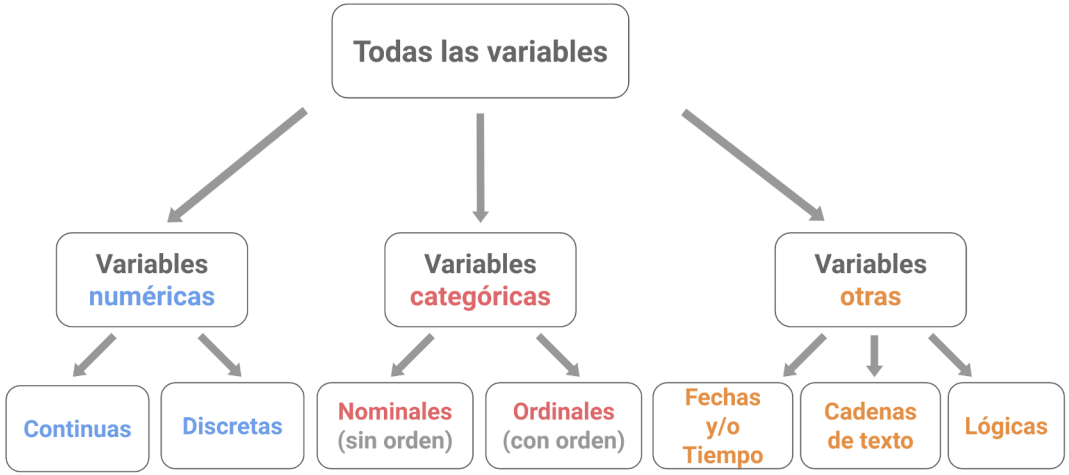
Estadística descriptiva con variables categóricas

1. Qué es una variable categórica y tipos existentes.
2. Estadísticas que se pueden calcular.
3. Concepto de probabilidad.

Variables categóricas

- En estadística, una variable categórica es una variable que puede tomar uno de un número limitado, y por lo general fijo, de posibles valores.
- Las variables categóricas también se denominan variables cualitativas o variables de atributos.
- Los valores de una variable categórica son categorías o grupos mutuamente excluyentes.
- Los datos categóricos pueden tener o no tener un orden lógico.



Fuente: Sebastian Garrido de Sierra

Variables categóricas

Ejemplo:

1. Numerico:
   a. Sexo (1 = Mujer, 2 = Hombre)
   b. Resultados de una encuesta (1 = De acuerdo, 2 = Neutral, 3 = En desacuerdo)
2. Texto:
   a. Formas de pago (Efectivo o Crédito)
   b. Configuraciones de una máquina (Bajo, Medio, Alto)
   c. Tipos de producto (Madera, Plástico, Metal)

Es importante identificar si una variable categórica es ordinal o no. Ya que en muchos casos será necesario transformar esa variable a números y **los números deberán reflejar ese orden.**

¿Cuáles variables del ejemplo son ordinales y cuáles no?

Variables categóricas

Ejemplo:

1. Numerico:
   a. Sexo (1 = Mujer, 2 = Hombre)
   b. Resultados de una encuesta (1 = De acuerdo, 2 = Neutral, 3 = En desacuerdo)
2. Texto:
   a. Formas de pago (Efectivo o Crédito)
   b. Configuraciones de una máquina (Bajo, Medio, Alto)
   c. Tipos de producto (Madera, Plástico, Metal)

Es importante identificar si una variable categórica es ordinal o no. Ya que en muchos casos será necesario transformar esa variable a números y **los números deberán reflejar ese orden.**

¿Cuáles variables del ejemplo son ordinales y cuáles no?

Ordinales: Resultados de una encuesta. Configuraciones de una máquina.

No ordinales: Sexo, formas de pago.

Depende: Tipos de producto. Acá el contexto del problema podría entregar información importante.

Estadísticas que se pueden calcular

Con las **variables categóricas no ordinales** es posible calcular algunas estadísticas como:

- Frecuencias.
- Moda: el valor que más se repite o con la frecuencia mayor.

