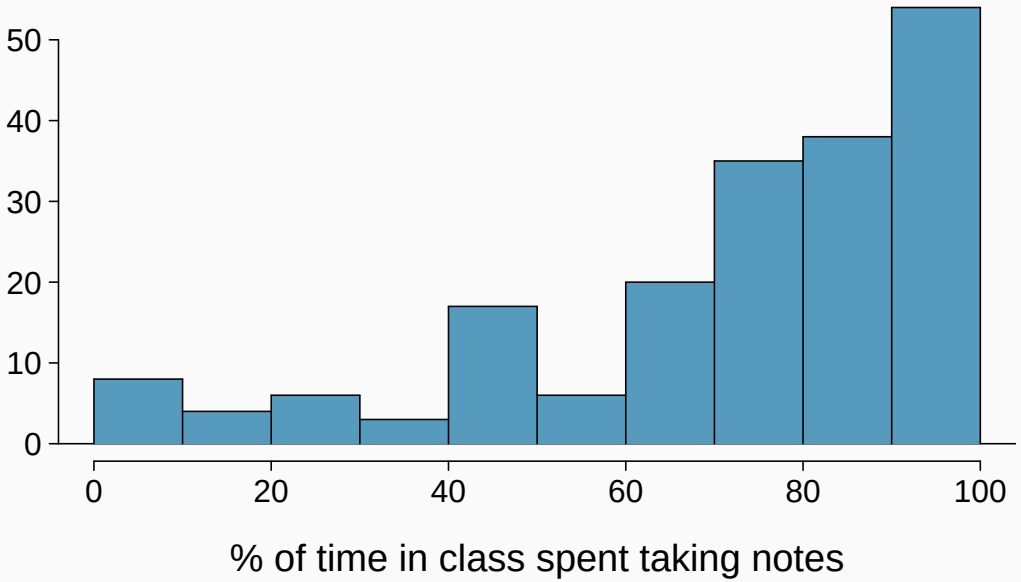


- Si la distribución es asimétrica o tiene valores atípicos extremos, el centro a menudo se define como la mediana
  - Sesgo a la derecha: media > mediana
  - Sesgo a la izquierda: media < mediana



Práctica

¿Cuál es más probable que sea cierto para la distribución del porcentaje de tiempo realmente dedicado a tomar notas en clase frente a Facebook, Twitter, etc.?



(a) media> mediana

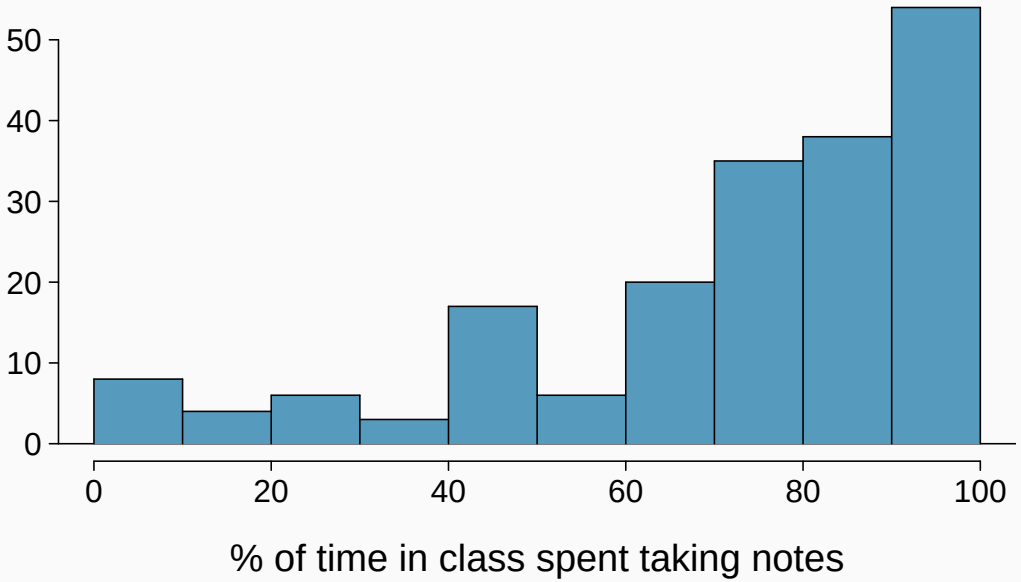
(b) media < mediana

(c) media ≈ mediana

(d) imposible de decir

Práctica

¿Cuál es más probable que sea cierto para la distribución del porcentaje de tiempo realmente dedicado a tomar notas en clase frente a Facebook, Twitter, etc.?



