Herramientas matemáticas y estadísticas
Segunda parte
Curso Propedéutico Especialidad en Divulgación de la Economía

Jorge de la Vega Góngora
jorge.delavegagongora@gmail.com

Museo Interactivo de Economía

Sesiones 10 a 12

Temas I

1. Introducción
2. Módulo 10. Herramientas Estadísticas I: Conceptos básicos de inferencia
   - Estadística
   - Probabilidad
   - Probabilidad Condicional
   - Variables aleatorias y mediciones básicas
   - Medidas de tendencia central y dispersión
   - Distribución normal
3. Módulo 11. Herramientas Estadísticas II: Gráficas y Regresión Lineal
   - Gráficas estadísticas
   - Correlación
   - Mínimos cuadrados ordinarios
   - Intervalos de confianza de los parámetros
4. Módulo 12. Herramientas Estadísticas III: Muestreo y diseño de encuestas
   - Muestreo
   - Esquemas de muestreo
   - Muestreo aleatorio simple
   - Encuestas
   - Diseño de cuestionarios

Introducción

El **Camino Real** *Persa fue una antigua carretera construida por el rey persa Darío I en el siglo V A.C. Darío construyó el camino para facilitar una comunicación rápida a través de su extenso imperio que abarcaba desde Susa hasta Sardes.*

Se dice que Euclides contestó a la pregunta del Rey Ptolomeo sobre una forma más sencilla de aprender matemáticas diciendo que "No hay Camino Real hacia la geometría".

Módulo 10. Herramientas Estadísticas I: Conceptos básicos de inferencia

¿Qué es la estadística?

La *Estadística* es el arte de hacer conjeturas numéricas acerca de preguntas complicadas.

¿Cuáles son los efectos de nuevos tratamientos médicos?

¿Qué es lo que causa el parecido entre padres e hijos, y qué tan fuerte es esa fuerza?

¿Porqué los Casinos nunca pierden?

¿Cuánta gente está empleada? ¿Desempleada?

¿Cuál será el tipo de cambio promedio del próximo sexenio?

Los temas anteriores son difíciles, y los métodos estadísticos ayudan a pensar adecuadamente en ellos.
Una característica común a todos estas preguntas es que hay *incertidumbre* en el contexto del problema.

¿Qué es la estadística?

La *Estadística* es el arte de hacer conjeturas numéricas acerca de preguntas complicadas.

¿Cuáles son los efectos de nuevos tratamientos médicos?

¿Qué es lo que causa el parecido entre padres e hijos, y qué tan fuerte es esa fuerza?

¿Porqué los Casinos nunca pierden?

¿Cuánta gente está empleada? ¿Desempleada?

¿Cuál será el tipo de cambio promedio del próximo sexenio?

Los temas anteriores son difíciles, y los métodos estadísticos ayudan a pensar adecuadamente en ellos.
Una característica común a todos estas preguntas es que hay *incertidumbre* en el contexto del problema.

¿Qué es la estadística?

La *Estadística* es el arte de hacer conjeturas numéricas acerca de preguntas complicadas.

¿Cuáles son los efectos de nuevos tratamientos médicos?

¿Qué es lo que causa el parecido entre padres e hijos, y qué tan fuerte es esa fuerza?

¿Porqué los Casinos nunca pierden?

¿Cuánta gente está empleada? ¿Desempleada?

¿Cuál será el tipo de cambio promedio del próximo sexenio?

Los temas anteriores son difíciles, y los métodos estadísticos ayudan a pensar adecuadamente en ellos.
Una característica común a todos estas preguntas es que hay *incertidumbre* en el contexto del problema.

¿Qué es la estadística?

La *Estadística* es el arte de hacer conjeturas numéricas acerca de preguntas complicadas.

¿Cuáles son los efectos de nuevos tratamientos médicos?

¿Qué es lo que causa el parecido entre padres e hijos, y qué tan fuerte es esa fuerza?

¿Porqué los Casinos nunca pierden?

¿Cuánta gente está empleada? ¿Desempleada?

¿Cuál será el tipo de cambio promedio del próximo sexenio?

Los temas anteriores son difíciles, y los métodos estadísticos ayudan a pensar adecuadamente en ellos.
Una característica común a todos estas preguntas es que hay *incertidumbre* en el contexto del problema.

¿Qué es la estadística?

La *Estadística* es el arte de hacer conjeturas numéricas acerca de preguntas complicadas.

¿Cuáles son los efectos de nuevos tratamientos médicos?

¿Qué es lo que causa el parecido entre padres e hijos, y qué tan fuerte es esa fuerza?

¿Porqué los Casinos nunca pierden?

¿Cuánta gente está empleada? ¿Desempleada?

¿Cuál será el tipo de cambio promedio del próximo sexenio?

Los temas anteriores son difíciles, y los métodos estadísticos ayudan a pensar adecuadamente en ellos.
Una característica común a todos estas preguntas es que hay *incertidumbre* en el contexto del problema.

¿Qué es la estadística?

La *Estadística* es el arte de hacer conjeturas numéricas acerca de preguntas complicadas.

¿Cuáles son los efectos de nuevos tratamientos médicos?

¿Qué es lo que causa el parecido entre padres e hijos, y qué tan fuerte es esa fuerza?

¿Porqué los Casinos nunca pierden?

¿Cuánta gente está empleada? ¿Desempleada?

¿Cuál será el tipo de cambio promedio del próximo sexenio?

Los temas anteriores son difíciles, y los métodos estadísticos ayudan a pensar adecuadamente en ellos.

Una característica común a todos estas preguntas es que hay *incertidumbre* en el contexto del problema.

¿Qué es la estadística?

La *Estadística* es el arte de hacer conjeturas numéricas acerca de preguntas complicadas.

¿Cuáles son los efectos de nuevos tratamientos médicos?

¿Qué es lo que causa el parecido entre padres e hijos, y qué tan fuerte es esa fuerza?

¿Porqué los Casinos nunca pierden?

¿Cuánta gente está empleada? ¿Desempleada?

¿Cuál será el tipo de cambio promedio del próximo sexenio?

Los temas anteriores son difíciles, y los métodos estadísticos ayudan a pensar adecuadamente en ellos.
Una característica común a todos estas preguntas es que hay *incertidumbre* en el contexto del problema.

¿Qué es la estadística?

Estadística

La Estadística se puede definir como la disciplina científica que estudia los temas de colección, organización y análisis de datos.

- Proviene de la palabra latina *status* referente al estado o condición de algo.
- Usada desde 1748 por estadístico alemán G. Achenwall, la estadística se enfocaba a la colección de datos sobre el Estado, como los económicos o demográficos, y a su respectiva agregación y análisis.
- Con el tiempo y con el desarrollo de la teoría de la probabilidad en los juegos de azar, principalmente por matemáticos en los siglos XVII y XVIII, los estadísticos comenzaron a incorporar las distribuciones de probabilidad como parte de los modelos que explicaban el comportamiento de los datos, ampliando la capacidad analítica, explicativa y predictiva de la Estadística.

¿Porqué se requiere la estadística? I

Sin datos, sólo eres otra persona con una opinión.

Las decisiones no deben basarse en anécdotas.

Actualmente es difícil funcionar en el mundo sin una comprensión básica de la estadística. Ejemplos de artículos periodísticos como los siguientes:

- El cólico infantil puede estar ligado a los padres: el artículo corresponde al reporte de un estudio de la relación entre el llanto excesivo y la depresión de los padres El estudio incluyó más de 7,600 bebés y sus padres y concluyó que el bebé tiende a llorar más si el padre reportó síntomas de depresión antes del nacimiento del bebé.

¿Porqué se requiere la estadística? II

- Pocos se ven como viejos, sin importar su edad: describe los resultados de una encuesta grande de 2969 adultos. Aquellos entrevistados se les preguntó a qué edad una persona debería ser considerada vieja. Los resultados mostraron que hay diferencias notables dependiendo de la edad de quien responde. La edad promedio considerada para la vejez fue 60 entre los encuestados con 18-29 años, mientras 69 para los del rango 30-49, 72 para los del rango 50-64 y 74 para aquellos de 65 y viejos.
- Si te dieran $20,000, ¿qué harías?: reporta un aspecto de un estudio sobre compras y ahorro del consumidor.

Para ser un adecuado consumidor de las noticias mencionadas, se debe ser capaz de lo siguiente:

- Extraer información de gráficas y tablas
- Seguir argumentos numéricos
- Comprender los conocimientos básicos de cómo los datos se conjuntan, sumarizan y analizan para extraer conclusiones estadísticas.

¿Para qué sirve la estadística?

Proceso Estadístico para el análisis de datos

El proceso de análisis de datos consta de los siguientes pasos:

1. Comprender la naturaleza del problema
2. Decidir qué medir y cómo medirlo
3. Recolectar información
4. Resumir los datos en información relevante y hacer un análisis preliminar
5. Llevar a cabo un análisis formal de datos, desarrollando modelos e inferencias.
6. Intepretar los resultados

Inferencia estadística

La **inferencia estadística** consiste en generalizar los resultados de una muestra a la población de la cual se obtuvo.

Probabilidad

- Todos los días hacemos decisiones bajo incertidumbre o nos enfrentamos a situaciones cuyo resultado final no podemos conocer de antemano:
  - ¿Cuánto tiempo tardaré en llegar hoy a la oficina?
  - Compro dólares hoy o los compro mañana?
  - Para mi hipoteca, ¿debo elegir una tasa de interés fija o variable?
- Podemos responder preguntas como estas usando las ideas y métodos de la probabilidad

Probabilidad

La probabilidad es el estudio sistemático de la incertidumbre, y nos permite definir una medida para darnos una idea cuantitativa de la incertidumbre asociada a una situación.

Interpretación de la probabilidad

- La probabilidad se puede interpretar de varias formas:
  - **Subjetiva**: medida personal de la creencia de que un evento ocurrirá. Un evento incierto en este contexto es cualquier evento del que se tiene cierta ignorancia sobre su resultado, eg: ¿cuánto dinero tengo en la bolsa?
  - **Frecuentista**: una probabilidad es la proporción de largo plazo de la ocurrencia de un evento, dado que se repiten bajo las mismas circunstancias.
- Desde el punto frecuentista, un **experimento** es un procedimiento que puede, al menos en teoría repetirse una cantidad infinita de veces y que tiene un conjunto bien definido de posibles resultados. Un **evento** es la ocurrencia de un subconjunto de esos posibles resultados.
- El **espacio muestral** es el conjunto de todos los posibles resultados de un experimento.
- Asignar una probabilidad a un evento es un intento de cuantificar su posibilidad de ocurrencia. Se asignan números entre 0 y 1.

Ley de los grandes números

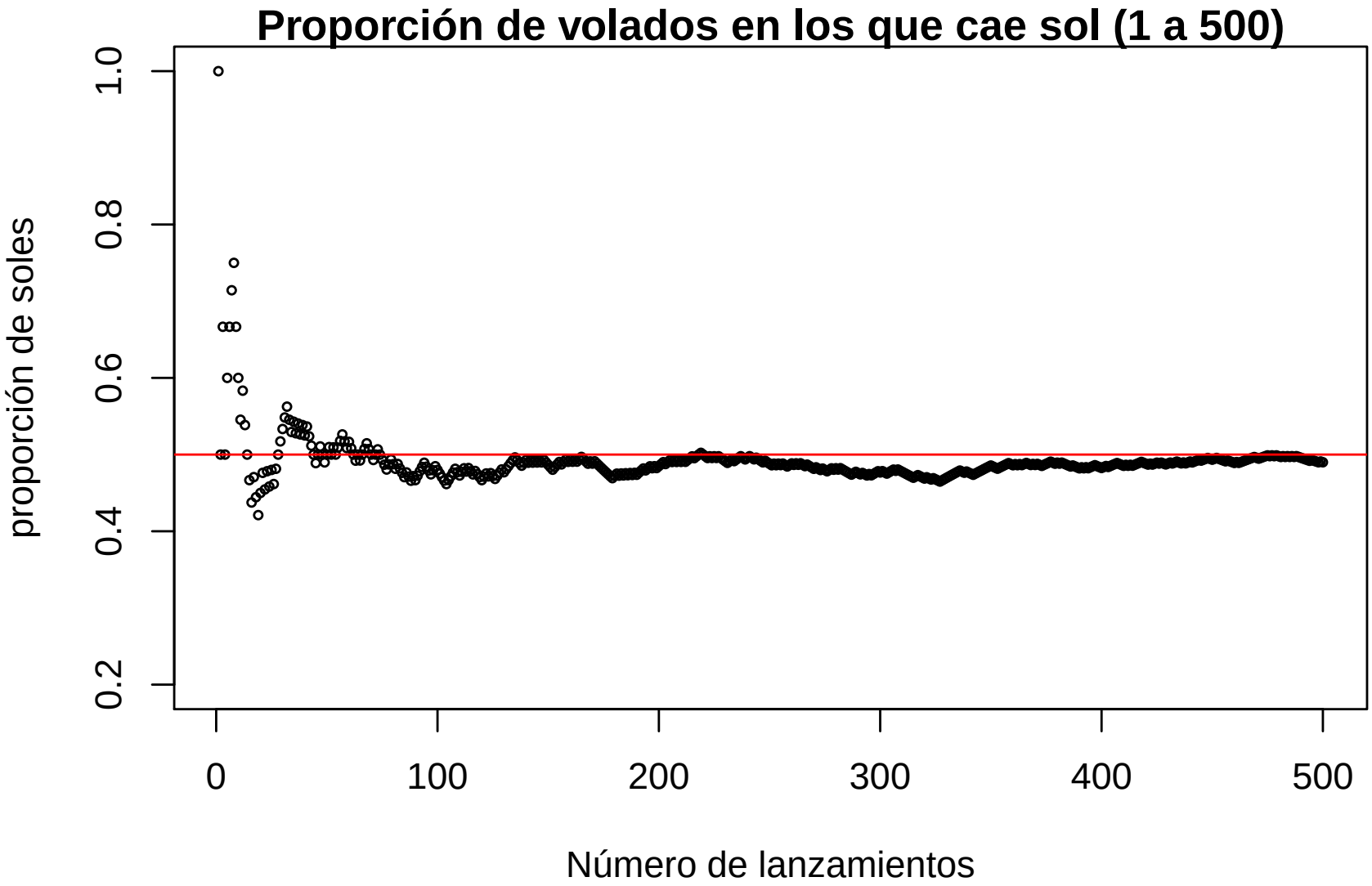
- Jakob Bernoulli escribió en 1705 (año de su muerte) el libro *Ars Conjectandi*, en el que desarrolló la teoría de la probabilidad.
- Bernoulli planteó el siguiente experimento ideal:

  En una gran urna hay 3,000 bolas negras y 2,000 blancas. Si se extraen bolas de la urna, anotando su color y devolviéndolas a la urna, puede afirmarse que cuanto mayor seal el número de extracciones más se aproximará a $2/3$ la proporción de bolas blancas extraídas.

- la afirmación anterior se conoce como la **ley de los grandes números** y es una de las bases de la estadística matemática.

Ejemplo de interpretación frecuentista

```
 [1] 1 0 1 0 1 1 1 1 0 0 0 1 0 0 0 0 1 0 0 1 1 0 1 0 1 0 1 1 1 1
```



Reglas de probabilidad I

Ejemplo: Lanzamiento de un dado

- Consideremos como experimento aleatorio el lanzamiento de un dado.
  - El **espacio muestral** de posibles resultados es el conjunto $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  - Los **eventos** son subconjuntos $A \subset \Omega$ Por ejemplo:
    $A =$ Resultado Par $= \{2, 4, 6\}$, $B =$ Resultado impar $= \{1, 3, 5\}$.
  - El conjunto de todos los eventos posibles del espacio muestral es el conjunto potencia de Ω, $\mathcal{P}(\Omega)$ Recuerden que si Ω tiene n elementos, el número de subconjuntos posibles es 2^n.
  - El conjunto vacío $\varnothing$ es el **evento imposible**. Por ejemplo $\varnothing = A \cap B$ es el resultado de que el dado sea par e impar a la vez.

Reglas de probabilidad II

Propiedades de la probabilidad

1. Cualquier evento tiene probabilidad entre 0 y 1. Además,
   - $P(\varnothing) = 0$,
   - $0 \leq P(A) \leq 1$ para $A \subseteq \Omega$, y
   - $P(\Omega) = 1$.
2. Si los resultados de un evento no pueden ocurrir simultáneamente (es decir, son disjuntos), la probabilidad de que ocurra cualquiera de ellos es la suma de las probabilidades:
$$\text{Si } A \cap B = \varnothing, P(A \cup B) = P(A) + P(B)$$
3. La probabilidad de que un resultado no ocurra es igual a uno menos la probabilidad de que el resultado ocurra:
$$P(A^c) = 1 - P(A)$$

Ejemplo

Ejemplo

- Consideremos un experimento aleatorio para investigar si hombres o mujeres son mas propensos a escoger un carro híbrido versus uno de combustión interna en una agencia de Honda. El Honda Civic está disponible en híbrido (H) o convencional (C).
- Se selecciona un cliente al azar que vaya a comprar un Honda Civic.
- El espacio muestral se puede representar como $\Omega = \{HH, HC, MH, MC\}$.
- Se puede pensar que no hay preferencias distintas entre hombres y mujeres, y a cada uno le da igual si el carro es híbrido o convencional. En ese caso la probabilidad de cada evento debería ser $1/4$.

Ejercicio de práctica

Una compañía constructora está trabajando en tres centros comerciales en sitios diferentes. Definan los eventos E_1, E_2 y E_3 como

$E_i =$ El centro en el Sitio i se terminará en la fecha establecida en el contrato

A través de un diagrama de Venn y representando los tres eventos como conjuntos, sombrar el área de cada uno de los siguientes eventos:

- Al menos un centro comercial se completa.
- Todos los centros comerciales se completan.
- Ninguno de los centros comerciales se completan.
- Sólo se completa el sitio 1 en la fecha establecida en el contrato.
- O el sitio 1 se termina a tiempo o ambos de los otros dos sitios se completan.

Independencia

El concepto de independencia es fundamental en diversos modelos de estadística.

Independencia

- Dos eventos son independientes si la posibilidad de que uno ocurra no se ve afectada por el conocimiento de que el otro ha ocurrido.
- Si hay más de dos eventos bajo consideración, serán independientes si el conocimiento de la ocurrencia de alguno de ellos no cambia las probabilidades de que cualquiera de los otros ocurra.

En términos de independencia, una cuarta propiedad es:

4. Regla de multiplicación: Si A y B son independientes,

$$P(A \cap B) = P(A)P(B)$$

Ejemplo de independencia

En los lanzamientos repetidos de una moneda 'honesta', de un dado, o de extracción de bolas de una urna *con reemplazo*, los eventos son independientes.

Probabilidad Condicional

El concepto de probabilidad condicional es fundamental en las aplicaciones estadísticas en general.

Probabilidad condicional

Cuando se tiene información adicional de que un evento B ya ocurrió, puede afectar el conocimiento de la ocurrencia de otro evento. A esta probabilidad se le llama **probabilidad condicional**.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{si } P(B) > 0$$

En términos de probabilidad condicional, si A y B son independientes,

$$P(A|B) = P(A).$$

Ejemplo de probabilidad condicional

El periódico *USA Today* del 6 de junio del 2000 dió información sobre el uso del cinturón de seguridad por género, basados en una encuesta a ciudadanos americanos:

	Hombres	Mujeres
Usa regularmente el cinturón de seguridad	0.10	0.175
No usa regularmente el cinturón de seguridad	0.40	0.325

- ¿Cuál es la probabilidad de que un adulto seleccionado al azar utilice regularmente el cinturón de seguridad?
- ¿Cuál es la probabilidad de que el adulto seleccionado utilice regularmente el cinturón dado que es un hombre?
- ¿Cuál es la probabilidad de que el sujeto seleccionado sea una mujer, dado que no usa el cinturón regularmente?

Mediciones estadísticas

Los adultos aman las cifras. Cuando les dices que tienes un nuevo amigo, ellos nunca preguntan sobre los temas esenciales. Ellos nunca te dicen: "¿Cómo suena su voz? ¿Cuáles son los juegos que le apasionan? ¿Colecciona mariposas?" En lugar de eso demandan: "¿Cuántos años tiene? ¿Cuántos hermanos? ¿Cuánto pesa? ¿Cuánto ganan sus padres?" Sólo de estas cifras ellos piensan que han aprendido algo sobre tu amigo.

El principito.

Mediciones estadísticas I

- Antoine Lavoisier participó en la definición del sistema métrico decimal y en 1791 escribió un informe titulado *Resumen de diversas obras de aritmética política*.
- La nueva república francesa necesitaba este informe ya que los impuestos se basaban en la propiedad, en la tierra realmente cultivada y en el ganado sujeto a recaudación.
- Lavoisier intentó calcular cuánta tierra de labor había en Francia, recopilando datos sobre el consumo anual de alimentos y alcohol, tanto en las ciudades como en los pueblos y luego calculó cuánta tierra era necesaria para producir todos esos alimentos y bebidas.
- Gracias a su estudio, se sabe que en 1790 Francia tenía 25 millones de habitantes, de los que 8 millones vivían en ciudades y de otros 8 millones que trabajaban en la viticultura.
- Lavoisier solicitó la creación de una oficina estadística para realizar registros regulares sobre la agricultura, comercio y población. De aquí se derivan los institutos de estadística, como el INEGI.

Mediciones estadísticas II

- Hay muchas preguntas que se pueden plantear que dan origen a la necesidad de medir cantidades que pueden considerarse inciertas:
  - ¿Cómo se distribuye el ingreso en México?
  - ¿Cuál es el peso de los mexicanos?
  - ¿Cuál es la proporción de hombres y mujeres?
  - ...

Variables Aleatorias I

Variable aleatoria

Una **variable aleatoria** es una característica que cambia dependiendo de sujeto a sujeto en un estudio o experimento. Por ejemplo, en una encuesta, respuestas a las siguientes preguntas se pueden considerar como variables:

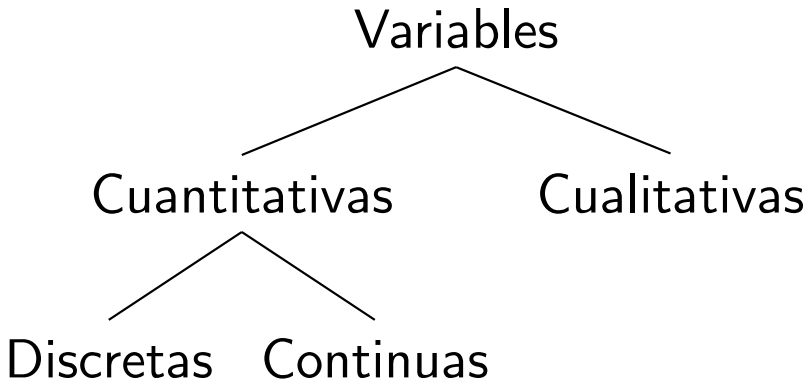
- ¿Cuál es su edad?
- ¿Cuántas personas forman su familia?
- ¿Cuál es su ingreso?
- ¿Cuál es el ingreso total de la familia?
- ¿Está casado(a)?
- ¿Tiene trabajo?

Formalmente, una variable aleatoria es una función del espacio muestral a los números reales (o un subconjunto de éstos):

$$X : \Omega \to \mathbb{R}$$

Tipos de variables

Las variables se pueden clasificar de la siguiente manera:



Ejemplos

- $X_i = \begin{cases} 1 & \text{si suben las tasas en el día } i \\ 0 & \text{si no suben las tasas en el día } i \end{cases}$
- $X =$ Número de clientes que usan su reservación en un vuelo de n asientos.
- $Y =$ Edad de personas que asisten al curso propedéutico de la especialidad "Divulgación en la Economía".
- $Z =$ ingreso de las personas residentes en México mayores de 65 años
- $\pi_t =$ inflación a diciembre del año t.
- $O =$ ocupación de la persona que aporta el principal ingreso en un hogar.

Ejemplos de variables aleatorias: modelo Bernoulli I

Estas variables son las más simples que se presentan en la práctica.

variables Bernoulli

- Son variables que sólo puede tomar dos valores: 0 o 1 (si o no; éxito o fracaso).
- $P(X = 1) = p$, $P(X = 0) = 1 - p$.
- Notación: $X \sim Ber(p)$ se lee: X tiene una distribución de Bernoulli con probabilidad de éxito igual a p.

Ejemplos de variables aleatorias: modelo Bernoulli II

Ejemplos de variables aleatorias Bernoulli

- En un procedimiento de auditoría, se puede clasificar un documento como Normal (0) o Anómalo (1).
- $X = 1$ si el volado da un sol y $X = 0$ si es un águila.
- En la revisión de productos, $X = 1$ si el artículo es defectuoso y $X = 0$ si no tiene defectos.
- Para un banco una persona puede o no se considerada para otorgarle un crédito.

Variables Binomiales I

- Si consideramos por ejemplo un foco, este puede tener o no defecto. ¿Qué pasa si consideramos un lote de 20 focos y extraemos 3 de ellos para determinar si el lote está en buenas condiciones?
- ¿De cuántas maneras se pueden extraer tres focos del lote de 20? Este es el número de subconjuntos de tamaño 3 de un conjunto con 20 elementos.
- Recordemos que el número de subconjuntos de tamaño x de un conjunto de n elementos es $\binom{n}{x}$.
- $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ indica las combinaciones de x elementos de un total de n elementos, y $n! = n(n-1)(n-2)\cdots 1$ es el **factorial**. Por definición $0! = 1$

Variables Binomiales II

Variables aleatorias Binomiales

Una variable binomial X se asocia a un experimento que consiste en la realización de n ensayos **idénticos e independientes**.

- Se puede considerar como la suma de n variables Bernoulli.
- Notación: $X \sim Bin(n, p)$, donde p es la probabilidad de éxito.
- La función de probabilidad es:

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x \in \{0, 1, 2, \ldots, n\}$$

- Otra función importante es la función de distribución acumulativa, que calcula la probabilidad acumulada hasta un valor x:

$$F(x) = P(X \leq x) = \sum_{i=0}^{x} P(X = i)$$

- La media de la variable es np y su varianza $np(1-p)$.

Ejemplos de variables aleatorias: modelo Binomial I

- Basado en registros históricos, la tienda Zara estima que la probabilidad de que un cliente que entre a su tienda compran un artículo con probabilidad 0.3 de manera independiente. Si a la tienda entran 150 clientes en un día, ¿Cuál es la probabilidad de que vendan no más 45 artículos?
- ¿Más de 45 artículos?
- ¿Entre 20 y 30 artículos?
- ¿Cuántos artículos en promedio se venden ese día?

Modelo binomial para precios

- Se puede considerar que de un periodo a otro, un precio puede subir con probabilidad p un peso y disminuir con probabilidad $1-p$ un peso. Entonces el precio después de n periodos es una variable binomial.
-

Medidas de tendencia central I

- Cuando se describen datos numéricos, es común reportar un valor que sea representativo de los datos. Ese número describe dónde aproximadamente están localizados los datos y se llama una **medida de tendencia central**. Los tres números más comunes son la **media**, la **mediana** y la **moda**.

Medidas de tendencia central

- La **media muestral** consiste de las observaciones $x_1, x_2, \ldots, x_n$ se define como

$$\bar{x} = \frac{\sum x_i}{n}$$

- La **mediana muestral** se obtiene ordenando primero las n observaciones de la menor a la mayor (incluyendo valores repetidos). Entonces la mediana es:

$$\text{mediana} = \begin{cases} \text{el valor de enmedio, si } n \text{ es impar} \\ \text{el promedio de los dos valores de enmedio, si } n \text{ es par} \end{cases}$$

- La **moda muestral** es el valor que más se repite en el conjunto de datos.

Medidas de tendencia central II

Ejemplo

El número de accesos por hora a la página web del MIDE se midió durante 40 horas repartidos en varios días y se obtuvo la siguiente muestra:

0 0 0 0 0 0 3 4 4 4 5 5 7 7 8 8 8 12 12 13 13 13 14 14
16 18 19 19 20 20 21 22 23 26 36 36 37 42 84 331

Obtener la media, la mediana y la moda.

Medidas de dispersión I

- Medida de dispersión: ¿qué tan diferentes son los datos?

Medidas de dispersión

- El rango de un conjunto de datos se define como la diferencia entre la observación más grande y la más chica.
- La varianza mide la suma de las desviaciones cuadráticas de cada dato con la media de los datos y se divide entre el número de observaciones menos 1:

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

La desviación estándar es la raíz cuadrada de la varianza: $s = \sqrt{s^2}$.

Ejemplo: El índice Big Mac compara el costo de una hamburguesa Big Mac en varias partes del mundo.

Medidas de dispersión II

País	Precio de Big Mac en USD
Argentina	3.02
Brasil	4.67
Chile	3.28
Colombia	3.51
Costa Rica	3.42
México	2.39
Perú	2.76
Uruguay	2.87

Mediciones poblacionales

- Las medidas de tendencia central y de dispersión basadas en muestra tienen su equivalente cuando se considera toda la población, y para representarlas usualmente se utilizan letras griegas.
- Por ejemplo, la media población se denota usualmente por μ y la desviación estándar de la población se denota por σ.

Variables aleatorias Normales

- Muchas variables aleatorias continuas tienen un comportamiento 'normal'.
- Por ejemplo, consideremos una población de bebes nacidos a término completo en un hospital en un cierto año y la variable de interés es $X =$ peso del bebé. Supongamos que nacieron 2000 bebés, de los que se muestran los pesos de los primeros 20:

```
 [1] 2.75 3.07 2.96 3.44 3.06 3.16 2.71 3.36 2.59 2.82 3.04 3.05 2.90 3.37
[15] 3.06 2.99 2.81 3.26 2.54 4.16
```

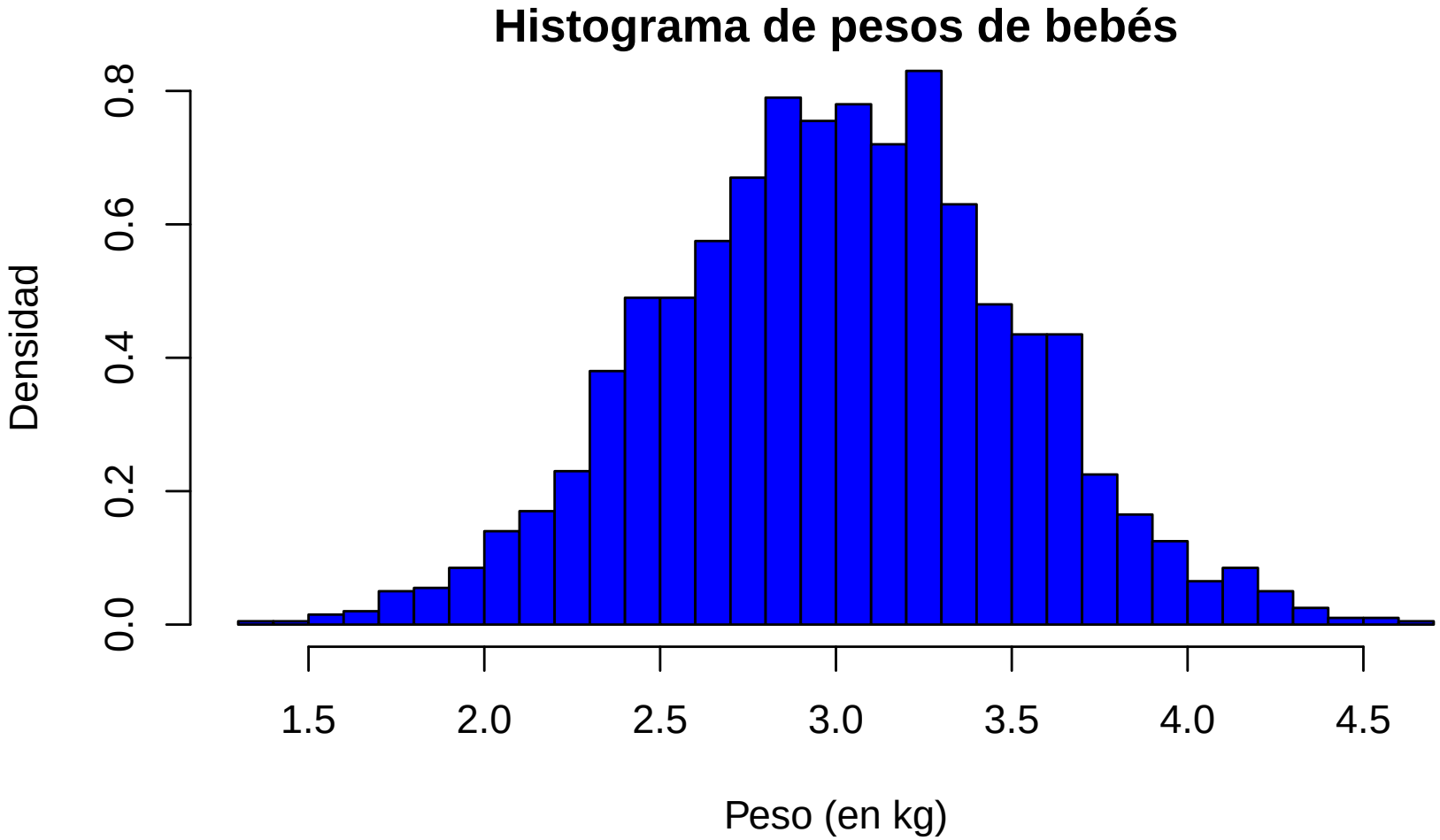
- Una forma de describir cómo se distribuyen estos datos es agrupando los datos por intervalos y obteniendo su frecuencia relativa. Las frecuencias relativas se utilizan para calcular la **densidad** de los datos:

$$\text{densidad} = \frac{\text{frecuencia relativa}}{\text{longitud del intervalo}}$$

Por ejemplo, entre 2 y 2.20 kilos se observaron un total de 62 bebes, así que la densidad es $\frac{62/2000}{2.20-2} = \frac{0.031}{0.20} = 0.155$

Variables aleatorias Normales

Al graficar los valores de la densidad en sus respectivos intervalos, se obtiene un *histograma* de los datos que nos da una idea de cómo se distribuyen.
El peso medio de los bebés es de $3.01kg$ y su varianza de $0.25kg^2$. La desviación estándar tienen las mismas unidades que la variable original: $0.5kg$s.