mediana: 80%
media: 76%

(a) media> mediana

(b) media < mediana

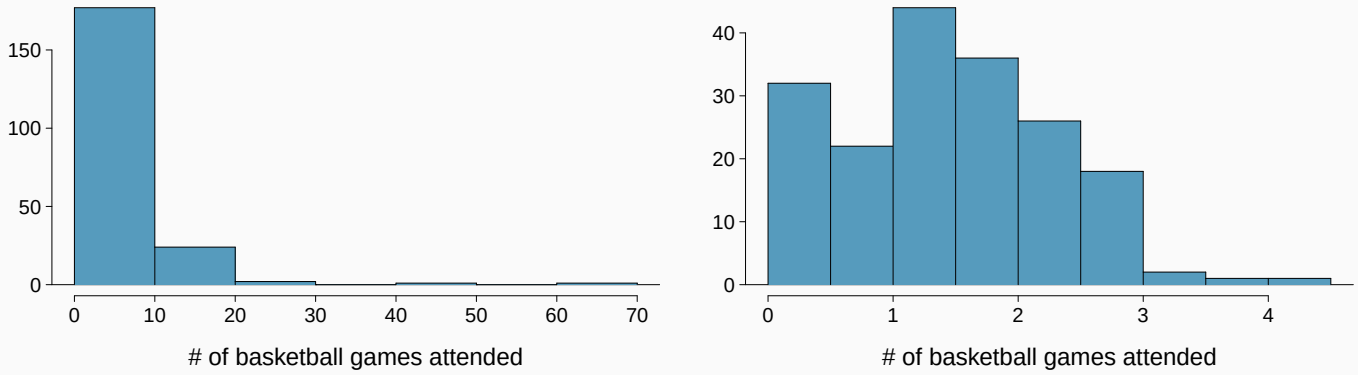
(c) media ≈ mediana

(d) imposible de decir

Datos extremadamente sesgados

Cuando los datos están extremadamente sesgados, transformarlos podría facilitar el modelado. Una transformación común es la transformación logarítmica.

Los histogramas de la izquierda muestran la distribución del número de partidos de baloncesto a los que asistieron los estudiantes. El histograma de la derecha muestra la distribución del registro del número de juegos asistidos.



Ventajas y desventajas de las transformaciones

- Los datos sesgados son más fáciles de modelar cuando se transforman porque los valores atípicos tienden a volverse mucho menos prominentes después de una transformación adecuada.

# de juegos	70	50	25	$\cdots$
log(# de juegos)	4.25	3.91	3.22	$\cdots$

- Sin embargo, los resultados de un análisis en unidades logarítmicas de la variable medida pueden ser difíciles de interpretar.

¿Qué otras variables esperaría que estuvieran extremadamente sesgadas?

Ventajas y desventajas de las transformaciones

- Los datos sesgados son más fáciles de modelar cuando se transforman porque los valores atípicos tienden a volverse mucho menos prominentes después de una transformación adecuada.

# de juegos	70	50	25	$\cdots$
log(# de juegos)	4.25	3.91	3.22	$\cdots$

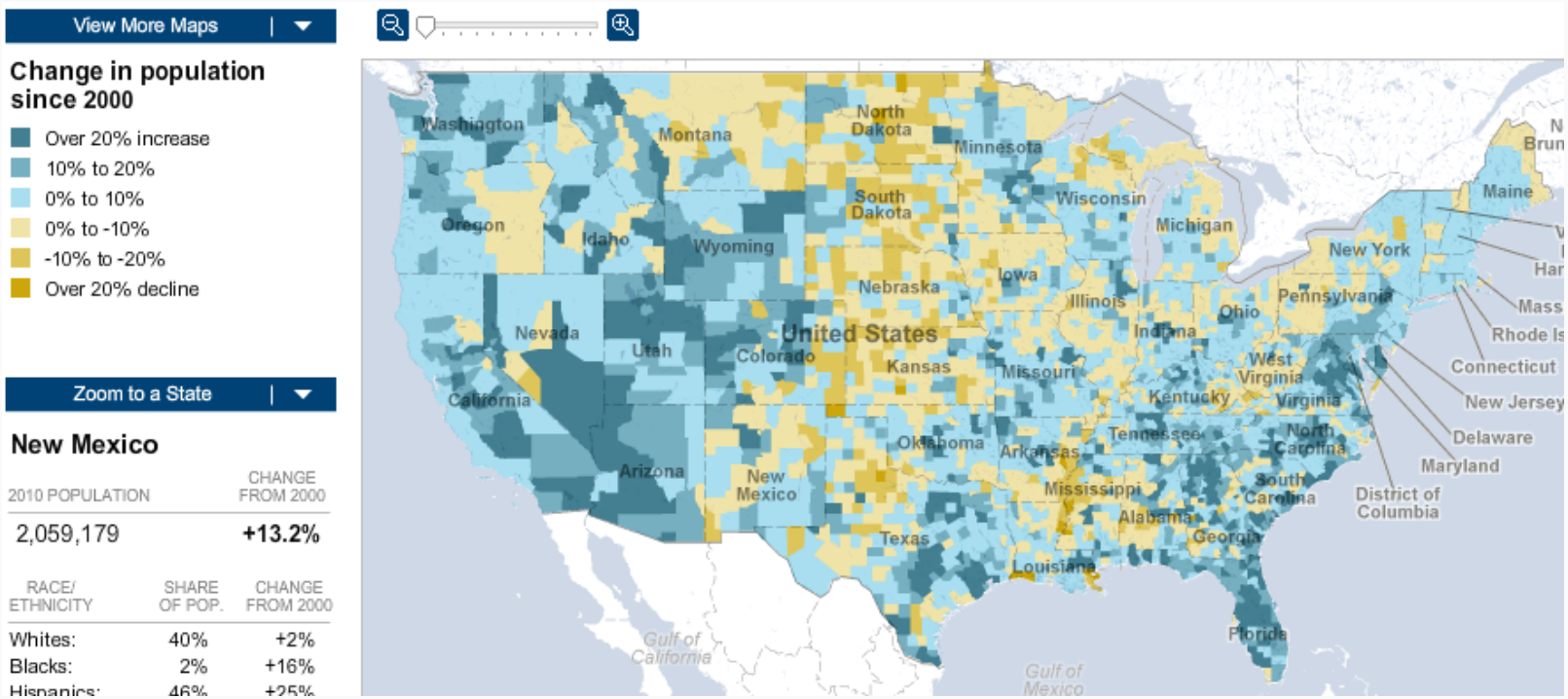
- Sin embargo, los resultados de un análisis en unidades logarítmicas de la variable medida pueden ser difíciles de interpretar.

¿Qué otras variables esperaría que estuvieran extremadamente sesgadas?

Salario, precio de la vivienda, etc.

Mapas de intensidad

¿Qué patrones son evidentes en el cambio de población entre 2000 y 2010?