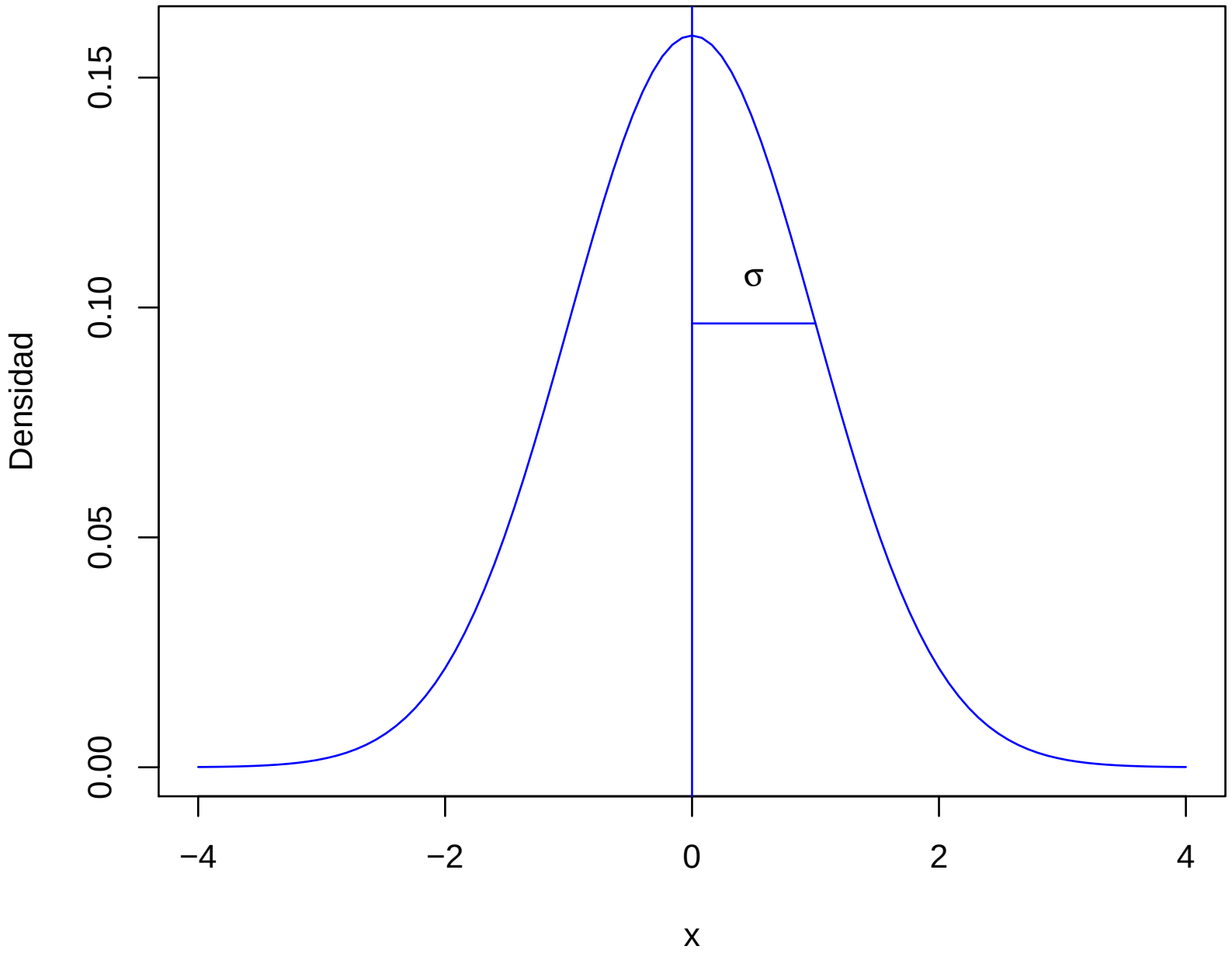
Distribución normal I

- En Econometría y Economía, uno de los tipos de variables aleatorias más usados es la variable aleatoria normal.
- El supuesto de que una variable aleatoria es normal simplifica el cálculo de probabilidades.
- Una variable aleatoria normal puede tomar cualquier valor en los reales. Su función de densidad es de la forma:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

donde μ y σ^2 son parámetros para la media y varianza como veremos más adelante.

Distribución normal II



Relación entre área y frecuencia relativa I

- El área de cualquier rectángulo en un histograma se puede interpretar como la probabilidad de observar un valor de la variable en el intervalo correspondiente:

$$\text{densidad} = \frac{\text{frecuencia relativa}}{\text{longitud del intervalo}}$$

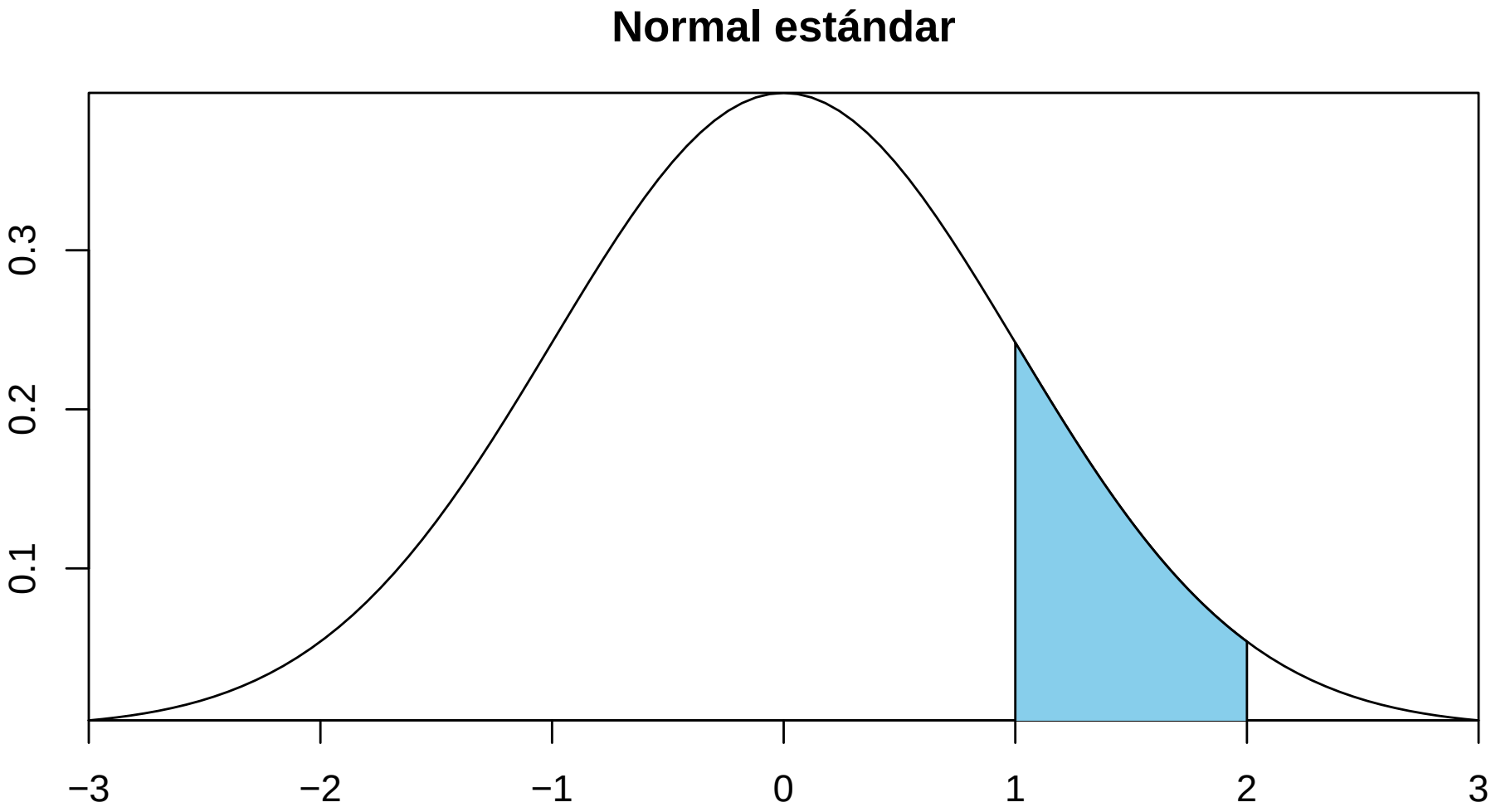
Como el rectángulo correspondiente a un intervalo tiene altura igual a la densidad, el área del rectángulo es:

$$\begin{aligned} \text{area} &= (\text{altura})(\text{longitud del intervalo}) \\ &= (\text{densidad})(\text{longitud del intervalo}) \\ &= \left(\frac{\text{frecuencia relativa}}{\text{longitud del intervalo}}\right)(\text{longitud del intervalo}) \\ &= \text{frecuencia relativa} \end{aligned}$$

- Lo anterior significa que el área del rectángulo arriba de cada intervalo es igual a la frecuencia relativa de los valores que caen en el intervalo.
- De esta forma podemos calcular la probabilidad de cualquier intervalo de una distribución normal.

Relación entre área y frecuencia relativa I

Por ejemplo, para calcular $P(1 < X < 2)$ en una distribución normal estándar, se calcula el área debajo de la curva entre 1 y 2:



Cálculo de probabilidades en una distribución normal

Distribución normal estándar

Si X es una variable aleatoria normal con media μ y varianza σ^2, entonces $Z = \frac{X-\mu}{\sigma}$ es una variable aleatoria con media 0 y varianza 1. Z sigue una distribución **normal estándar**

Ejemplos

Calcular las siguientes probabilidades: (Z es normal estándar):

- $P(-1 < Z < 1)$
- $P(-2 < Z < 2)$
- $P(-3 < Z < 3)$
- $P(X > 2)$ si X es normal con media 5 y varianza 2
- $P(-1 < X < 2)$ si X es normal con media 3 y varianza 1

Inferencia Estadística I

- Un problema muy común en economía es la inferencia de hipótesis. Consideremos un ejemplo.

Ejemplo de inferencia

En una encuesta sobre desempleo que se realiza en una ciudad de 8 millones de habitantes, se elige una muestra al azar de 2,000 personas. De éstas 700 se declaran desempleadas. ¿Se puede inferir algo sobre el índice de desempleo de la zona? Lo que se desea es obtener un resultado con un *nivel de confianza* del 95 % (o equivalentemente, con un *nivel de significancia* del 5 %).

- De la muestra obtenida se puede estimar la proporción de desempleados: $\hat{p} = \frac{700}{2000} = 0.35$ pero este estimador, por ser el de una muestra y no el de la población, cuenta con un margen de error.

Propiedades generales de la distribución muestral de $\hat{p}$

Sea $\hat{p}$ la proporción de éxitos en una muestra aleatoria de tamaño n de una población cuya proporción de éxitos es p. Denotemos la media de $\hat{p}$ como $\mu_{\hat{p}}$ y su desviación estándar como $\sigma_{\hat{p}}$. Entonces se cumplen las siguientes reglas para la estimación de una proporción:

1. $\mu_{\hat{p}} = p$
2. $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
3. Cuando n es grande, y p no está muy cerca de 0 o de 1, la distribución de $\hat{p}$ es aproximadamente normal.

Entonces un intervalo de confianza del 95 % para el verdadero valor del parámetro p esta dado por:

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Inferencia Estadística III

Entonces para nuestro ejercicio, un intervalo del 95 % de confianza para el índice de desempleo es:

$$\left(0.35 - 1.96\sqrt{\frac{0.35(0.65)}{2000}}, 0.35 + 1.96\sqrt{\frac{0.35(0.65)}{2000}}\right) = (0.33, 0.37)$$

Módulo 11. Herramientas Estadísticas II: Gráficas y Regresión Lineal

Tipos de gráficas estadísticas I

- Las gráficas pueden ser fundamentales para poner orden en el caos del exceso numérico.
- Cada tipo de dato tiene una mejor manera de representarse gráficamente. Entre las gráficas estadísticas más comunes se encuentran las siguientes.
- Para datos de tipo cualitativo, en donde se representan frecuencias:
  - Gráficas de barras
  - Gráficas de pie
  - Gráficas de puntos (dotplots)
  - Rama y hoja (stem and leaf)
- Para datos cuantitativos ya sean continuos o discretos
  - histogramas
  - boxplots
  - lineas
- Para dos o más variables:
  - Gráficas de dispersión (scatterplots)
  - Gráficas de burbujas
  - Series de tiempo
  - Gráficas de mosaico

Tipos de gráficas estadísticas II

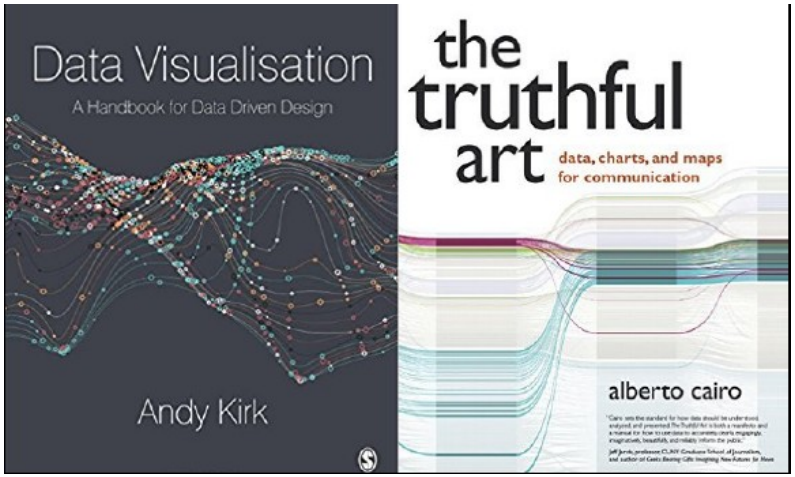
  - Gráficas trellis

- Sin embargo, en la actualidad hay muchos tipos de gráficas, tanto estáticas como dinámicas, que sirven para visualizar información (aunque no siempre de manera efectiva).