http://projects.nytimes.com/census/2010/map

Datos categóricos

Tablas de contingencia

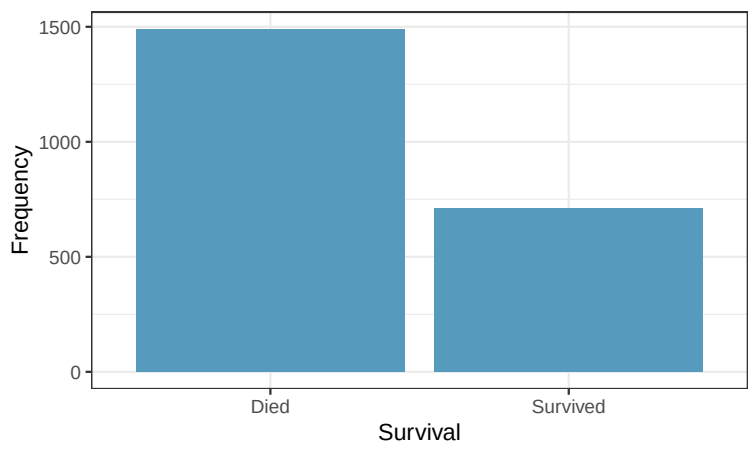
Una tabla que resume los datos de dos variables categóricas se denomina tabla de contingencia.

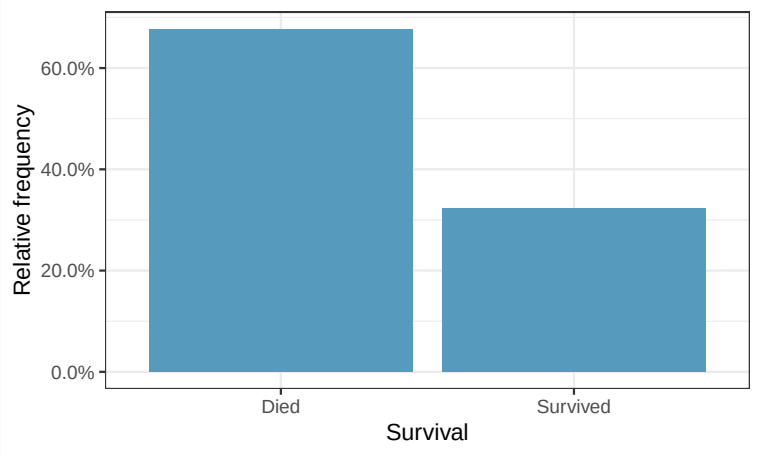
La siguiente tabla de contingencia muestra la distribución de supervivencia y edades de los pasajeros del Titanic.

		Supervivencia		
		Murió	Sobrevivió	Total
Edad	Adulto	1438	654	2092
	Niño	52	57	109
	Total	1490	711	2201

Gráficos de barras

Un gráfico de barras es una forma común de mostrar una única variable categórica. Un gráfico de barras donde se muestran proporciones en lugar de frecuencias se llama gráfico de barras de frecuencias relativas.





¿En qué se diferencian los gráficos de barras de los histogramas?

Los diagramas de barras se usan para mostrar distribuciones de variables categóricas, los histogramas se usan para variables numéricas. El eje x en un histograma es una recta numérica, por lo tanto, el orden de las barras no se puede cambiar. En un gráfico de barras, las categorías se pueden enumerar en cualquier orden (aunque algunos ordenamientos tienen más

Eligiendo la proporción apropiada

¿Parece haber una relación entre la edad y la supervivencia de los pasajeros del Titanic?

		Supervivencia		
		Murió	Sobrevivió	Total
Edad	Adulto	1438	654	2092
	Niño	52	57	109
	Total	1490	711	2201

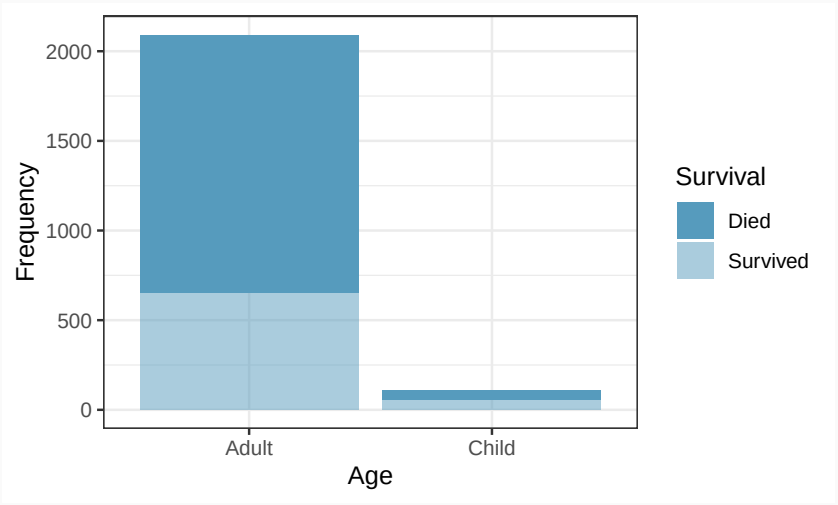
Para responder a esta pregunta, examinamos las proporciones de las filas:

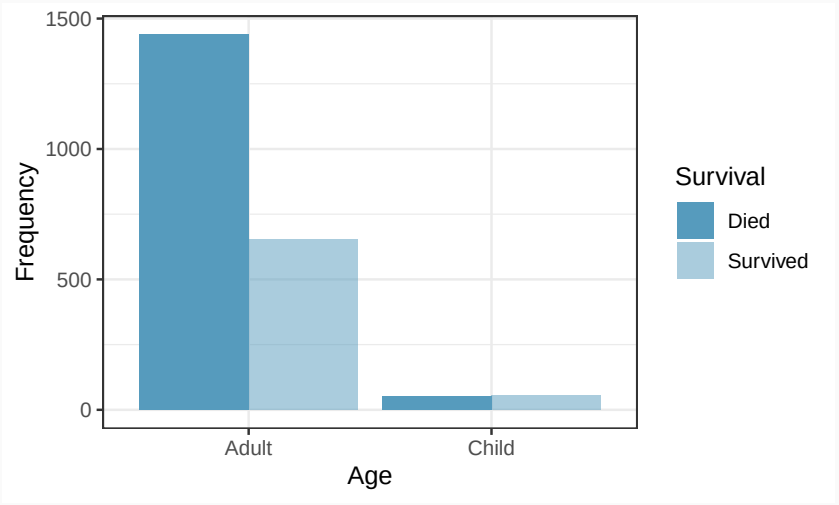
- % Adultos que sobrevivieron: $654 \,/\, 2092 \approx 0.31$
- % Niños que sobrevivieron: $57 \,/\, 109 \approx 0.52$

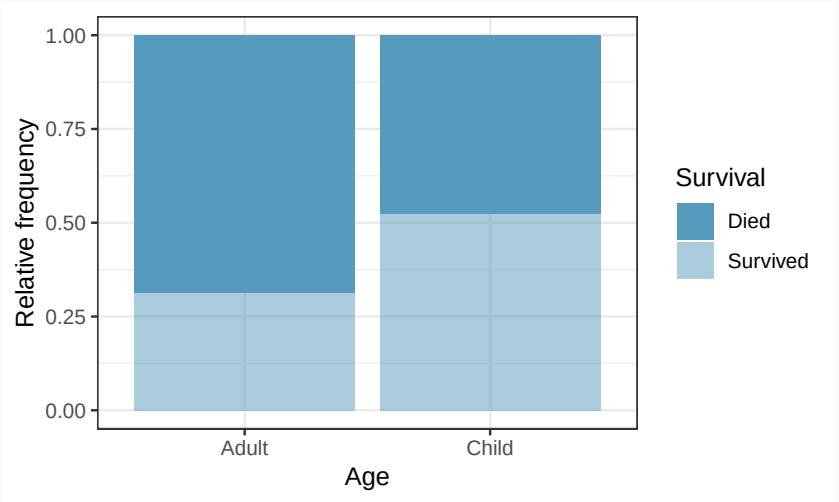
Gráficos de barras con dos variables

- Gráfico de barras apiladas: Visualización gráfica de la información de la tabla de contingencia, para conteos.
- Gráfico de barras lado a lado: Muestra la misma información colocando barras uno al lado del otro, en lugar de uno encima del otro.
- Gráfica de barras apiladas estandarizadas: visualización gráfica de la tabla de contingencia información, para proporciones.

¿Cuáles son las diferencias entre las tres visualizaciones que se muestran a continuación?

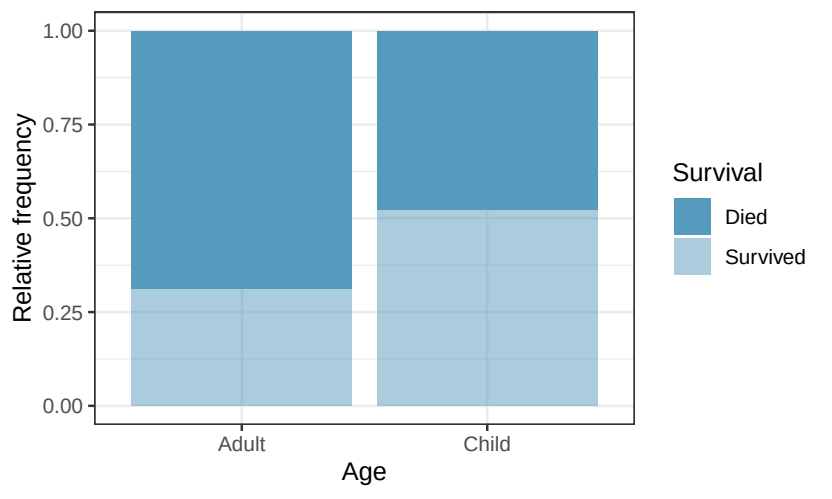


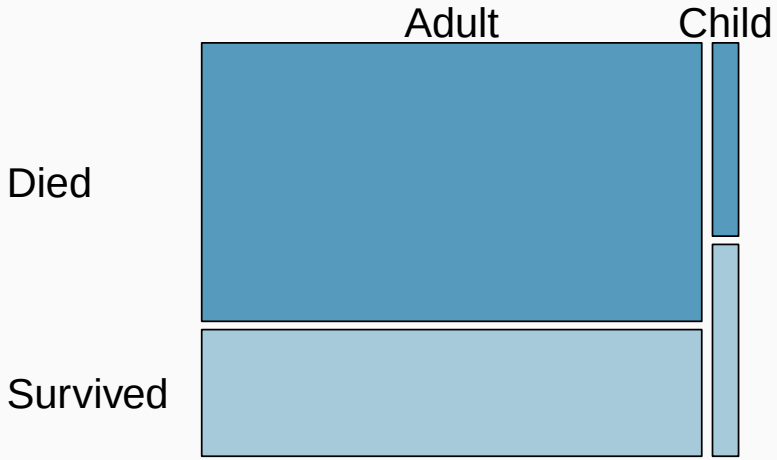




Gráficas de mosaico

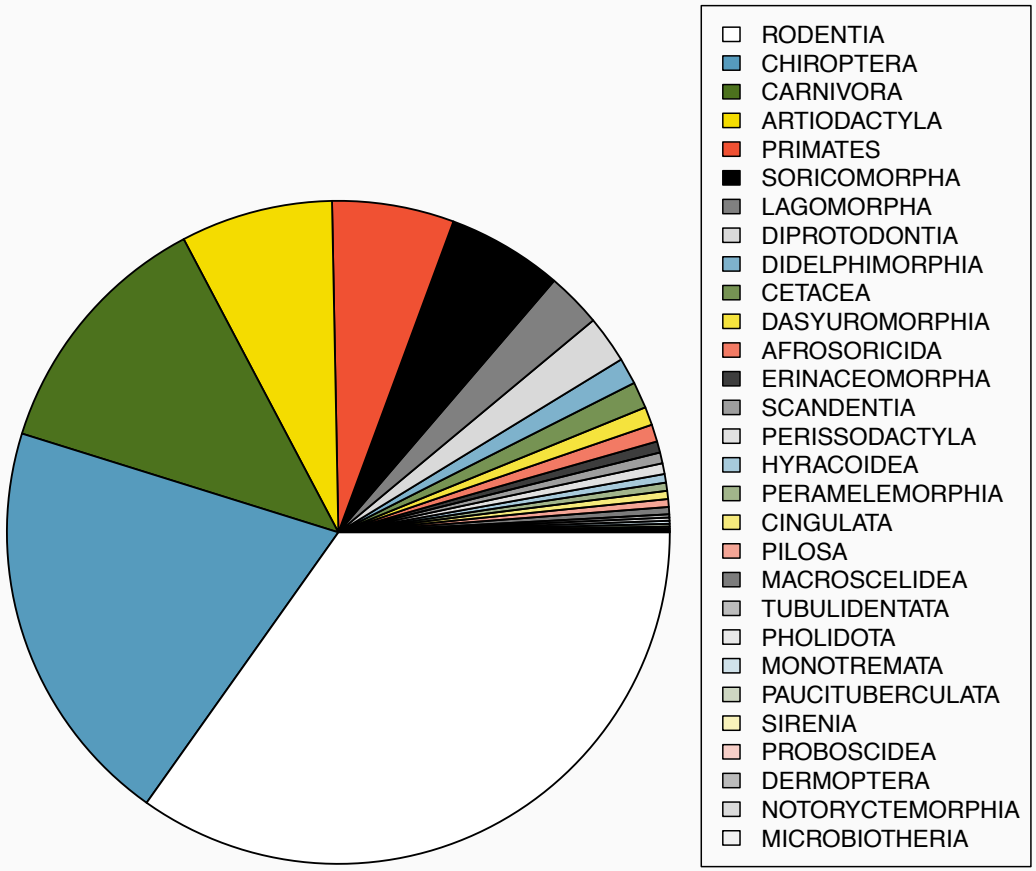
¿Cuál es la diferencia entre las dos visualizaciones que se muestran a continuación?