Recursos en temas de visualización y gráficas estadísticas I

- Curso gratuito de Alberto Cairo en Journalismcourses.org

- Libros de Alberto Cairo y Andy Kirk sobre datos y visualización:



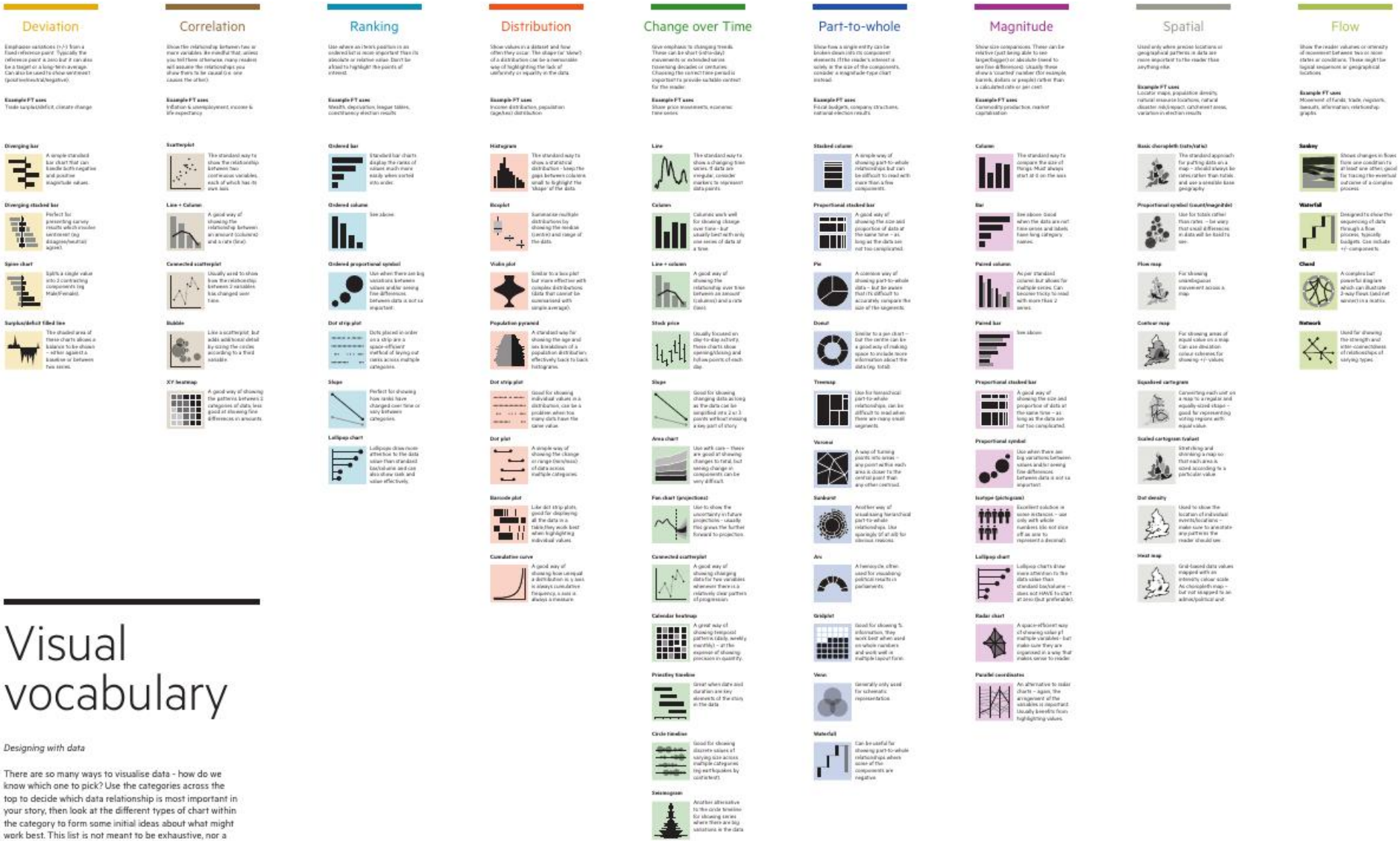
Recursos en temas de visualización y gráficas estadísticas II

- Libros sobre gráficas estadísticas

- Videos de Hans Roling en TED
- Gapminder.org
- Portal BID: Números para el desarrollo

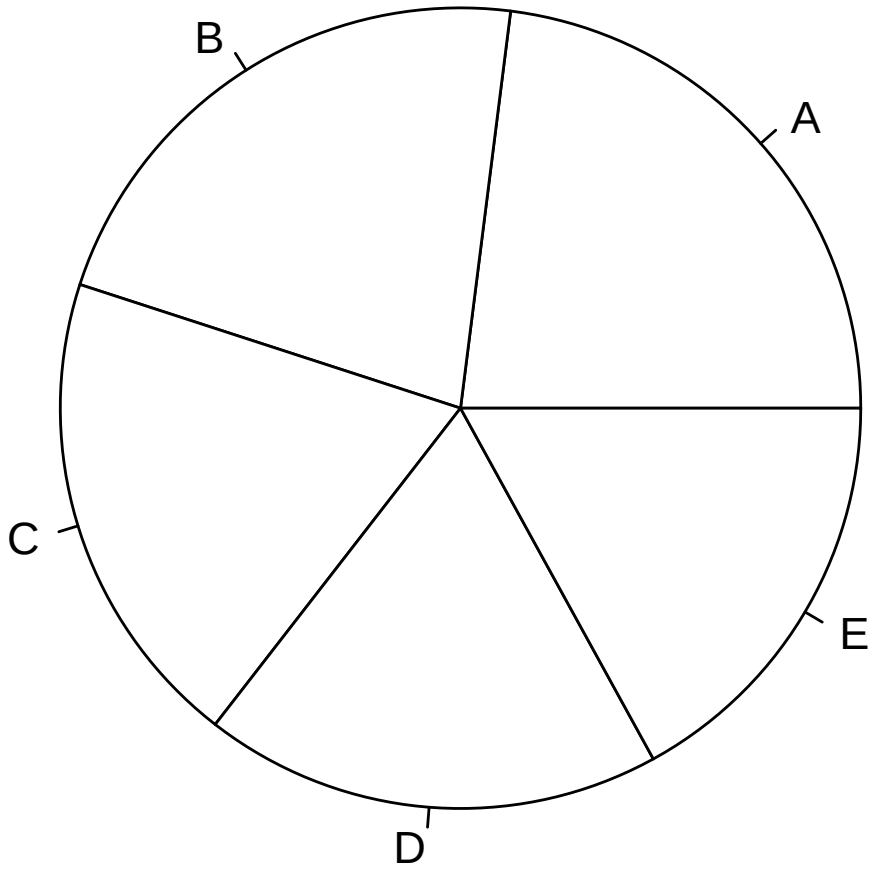
Métodos gráficos

Consultar: Guía gráfica del FT



Gráficas efectivas

Consideren la siguiente gráfica de pie con 5 segmentos. Traten de colocar las etiquetas de los segmentos del mayor al menor.



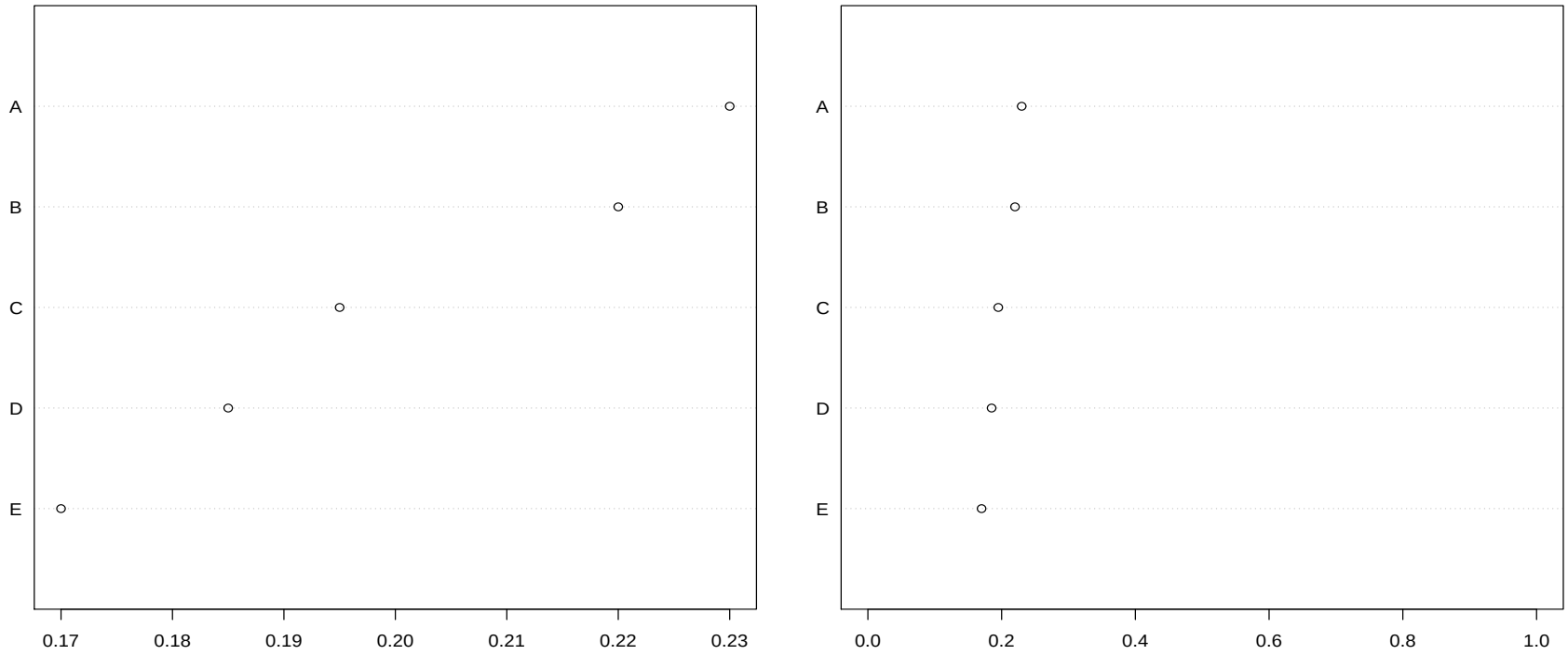
Gráficas efectivas

- La mayoría de las personas tienen problemas para hacer la correcta determinación del orden de las áreas.
- La tarea puede ser mucho más fácil en otros tipos de graficas
- William Cleveland (1984) introdujo las gráficas de puntos basado en resultados de percepción humana y de cómo se realiza la decodificación de información gráfica.

Gráficas efectivas

Una gráfica es más **efectiva** que otra si la información cuantitativa se puede decodificar más fácilmente para la mayoría de sus observadores.
Esta definición supone que la razón por la que se utilizan las gráficas es comunicar información.
Una gráfica efectiva cuenta una historia sobre los datos que representa.

Los mismos datos en gráfica de puntos



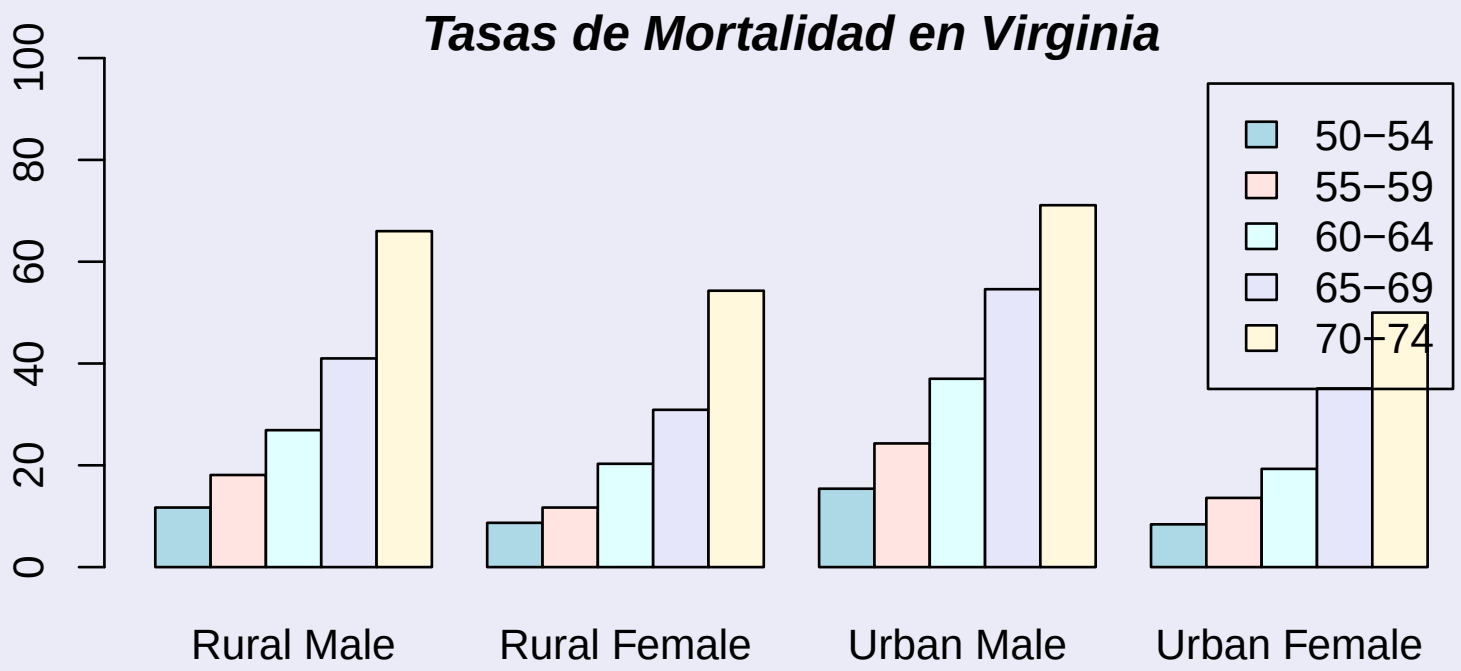
Los datos mostrados son los de la siguiente tabla:

A	23.0 %
B	22.0 %
C	19.5 %
D	18.5 %
E	17.0 %

Gráfica de barras

- Cuándo se usa: datos categóricos.
- Ejemplo: La siguiente tabla muestra las tasas de mortalidad en Virginia para diferentes grupos de personas

	Rural Male	Rural Female	Urban Male	Urban Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0



Una gráfica de barras convertida en infografía

La siguiente gráfica reemplaza las barras por cubetas de leche, pero también distorsiona las áreas. Las dos cubetas de 1980 representan 32 vacas mientras que el de 1970 representa 19 (y 32 no es el doble de 19).

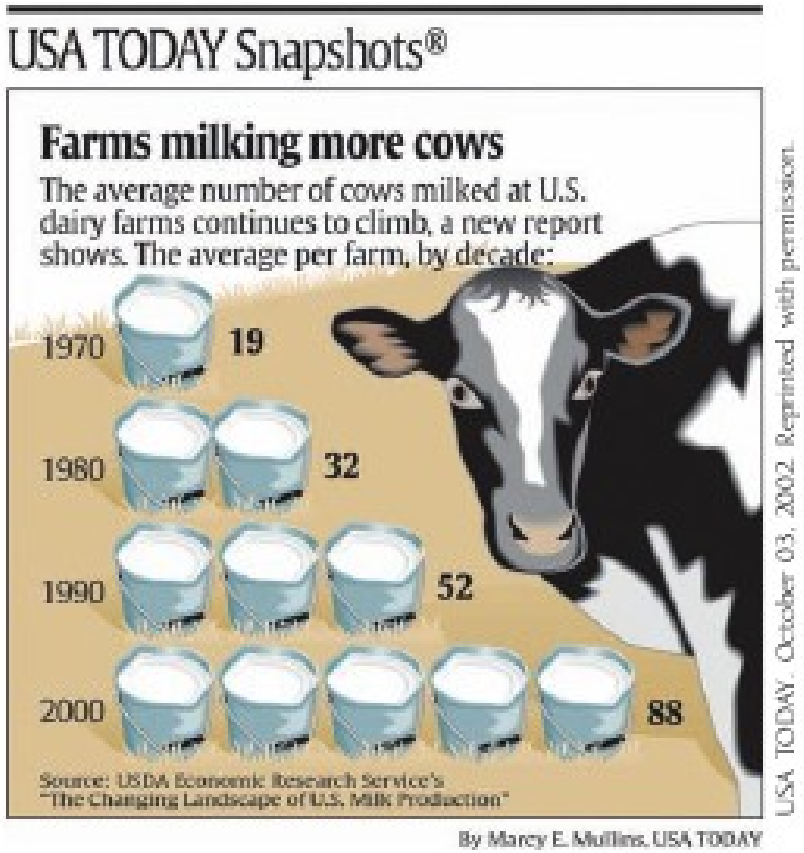
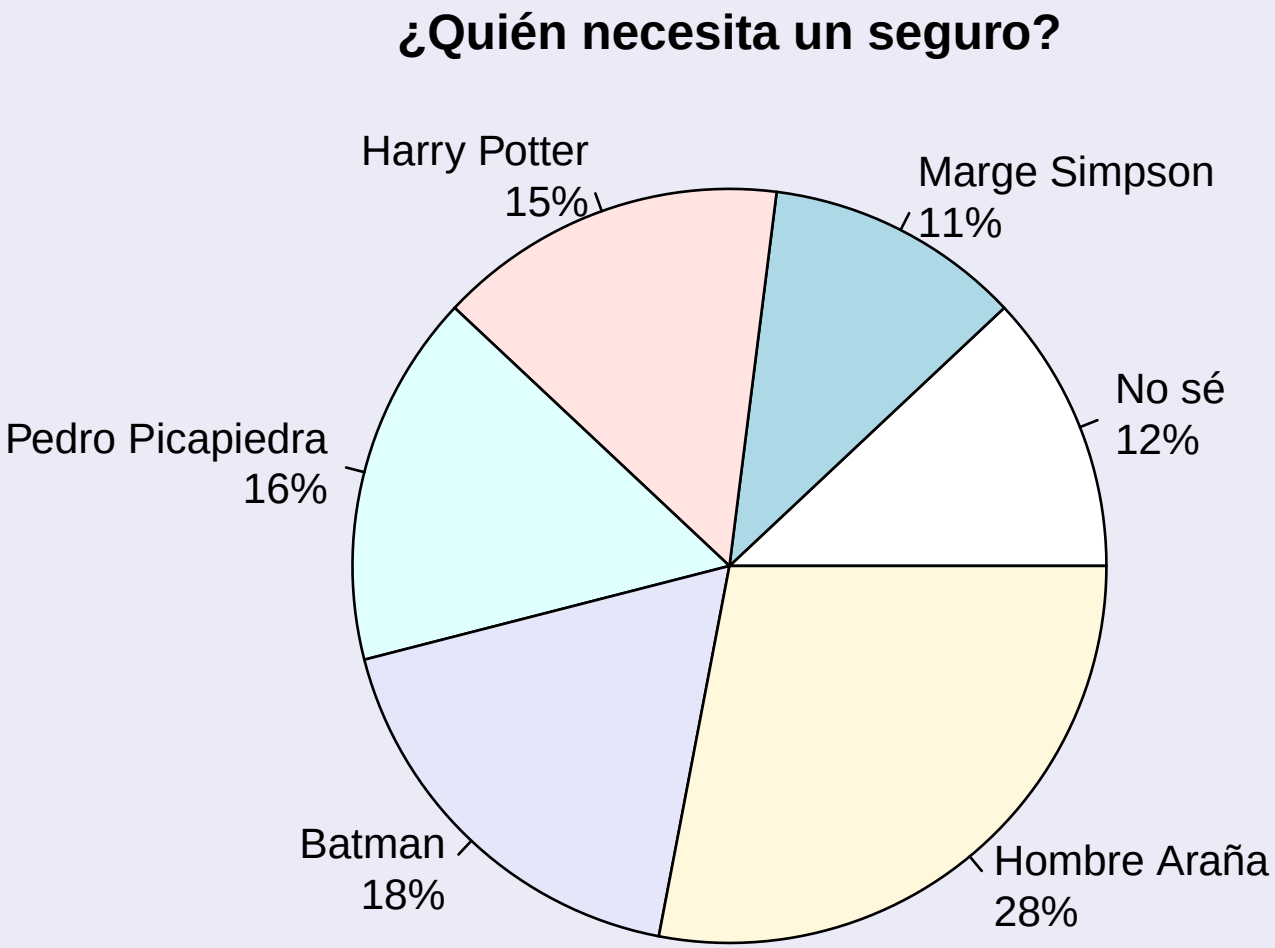