Gráficos circulares

¿Puedes decir qué orden abarca el porcentaje más bajo de especies de mamíferos?



Datos de http://www.bucknell.edu/msw3.

Datos numéricos entre grupos

Comparar datos numéricos entre grupos

Algunas de las investigaciones más interesantes pueden considerarse examinando datos numéricos entre grupos. Aquí se introducen dos métodos convenientes:

- diagramas de caja lado a lado
- histogramas huecos.

Datos de población

Median Income for 150 Counties, in $1000s

Population Gain						No Population Gain		
38.2	43.6	42.2	61.5	51.1	45.7	48.3	60.3	50.7
44.6	51.8	40.7	48.1	56.4	41.9	39.3	40.4	40.3
40.6	63.3	52.1	60.3	49.8	51.7	57	47.2	45.9
51.1	34.1	45.5	52.8	49.1	51	42.3	41.5	46.1
80.8	46.3	82.2	43.6	39.7	49.4	44.9	51.7	46.4
75.2	40.6	46.3	62.4	44.1	51.3	29.1	51.8	50.5
51.9	34.7	54	42.9	52.2	45.1	27	30.9	34.9
61	51.4	56.5	62	46	46.4	40.7	51.8	61.1
53.8	57.6	69.2	48.4	40.5	48.6	43.4	34.7	45.7
53.1	54.6	55	46.4	39.9	56.7	33.1	21	37
63	49.1	57.2	44.1	50	38.9	52	31.9	45.7
46.6	46.5	38.9	50.9	56	34.6	56.3	38.7	45.7
74.2	63	49.6	53.7	77.5	60	56.2	43	21.7
63.2	47.6	55.9	39.1	57.8	42.6	44.5	34.5	48.9
50.4	49	45.6	39	38.8	37.1	50.9	42.1	43.2
57.2	44.7	71.7	35.3	100.2		35.4	41.3	33.6
42.6	55.5	38.6	52.7	63		43.4	56.5	

En esta tabla se muestran a la izquierda, el ingreso familiar promedio (en miles de dólares) de una muestra aleatoria de 100 condados que tuvieron aumentos de población. A la derecha, se muestran los ingresos medianos de una muestra aleatoria de 50 condados que no tuvieron aumento de población.

Gráficos

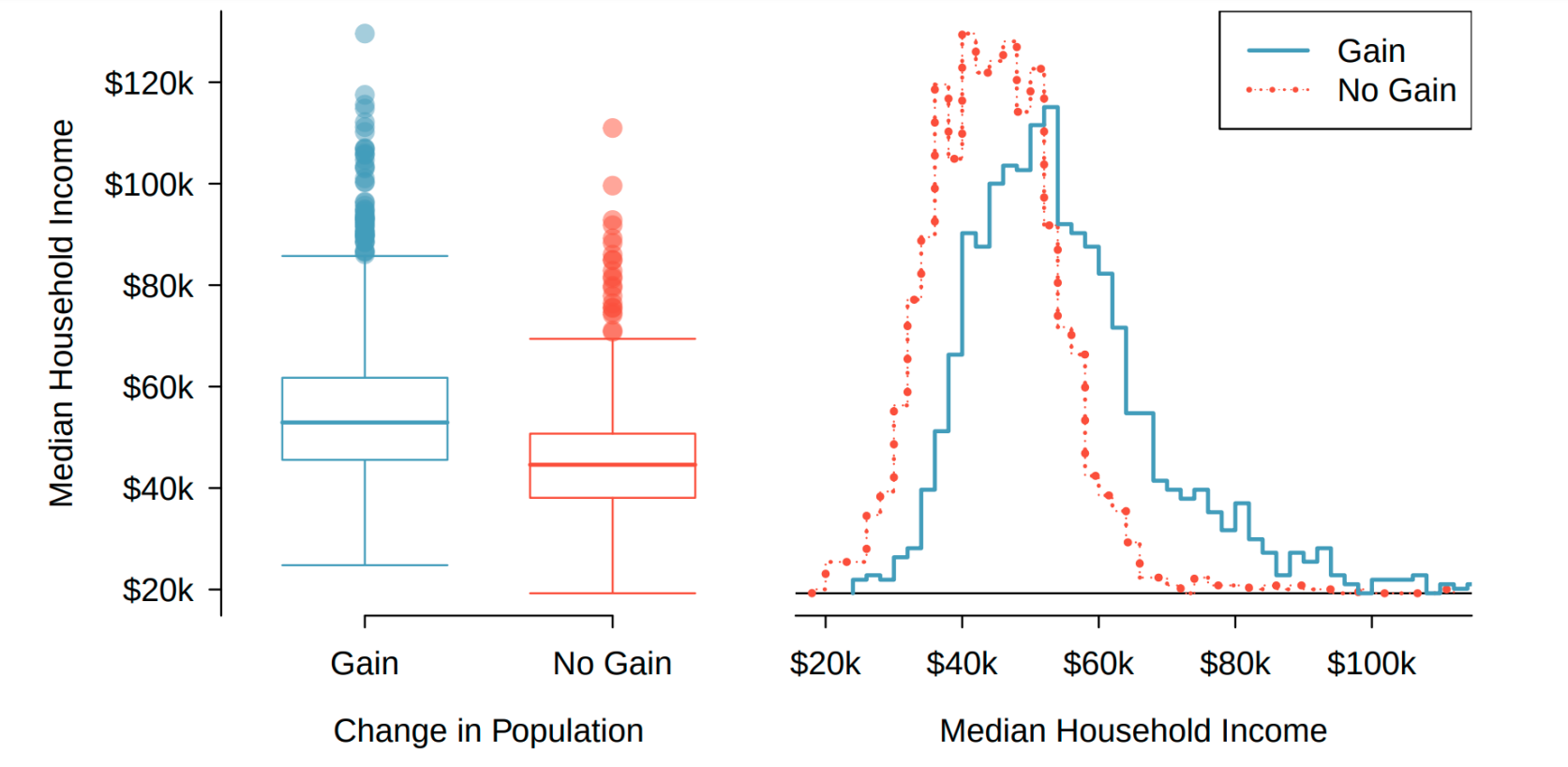


Diagrama de caja uno al lado del otro (panel izquierdo) e histogramas huecos (panel derecho) para la mediana del ingreso, donde los condados se dividen según si hubo una ganancia o pérdida de población.

Capítulo 2 del libro OpenIntro Statistics.
Las diapositivas están basadas en las slides desarrolladas por Mine Çetinkaya-Rundel de OpenIntro y el contenido del libro.