Figura: **Fuente: USA Today, Septiembre 17, 2009**

Gráficas de pie

- Cuando se usa: Las gráficas de pie son más efectivas cuando hay pocas categorías en los datos y si los datos no son muy importantes.
- Ejemplo:



Otro ejemplo de gráfica de pie

Gráfica de rama y hoja (stem and leaf)

- Cuando se usa: Conjuntos de datos numéricos con un número moderado de observaciones (no trabaja bien para conjuntos de datos grandes).
- Ejemplo:

```
 [1] 152  89 109  88 106  88 105  87 102  87 199  86 198  85 196  84 195
[18]  84 194  84 194  79 193  78 191  78 190  77 190  76 190  76 190  76
[35] 190  75 189  67 189  43

  The decimal point is 1 digit(s) to the right of the |

   4 | 3
   5 |
   6 | 7
   7 | 56667889
   8 | 4445677889
   9 |
  10 | 2569
  11 |
  12 |
  13 |
  14 |
  15 | 2
  16 |
  17 |
  18 | 99
  19 | 0000013445689
```

Gráfica de rama y hoja (stem and leaf)

Lo que hay que mirar en este tipo de gráfica:

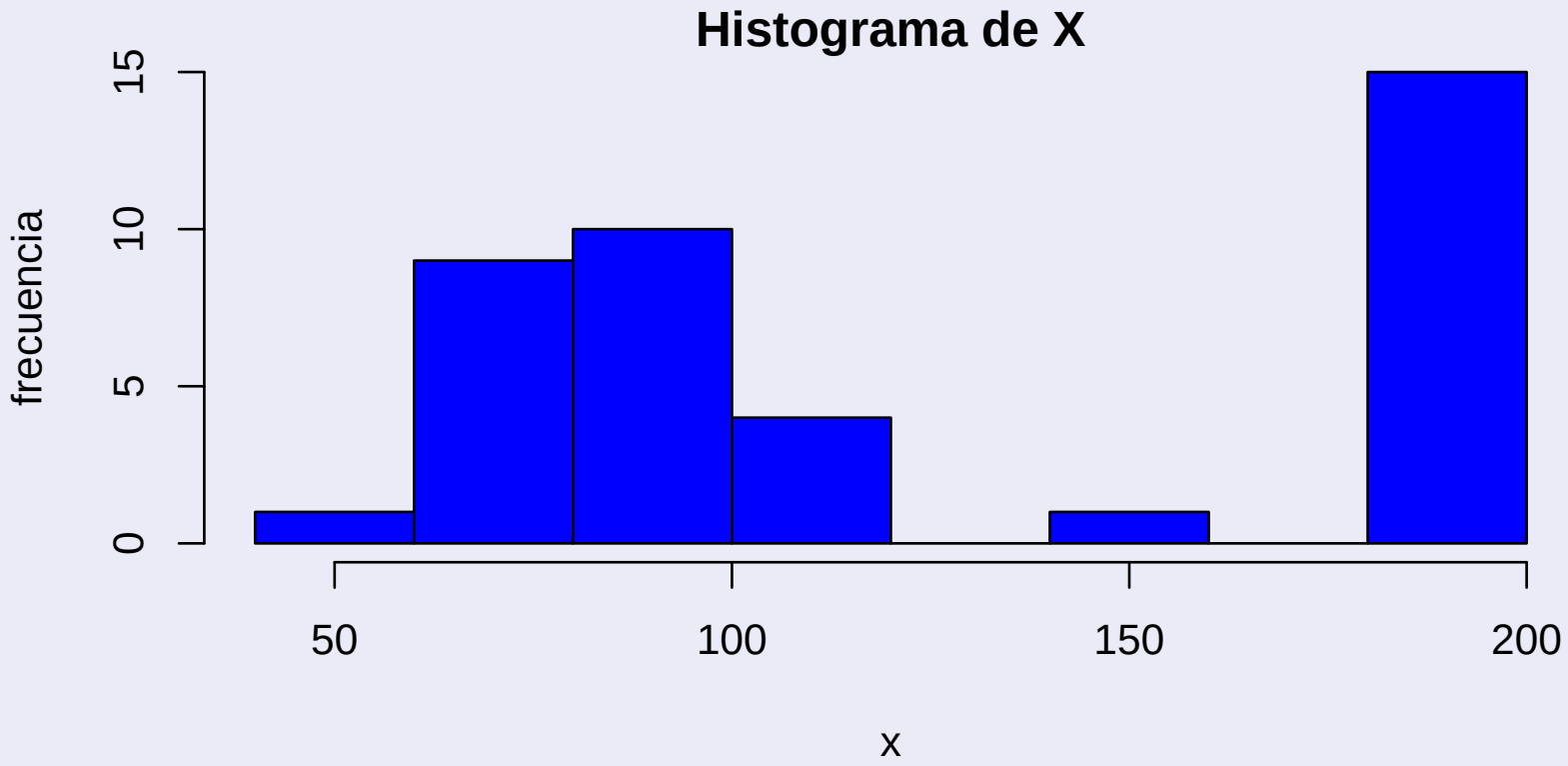
- Un valor típico o representativo en el conjunto de datos
- La extensión de la dispersión alrededor de un valor típico
- La presencia de huecos en los datos
- La extensión de la asimetría en la distribución de los datos.
- El número y ubicación de los picos.

Otras alternativas:
En lugar de que el tallo sean decenas y las hojas unidades, también se puede cambiar a otras escalas para simplificar la estructura: el tallo pueden ser miles y las hojas centenas, etc.

Histogramas

- Cuando se usa: Datos numéricos discretos o continuos (se pueden usar para gran cantidad de datos).
- Ejemplo:

```
 [1] 152  89 109  88 106  88 105  87 102  87 199  86 198  85 196  84 195
[18]  84 194  84 194  79 193  78 191  78 190  77 190  76 190  76 190  76
[35] 190  75 189  67 189  43
```



Histogramas I

Lo que hay que mirar en este tipo de gráfica:

- Centro o valor típico (unimodal o multimodal)
- Extensión de dispersión o variabilidad
- la forma de la distribución (sesgada, simétrica)
- ubicación y número de picos
- Presencia de huecos y valores atípicos o extremos (colas anchas, colas delgadas).

Otras consideraciones:

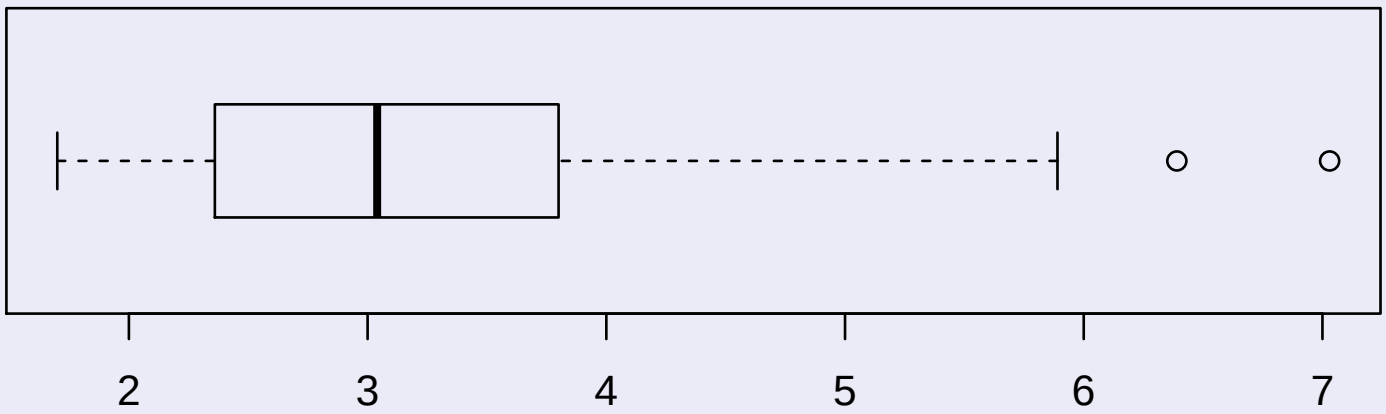
- Se pueden graficar frecuencias, frecuencias relativas o densidad.
- También se pueden graficar frecuencias acumuladas (función de distribución)
- Las barras no tienen que ser del mismo ancho, pueden tener anchos variables.

Gráficas de caja (boxplot)

Una gráfica de caja o boxplot es una versión simplificada de un histograma que provee información sobre el centro, la dispersión y simetría o sesgo de los datos, así como valores extremos. También permite comparar diferentes poblaciones de manera directa.

- Cuando se usa: Cuando se requiere representar los datos de manera simplificada sin conocer necesariamente la forma de la distribución, sólo sus características más generales. Se utiliza para datos cuantitativos y usualmente para comparar muestras de diferentes poblaciones.
- Ejemplo:

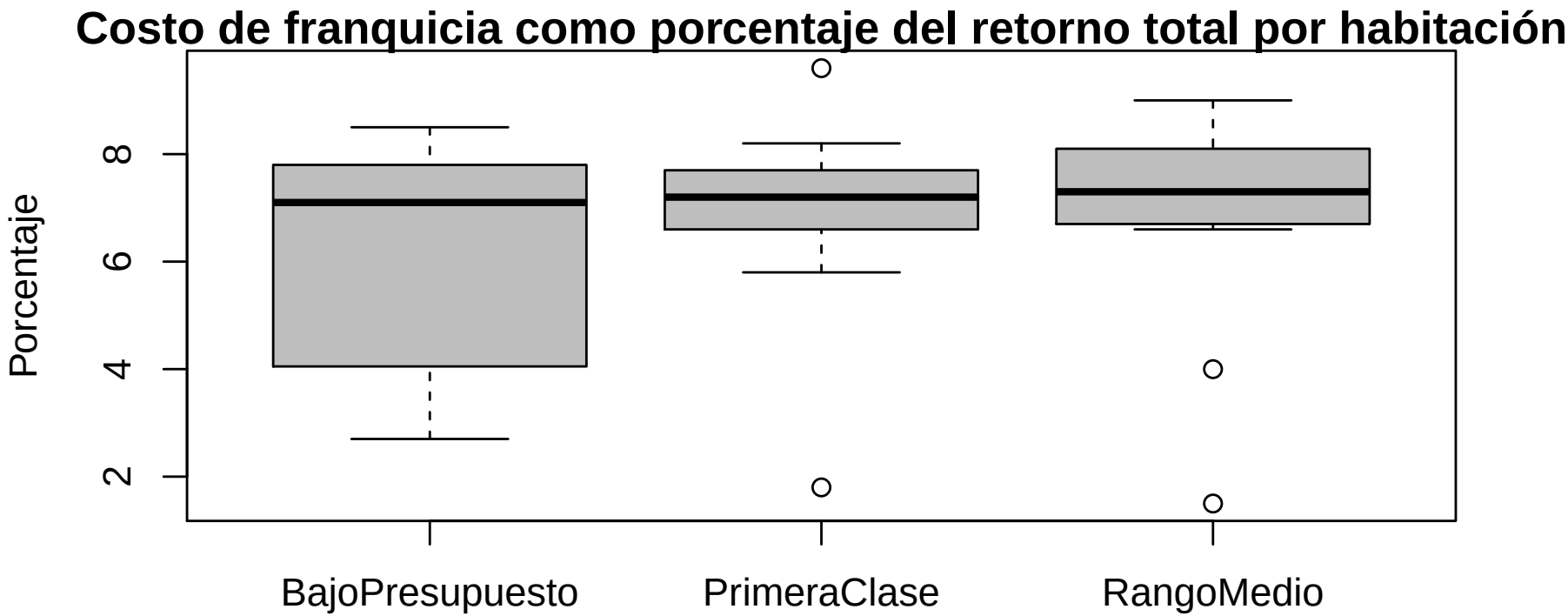
```
 [1] 3.57 3.01 3.97 4.67 3.80 3.64 3.28 1.83 3.51 3.42 3.92 5.89 3.04 2.36
[15] 4.92 1.72 3.89 5.20 2.21 3.98 3.54 3.24 3.06 1.99 2.48 3.54 7.03 2.28
[29] 2.76 2.09 2.66 2.31 2.93 3.03 2.37 2.91 1.83 5.57 6.39 2.31 1.93 3.80
[43] 2.72 1.70 2.87
```



Boxplot: comparación de poblaciones

En la siguiente gráfica se comparan varias poblaciones: Los siguientes datos comparan precios entre hoteles de bajo presupuesto (20 hoteles), de rango medio (14 hoteles) y de primera clase (15 hoteles). Los datos representan costo de la franquicia como proporción del retorno total por habitación.

```
              Tipo porcentaje
45    PrimeraClase        7.6
17 BajoPresupuesto        7.9
20 BajoPresupuesto        8.5
40    PrimeraClase        6.6
2  BajoPresupuesto        2.8
6  BajoPresupuesto        4.1
```

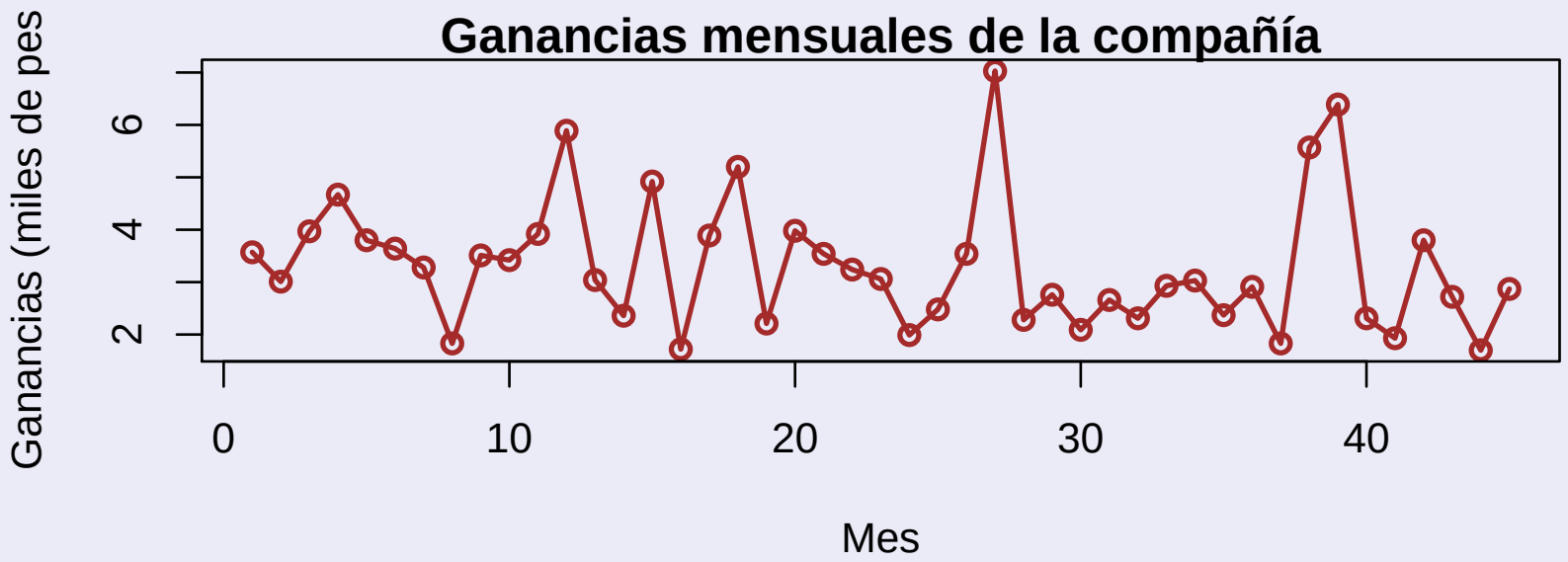


Gráficas de línea

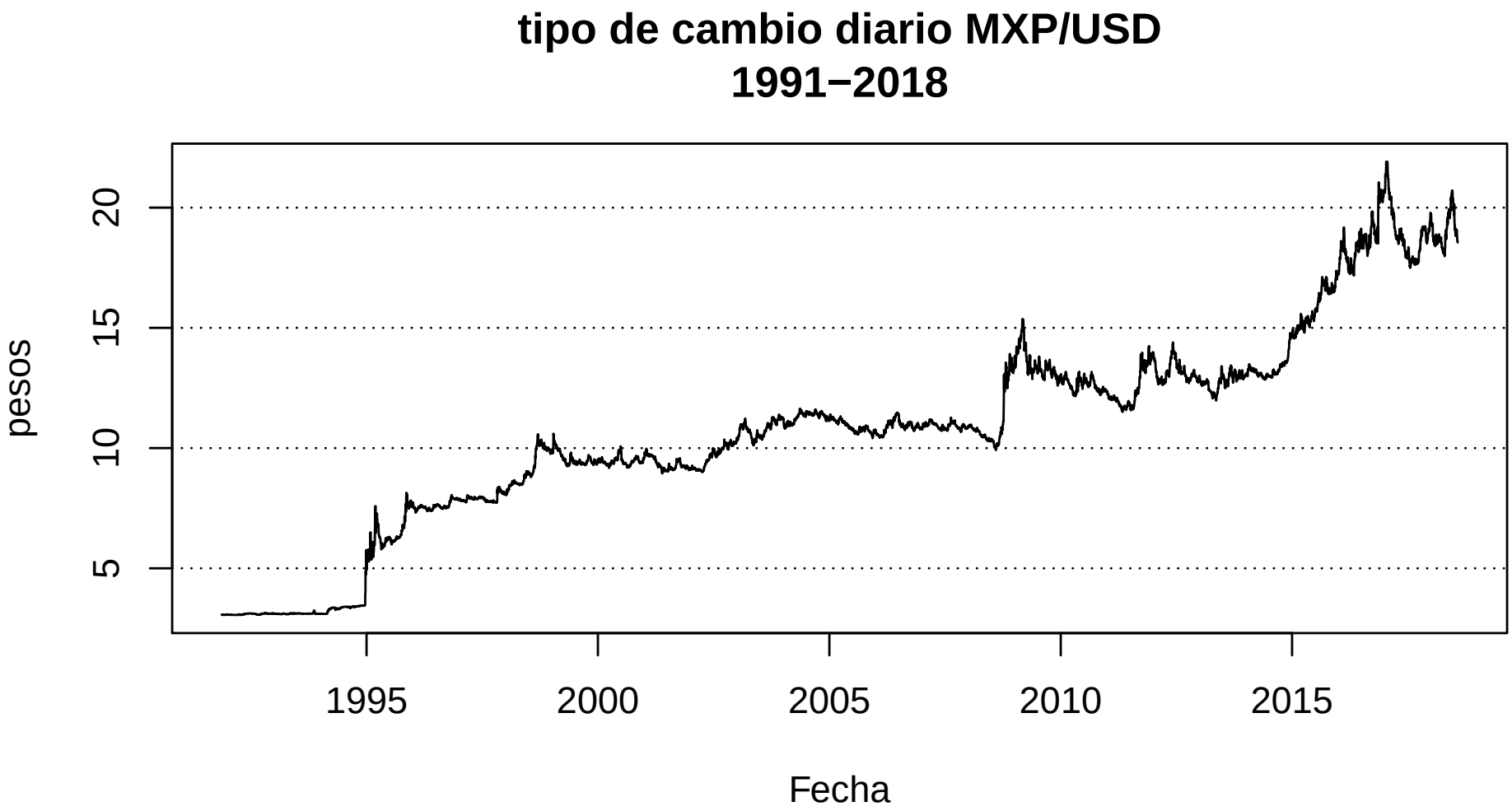
Uno de los tipos de datos más comunes en Economía son las series de tiempo. Son mediciones que tienen una dependencia del tiempo.

- Cuando se usa: Cuando los datos corresponden a una serie de tiempo y los datos tienen una dependencia temporal
- Ejemplo: Los siguientes datos corresponden a ganancias de una empresa (en miles) en cada mes

```
 [1] 3.57 3.01 3.97 4.67 3.80 3.64 3.28 1.83 3.51 3.42 3.92 5.89 3.04 2.36
[15] 4.92 1.72 3.89 5.20 2.21 3.98 3.54 3.24 3.06 1.99 2.48 3.54 7.03 2.28
[29] 2.76 2.09 2.66 2.31 2.93 3.03 2.37 2.91 1.83 5.57 6.39 2.31 1.93 3.80
[43] 2.72 1.70 2.87
```



Gráficas de linea: series de tiempo

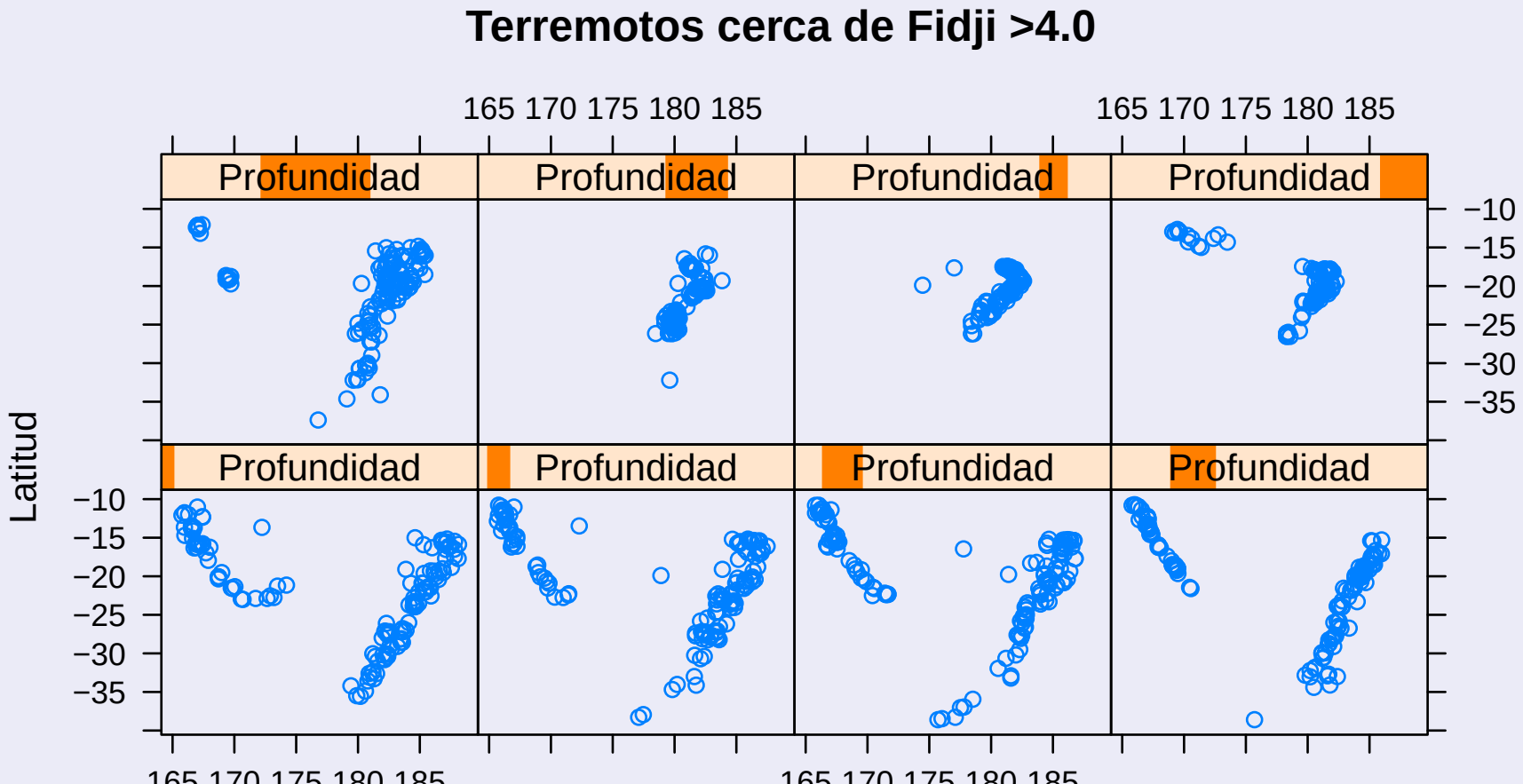


Una fuente relevante para series de tiempo financieras y económicas globales es Quandl.

Gráficas trellis (enrejado o retícula)

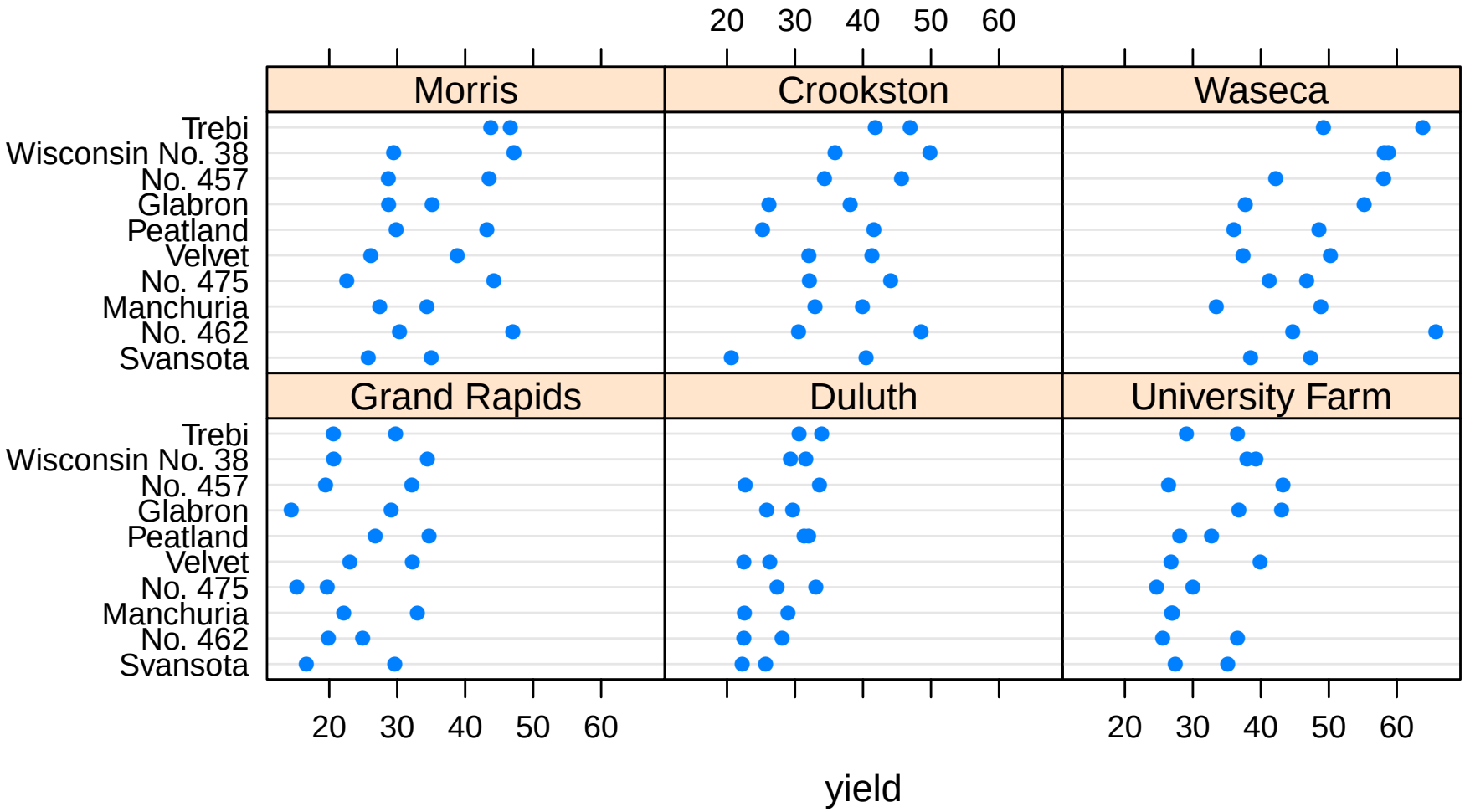
Las gráficas trellis o de retícula, permiten combinar variables continuas condicionadas con respecto a variables categóricas (o discretizadas). Cada tipo de gráfica se puede combinar con una retícula

- Cuando se usa: Cuando se combinan variables continuas con variables categóricas.
- Ejemplo: Los datos corresponden a latitud y longitud de 1,000 terremotos mayores a 4.0.



Gráficas Trellis, otro ejemplo

Los siguientes datos corresponden a rendimientos de 10 variedades de trigo en 6 estaciones de Minnesota.



Gráficas de dispersión I

Consideren la siguiente situación: Un economista trabaja en Coca Cola y su función es buscar la eficiencia.

- Analiza operaciones de entrega y servicio en máquinas tragamonedas de refrescos.
- Tiene una hipótesis: el tiempo $tiempo$, utilizado por un repartidor, en cargar y dar servicio a una máquina está relacionado con la cantidad de cajas de producto entregadas, $cajas$:

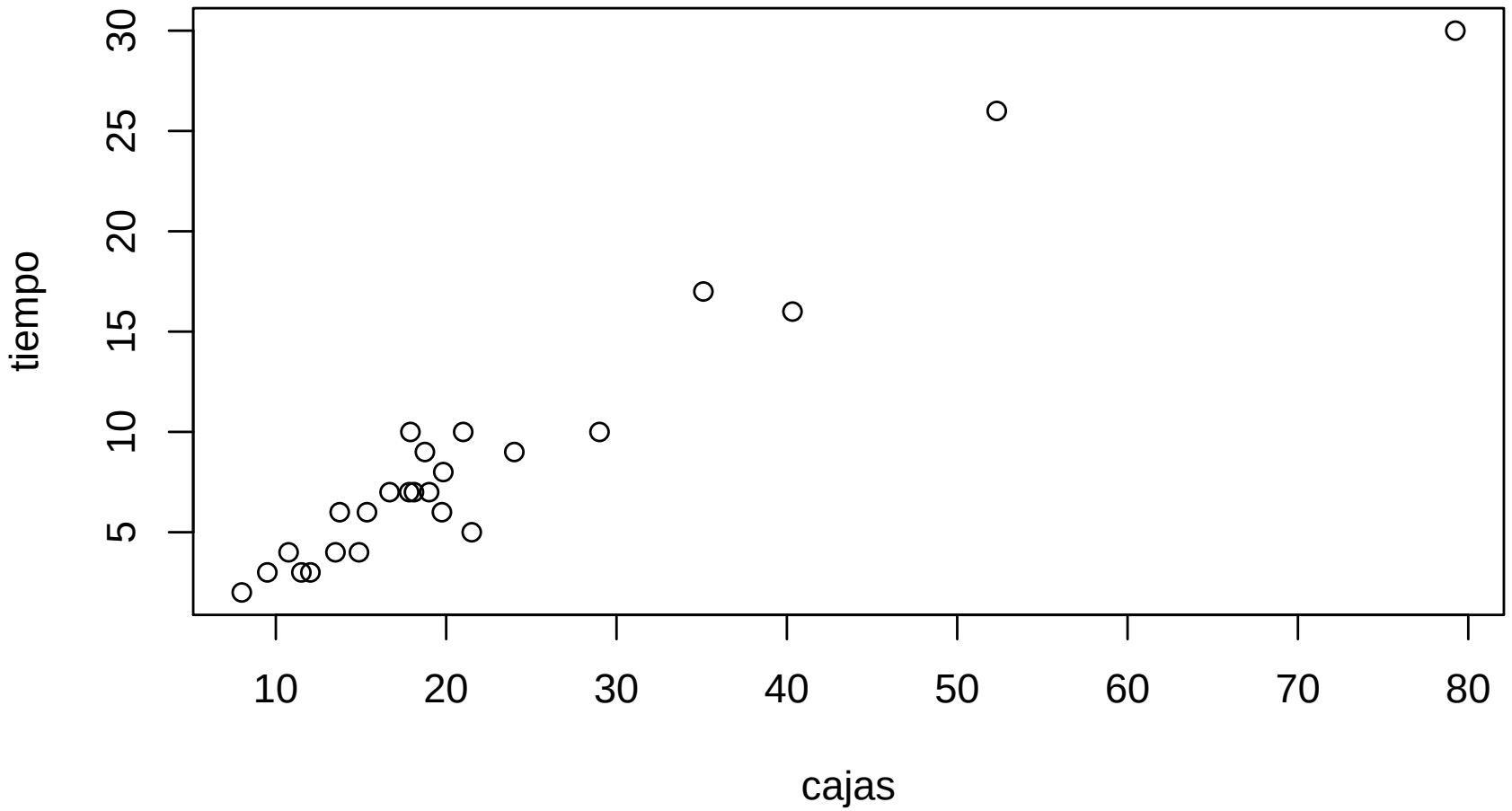
$$tiempo = f(cajas)$$

- No conoce la forma de la función f, así que requiere de un procedimiento que le permita verificar su teoría y poder darse una idea sobre la forma de f.
- Para corroborar su hipótesis, escoge 25 tiendas *al azar* que tengan máquinas tragamonedas y anota los datos en su libreta:

tienda	tiempo (min)	cajas
1	16.68	7
2	11.50	3
3	12.03	3
4	14.88	4
⋮	⋮	⋮
25	10.75	4

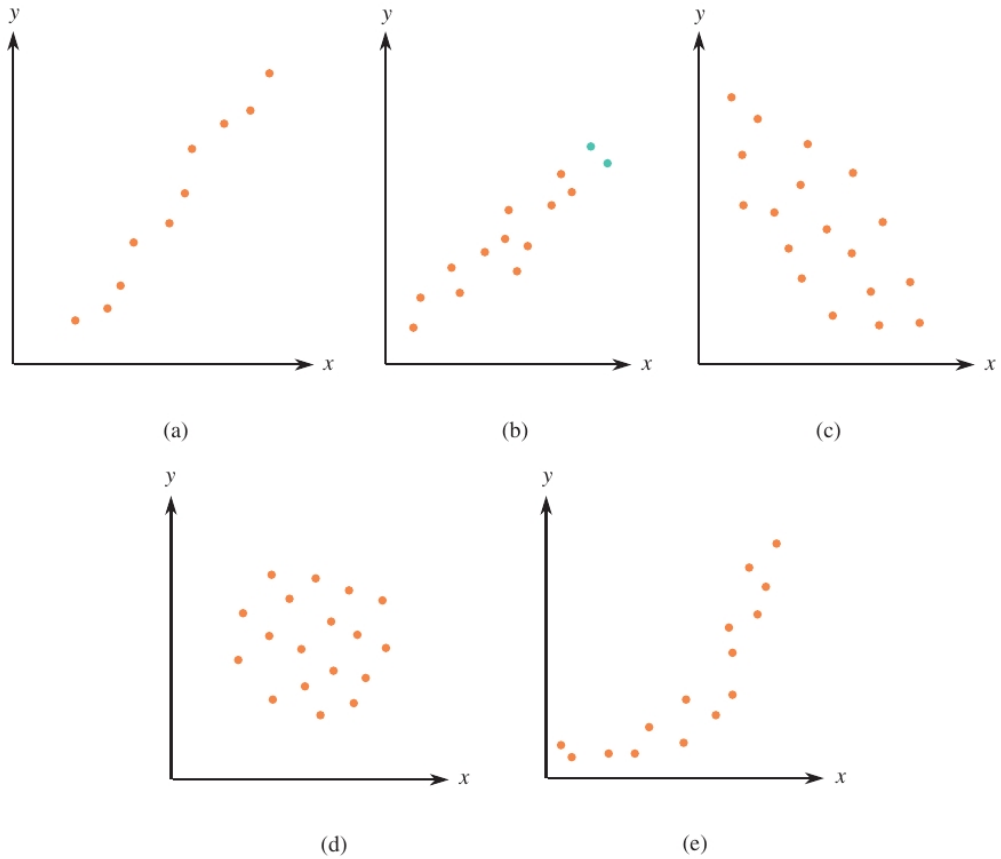
Gráficas de dispersión de puntos

Con los pares de datos elabora una gráfica de *dispersión de puntos* o *scatterplot*. Al parecer, el economista no estaba tan equivocado: mientras más cajas se tienen que entregar, más es el tiempo de servicio del repartidor. ¿Qué tipo de relación se puede apreciar en la gráfica?



Correlación I

- Usualmente los economistas buscan relaciones entre diferentes variables para identificar qué tipo de relación tienen: directa, inversa, lineal, no lineal, etc.
- Una gráfica de dispersión de puntos nos da una impresión visual de cómo y qué tan fuertemente pueden estar relacionadas dos variables.



Correlación II

- Una medida numérica de la fuerza de relación entre variables x y y es el **coeficiente de correlación**

Coeficiente de correlación

El coeficiente de correlación (de Pearson) para un conjunto de puntos $(x_i, y_i), \quad i = 1, 2, \ldots, n$ se define como:

$$r = \frac{\sum_{i=1}^{n} z_x z_y}{n-1}$$

donde $z_x = \frac{x_i - \bar{x}}{s_x}$ y $z_y = \frac{y_i - \bar{y}}{s_y}$. Este coeficiente tiene la propiedad de que $-1 \leq r \leq 1$ y mide el *grado de relación lineal* que hay entre dos variables X y Y de los que se tiene una muestra.

Algunas de las propiedades más importantes del coeficiente de correlación son las siguientes:

1. El valor de r no depende de la unidad de medida de las variables que lo conforman.

Correlación III

2. El valor de r no depende de cuál sea la variable considerada como dependiente o independiente
3. El valor de r está entre -1 y 1. Un valor cercano a 1 indica una relación fuerte positiva mientras que un valor cercano a -1 es una fuerte relación negativa. Se puede usar la siguiente tabla como un indicador de la fortaleza de la relación lineal entre variables:

$\lvert r\rvert \in [0.8, 1]$	Fuerte
$\lvert r\rvert \in [0.5, 0.8)$	Moderada
$\lvert r\rvert \in [0, 0.5)$	Débil

4. Una correlación de $r = 1$ sólo ocurre cuando todos los puntos en la gráfica caen exactamente sobre una linea recta con pendiente positiva. Similarmente, $r = -1$ ocurre cuando todos los puntos caen en una línea recta con pendiente negativa.
5. Correlación NO implica causalidad!

Ejemplo

Los datos correspondientes al repartidor de refresco son los siguientes:

```
[1] "cajas"
 [1] 16.68 11.50 12.03 14.88 13.75 18.11  8.00 17.83 79.24 21.50 40.33
[12] 21.00 13.50 19.75 24.00 29.00 15.35 19.00  9.50 35.10 17.90 52.32
[23] 18.75 19.83 10.75
[1] "tiempo"
 [1]  7  3  3  4  6  7  2  7 30  5 16 10  4  6  9 10  6  7  3 17 10 26  9
[24]  8  4
```

¿Cuál es la correlación?

```
zx <- (cajas-mean(cajas))/sd(cajas)
zy <- (tiempo-mean(tiempo))/sd(tiempo)
r <- sum(zx*zy)/24
r

[1] 0.9646146

cor(cajas, tiempo)

[1] 0.9646146
```

Definición del modelo lineal I

- En el ejemplo del repartidor de refrescos, los puntos caen, aproximadamente, sobre una linea recta.
- El economista postula, como modelo para explicar el tiempo de servicio, que la función f es de la forma

$$tiempo = f(cajas) = \alpha_0 + \alpha_1 cajas$$

- Sin embargo, la relación no es exacta, como se ve en el scatterplot, así que aún falta especificar algo más en el modelo para que sea realista.
- El elemento que falta es la diferencia entre el valor del tiempo observado y el valor del tiempo que está sobre la linea recta. Esta diferencia es interpretada como un *error estadístico.*
- Este error no es un error de medición, sino un error que representa incertidumbre. El error es en realidad una *variable aleatoria*.
- Finalmente, el modelo que se postula se puede representar de la siguiente forma:

$$tiempo = f(cajas) = \alpha_0 + \alpha_1 cajas + \epsilon$$

En la ecuación anterior, los elementos tienen nombres:

Definición del modelo lineal II

- $tiempo$ es llamada la variable de respuesta (o dependiente)
- $cajas$ es llamada la variable predictiva (o independiente)
- α_0 y α_1 son parámetros, para la ordenada al origen y para la pendiente, respectivamente
- ϵ es un error estocástico, y por lo tanto, debe tener una distribución de probabilidad.

Preguntas del modelo lineal

- El modelo anterior, aunque simple, comienza a generar preguntas interesantes:
  - ¿Cuál es la distribución más apropiada para los errores estadísticos?
  - ¿Cómo se estiman los parámetros del modelo?, es decir, de todos los posibles valores que pueden tomar α_0 y α_1, ¿cuáles deben elegirse o en base a qué criterio?
  - Una vez que se obtiene la ecuación de la "mejor" linea recta, ¿Para qué sirve el modelo? ¿Cómo me puede ayudar a tomar decisiones?

Distribución de los errores I

¿Cuál es la distribución más apropiada para los errores estadísticos?

- En la práctica, se espera que los errores estadísticos con respecto a un modelo sean tanto negativos como positivos. En promedio, se espera que los errores se anulen unos con otros, por lo que es razonable suponer que la media o esperanza de la distribución de ϵ sea 0. Esto se denota con el símbolo $E(\epsilon|cajas) = 0$.
- También es razonable, aunque no necesario, suponer que los errores se *dispersan* del mismo modo en cualquier parte de la recta. Esto implica que la varianza de los errores es *constante*. Esto se representa como $\mathrm{Var}(\epsilon|cajas) = \sigma^2$ donde σ^2 es una constante numérica.
- Por último, La *forma* de la distribución de los errores se puede modelar adecuadamente como si fuera una campana. Esto puede expresarse diciendo que los errores tienen distribución normal con media 0 y varianza σ^2:

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Distribución de los errores II

- Resumiendo, el modelo de regresión lineal es:

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 x_i + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}\left(0, \sigma^2\right) \\ \epsilon_1, \epsilon_2, \ldots, \epsilon_n & \quad \text{son independientes} \end{aligned}$$

- La siguiente pregunta es: ¿Cómo elegir la "mejor" línea recta? ¿En base a qué criterio?

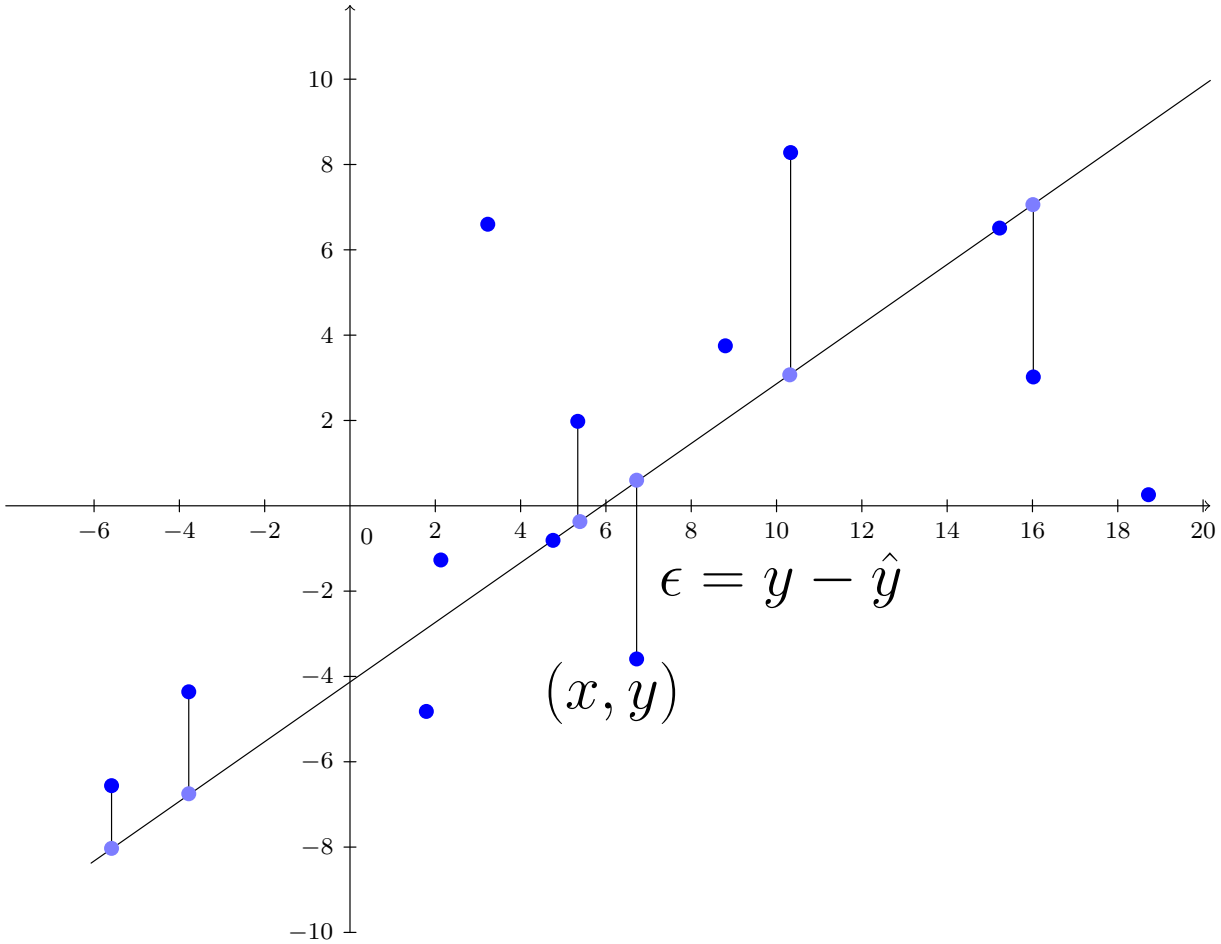
Estimación de los parámetros: mínimos cuadrados I

Mínimos cuadrados busca encontrar la recta que minimiza la suma de los cuadrados de los errores: $\sum_{i=1}^{2} \epsilon_i^2 = \sum_{i=1}^{2} (y_i - \hat{y}_i)^2$. En términos de los parámetros se busca minimizar la siguiente función de los parámetros α y β:

$$f(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Encontrar los valores de α y β que minimizan esta función requieren un poco de cálculo, por lo que consideraremos las ecuaciones dadas a continuación.

Estimación de los parámetros: mínimos cuadrados II



Estimación de los parámetros: mínimos cuadrados III

- Los valores de los parámetros que encontramos se pueden encontrar con las siguientes fórmulas:

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= \frac{SXY}{SXX} \end{aligned}$$

  donde $SXY = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$ y $SXX = \sum_{i=1}^{n}(x_i - \bar{x})^2$

- Para calcular $\hat{\alpha}$ y $\hat{\beta}$ se requiere conocer sólo 5 números: $n, \bar{x}, \bar{y}, SXX, SXY$.
- Una vez que se calculan los valores de los parámetros, hay dos valores para la respuesta: el **observado** y_i y el **estimado**, $\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$
- A la diferencia entre los valores observados y los estimados, que corresponden a valores numéricos de los errores ϵ_i se les llama residuales y se denotan como $e_i = y_i - \hat{y}_i$.
- Como los residuales estiman a los errores, la varianza de éstos se puede estimar con la varianza muestral de los residuales, por lo tanto:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Inferencia en regresión lineal simple

En nuestro ejemplo:

```
Call:
lm(formula = tiempo ~ cajas)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5811 -1.8739 -0.3493  2.1807 10.6342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.321      1.371   2.422   0.0237 *
cajas          2.176      0.124  17.546 8.22e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.181 on 23 degrees of freedom
Multiple R-squared:  0.9305,Adjusted R-squared:  0.9275
F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15
```

Intervalos de confianza para los coeficientes de un modelo de regresión lineal

Intervalo de confianza para la pendiente

Un intervalo de confianza del 95 % para la pendiente β es de la forma:

$$\hat{\beta} \pm (\text{valor críritico } t) s_{\hat{\beta}}$$

donde s_β es el error estándar del estimador y t es un valor crítico que se obtiene de las tablas t con $n-2$ grados de libertad. Por ejemplo, para distintos valores de n el valor t se muestra en la siguiente tabla:

```
      5       6       7       8       9      10      11      12
2.015048 1.943180 1.894579 1.859548 1.833113 1.812461 1.795885 1.782288
     13      14      15      16      17      18      19      20
1.770933 1.761310 1.753050 1.745884 1.739607 1.734064 1.729133 1.724718
     21      22      23      24      25      26      27      28
1.720743 1.717144 1.713872 1.710882 1.708141 1.705618 1.703288 1.701131
     29      30     >30
1.699127 1.697261 1.690000
```

En nuestro ejemplo, $n-2 = 23$, así que el intervalo es
$(2.176 - 1.713872(0.124), 2.176 + 1.713872(0.124) = (1.96, 2.39)$

Intervalos de confianza para los coeficientes de un modelo de regresión lineal

Intervalo de confianza para la ordenada al origen

Un intervalo de confianza del 95 % para la ordenada α es de la forma:

$$\hat{\alpha} \pm (\text{valor ciritico } t) s_{\hat{\alpha}}$$

donde s_α es el error estándar del estimador y t es un valor crítico.

En el ejemplo, el intervalo es
$(3.321 - 1.713872(1.371), 3.321 + 1.713872(1.371) = (0.97, 5.67)$

Ejemplo Regresión I

Gasto en investigación y desarrollo y tasa de crecimiento

Los siguientes datos corresponden a la inversión en investigación y desarrollo y su relación con el crecimiento de 8 industrias. Las variables son $x =$ gasto en investigación y desarrollo (en miles de dólares) y $y =$ tasa de crecimiento (en % anual)

```
[1] "x"
[1] 2024 5038  905 3572 1157  327  378  191
[1] "y"
[1]  1.90  3.96  2.44  0.88  0.37 -0.90  0.49  1.01
```

- ¿Un modelo de regresión lineal simple proveerá información útil para predecir la tasa de crecimiento a partir de el gasto en investigación y desarrollo?
- ¿Cuánto aumenta el crecimiento por cada mil dólares que se invirtien en investigación y desarrollo?

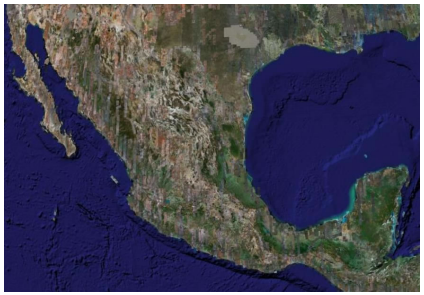
Módulo 12. Herramientas Estadísticas III: Muestreo y diseño de encuestas

Muestreo

- Una muestra es un subconjunto de una población, usada para determinar verdades a través de hipótesis sobre características de una población.
- ¿Porqué muestrear?
  - Recursos (tiempo, dinero) y carga de trabajo
  - Nos da resultados con precisión conocida que puede ser calculada matemáticamente
- El marco muestral es la lista de que se extrae la muestra:
  - Directorios
  - Listas de personas/objetos
  - Inventarios
  - Padrón electoral

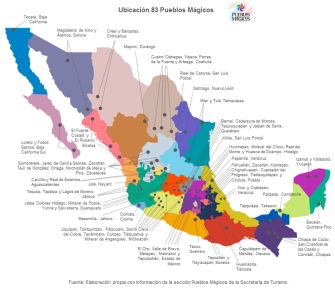
Muestreo

¿A qué quieres generalizar?



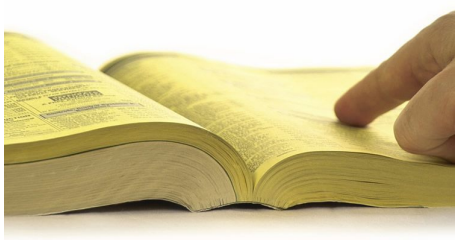
La población teórica

¿A qué población tienes acceso?



La población de estudio

¿Cómo se puede accesar a la muestra?



El marco muestral

¿Quién está en el estudio?

La muestra

Representatividad I

- Una de las características más deseadas de una muestra es que sea **representativa**: *que replique las características de la población a la que pertenece*.
- 3 factores que afectan la representatividad:
  - El procedimiento de muestreo
  - El tamaño de muestra
  - La participación (no respuesta)

Algunas fuentes de información relevantes sobre muestreo:

- Magic Town (1947), James Stewart: El investigador de opinión pública descubre un pueblo que tiene las mismas características que todo Estados Unidos: Grandview.

Esquema de muestreo

- Muestreo probabilista: cada elemento tiene una probabilidad de selección.
  - Muestreo aleatorio simple
  - Muestreo aleatorio estratificado
  - Muestreo de conglomerados (cluster sampling)
  - Muestreo multiestados
  - Muestreo multifases
  - Muestras complejas
- Muestreo no probabilista
  - Muestreo de conveniencia entrevistar a quien abra la puerta
  - Quota necesitamos que haya por lo menos 100 mujeres en la muestra

Muestreo aleatorio simple I

- Es la forma más simple de una muestra probabilista.
- Cada muestra de tamaño n tiene la misma probabilidad de ser seleccionada.
- El muestreo puede ser con reemplazo (la misma unidad se puede incluir más de una vez) o sin reemplazo (cada unidad en la muestra es diferente)

Ejemplo

Si la población es el conjunto $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$, se pueden formar:

- 8 diferentes muestras de tamaño 1 o de tamaño 7
- 28 muestras de tamaño 2 o 6.
- 56 muestras de tamaño 3 o 5.
- 70 muestras de tamaño 4.

Desventajas:

- Si el marco muestral es grande, el método se vuelve impráctico.
- Los subgrupos de minorías de interés en la población pueden no estar presentes en la muestra en números suficientes para su estudio.

Muestreo sistémico I

- Se usa como un proxy al muestreo aleatorio simple
- Se ordena la muestra con algún esquema de orden
- Se elige una muestra de tamaño n y se toma k el entero más cercano a N/n
- Se elige un número aleatorio entre 1 y k, y la muestra se compone de las unidades $R, R+k, R+2k, \ldots, R+(n-1)k$.
- Siempre el primer elemento se escoge al azar.

Ventajas

- La muestra es fácil de seleccionar
- Un marco muestral adecuado se puede especificar fácilmente
- La muestra se esparse adecuadamente sobre la problacióno de referencia

Desventajas:

- La muestra puede ser sesgada si hay periodicidades ocultas en los datos
- Es difícil establecer la precisión del estimador de los resultados de una encuesta

Muestreo sistémico II

Ejemplo

Para seleccionar una muestra de 45 productos y tomar sus precios, de una lista de 45,000 productos, entonces $k = 1000$. Si se elige como número aleatorio inicial $R = 597$, los productos con los números

$$597, 1597, 2597, \ldots 44597$$

son los que estarán en la muestra

Muestreo estratificado I

- Si la variable de interés a medir toma diferentes valores en diferentes subpoblaciones, se obtienen estimadores más precisos de la población a través de una muestra estratificada
- Los estratos son categorías disjuntas que agrupan a los elementos de una población y se pueden considerar como esas subpoblaciones.
- Cada elemento en un estrato tiene la misma probabilidad de ser seleccionado.

Las razones para usar un muestreo estratificado incluyen:

- Protegerse contra la posibilidad de tener una mala muestra. Utilizando la misma fracción de muestreo en cada estrato asegura una representación proporcionada de la muestra.
- Se pueden apliicar diferentes métodos de muestreo en cada estrato,
- Se quiere obtener datos de precisión conocida para cada subgrupo.
- pueden reducir los costos de la encuesta.
- Bien realizada, pude dar estimaciones más precisas para toda la población.

Desventajas

- En algunos casos (cuando se tienen muchos estratos o un tamaño de muestra mínimo por grupo) el muestreo estratificado puede requerir una muestra mucho más grande que otros métodos.

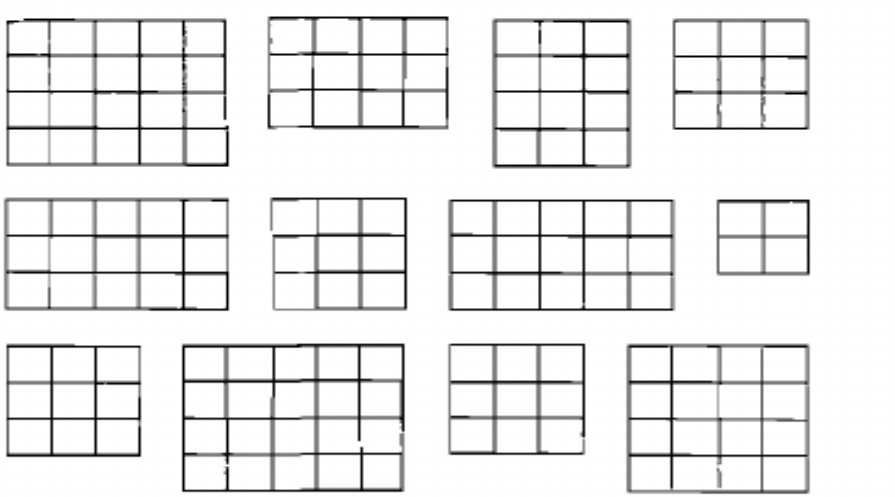
Ejemplo de muestreo estratificado I

Obtener una muestra de estudiantes de la facultad de Ingeniería de la UNAM para comparar las experiencias educacionales y de trabajo de hombres y mujeres. Como hay muchos más hombres que mujeres, se puede estratificar por género y obtener una muestra mayor de mujeres para lograr la misma precisión en los resultados que los que se obtienen para hombres.

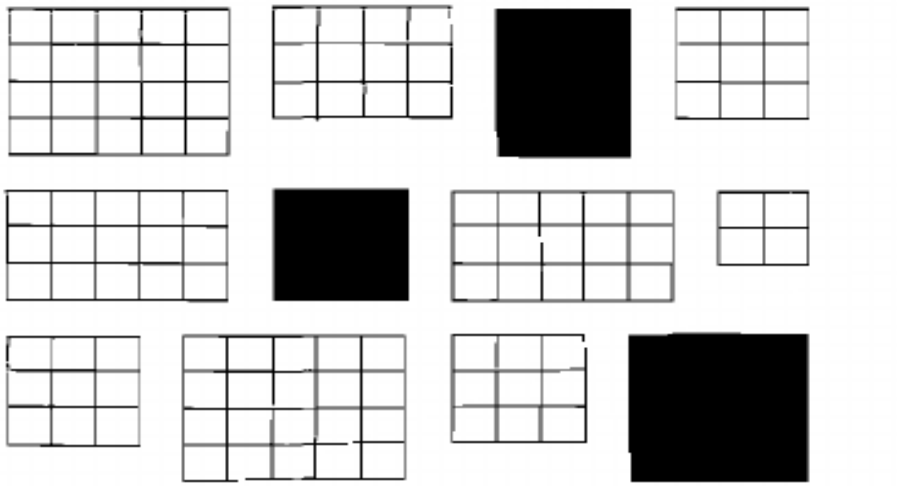
Muestreo por conglomerados I

- La población se divide en subgrupos (conglomerados) de unidades homogéneas.
- Se seleccionan una muestra aleatoria de los conglomerados.
- Se estudian o miden *todas* las unidades de los conglomerados seleccionados.

Muestreo por conglomerados II



Take an SRS of clusters; observe all elements within the clusters in the sample:



Muestreo por conglomerados III

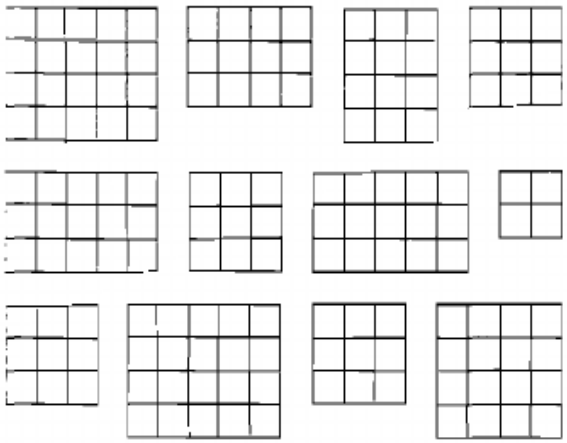
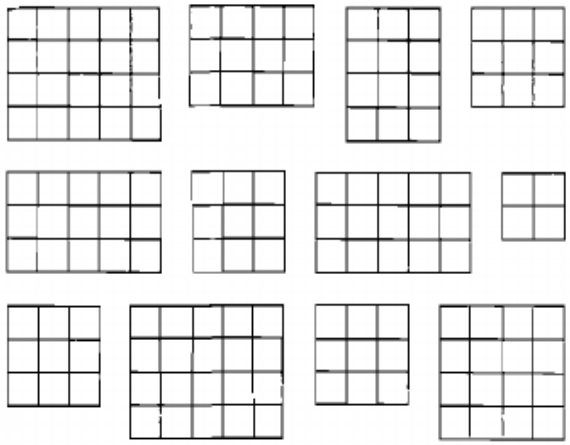
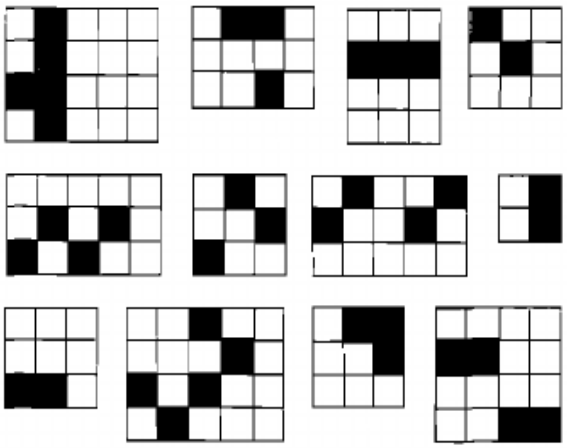
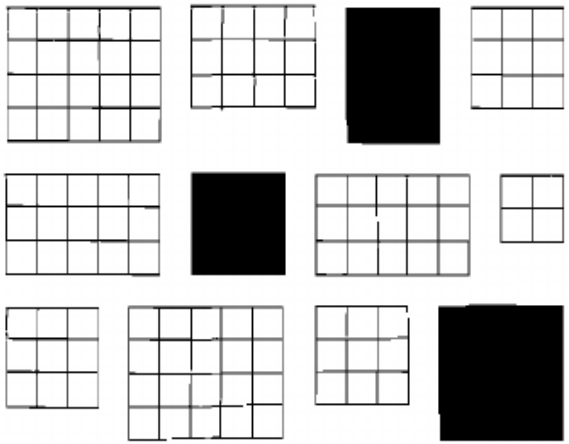
Ventajas:

- Reduce el costo de preparar el marco muestral
- Reduce costos administrativos (viajes, encuestadores, etc)

Desventajas

- El error muestral es mayor que el correspondiente a una muestra aleatoria simple del mismo tamaño.

Diferencias muestreo estratificado y por conglomerados

Stratified Sampling	Cluster Sampling
Each element of the population is in exactly one stratum.	Each element of the population is in exactly one cluster
Population of H strata; stratum h has N_h elements:	One-stage cluster sampling; population of N clusters:
	
Take an SRS from *every* stratum:	Take an SRS of clusters; observe all elements within the clusters in the sample:
	

- En la muestra, todos los estratos están representados, no así los conglomerados.
- Con muestreo estratificado, los mejores resultados se obtienen cuando los elementos en el estrato son *homogéneos*. En los conglomerados. los mejores resultados ocurren cuando los elementos en ellos son internamente *heterogéneos*.

Proceso de muestreo

El proceso de muestreo tiene varias etapas:

1. Definir la población de interés
2. Especificar el *marco muestral*, un conjunto de elementos, o los posibles eventos a medir.
3. Especificar el *método de muestreo* para seleccionar a los elementos o eventos del marco muestral
4. Determinar el tamaño de muestra
5. Implementar el plan de muestreo
6. Obtener los datos
7. Revisar el proceso de muestreo

Elementos a considerar en el diseño de un cuestionario I

- Identificar qué es lo que se quiere saber:
  - Definir los objetivos
  - Definir el alcance
- Siempre probar las preguntas antes de llevar a cabo la encuesta
- Mantener las preguntas simples y claras
- Usar preguntas específicas en lugar de generales, siempre que sea posible
- Relacionar las preguntas al concepto de interés
- Decidir si usar preguntas cerradas o abiertas
  - Cerradas: se eligen sólo algunas de un conjunto de opciones
  - Abiertas: El respondente no está limitado a una respuesta de un conjunto específico
- Reportar la pregunta hecha, no interpretaciones
- Evitar preguntas sesgadas que motiven al respondente a decir lo que se quiere oir
- Usar elección forzada en lugar de preguntas acuerdo/desacuerdo
- Preguntar sólo un concepto en cada pregunta
- Poner atención al efecto en el orden de las preguntas.

Algunas referencias complementarias

- Cole Nussbaumer Knaflic (2015) *Storytelling with Data: A data visualization guide for business professionals*, Wiley.
- Naomi B. Robbins (2013) *Creating more effective Graphs*, Chart House.
- Joel Best (2009) *Uso y Abuso de las Estadísticas* Ed. Cuatro Vientos
- Gerardo Herrera Corral (2018) *El azaroso arte del engaño: Historias del mundo de la casualidad y la estadística*, Taurus
- Cursos en Coursera, edX, Udemy, JournalismCourses

¡Muchas gracias!