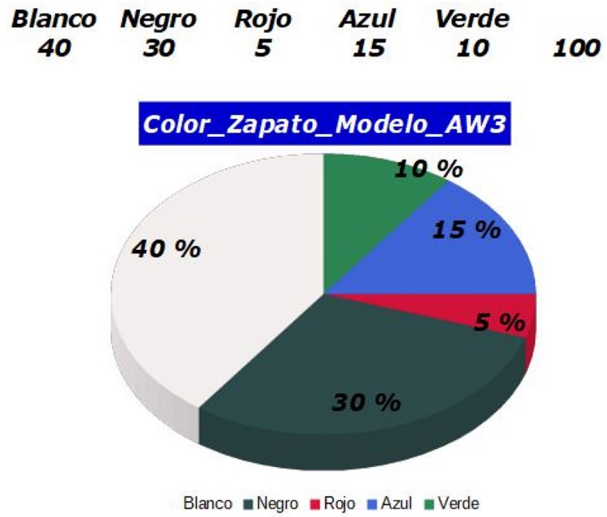
No tiene sentido calcular promedios, desviación estándar, cuartiles, etc. Esto reduce bastante el análisis que podemos hacer con respecto a la variable.

A pesar de esto, es posible hacer algunos gráficos interesantes como: gráfico de torta y gráfico de barras.

Estadísticas que se pueden calcular

El gráfico de torta:

Acá se usó solo una variable.

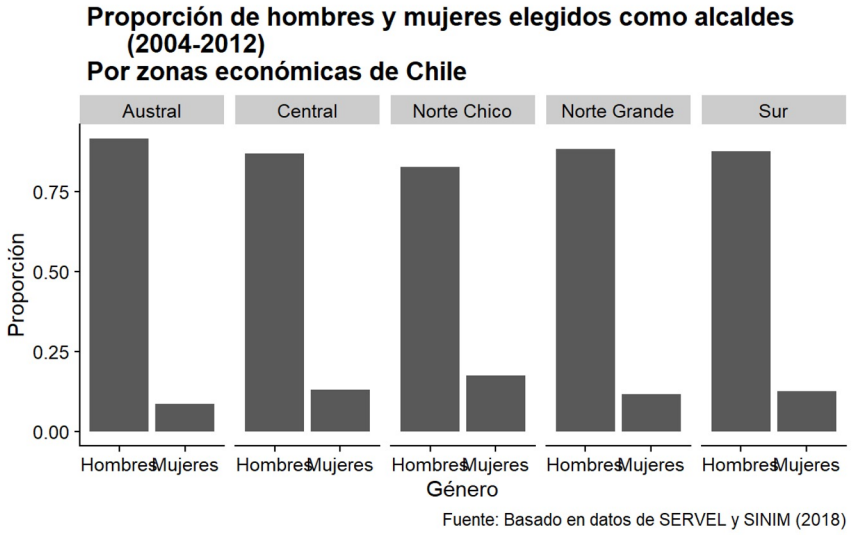


El gráfico de barras:

Aca se usaron 2 variables al mismo tiempos.

Sexo y zona económica.

Notar que el eje Y está en porcentaje, no es obligatorio, pero a veces recomendable si esto permite una mejor visualización.



Estadísticas que se pueden calcular

Con las **variables categóricas ordinales** se puede hacer mucho más. Al transformarlas en número, es posible calcular varias estadísticas típicas y, como son ordinales, si pueden representar comportamientos de manera fidedigna.

Veamos el ejemplo anterior: Configuraciones de una máquina (Bajo, Medio, Alto). Acá, los valores categóricos podrían tomar los siguientes valores:

Bajo = 0 ; Medio = 1 ; Alto = 2

Los valores numéricos reflejan una cercanía que se cumple de igual manera en el caso de los valores categóricos. *Bajo* esta más cerca de *Medio* y más lejos de *Alto*. En el caso numérico, el *0* está más cerca del *1* y más lejos del *2*.

Hay que considerar que algunos cálculos pueden generar números decimales. Estos resultados pueden aproximarse al momento de hacer una interpretación del resultado en la escala original (valores categóricos).

Estadísticas que se pueden calcular

Finalmente, no solamente es posible pasar variables categóricas a numéricas, sino que a veces es recomendable pasar de variables numéricas a categóricas. Esto se llama: **Intervalos de clase.**

Los intervalos de clase son divisiones que se hacen a la variable numérica. En específico al rango numérico que esta tiene. Veamos un ejemplo:

- Como se ve, las clases se transforman en categorías ordinales.
- Ahora, se puede tratar como cualquier variable categórica ordinal.
- Esto es útil para resumir la información. Además, si poseemos información adicional del problema, esta división nos puede ayudar a encontrar patrones interesantes que se nos haría más difícil de encontrar si no hiciéramos esta división.

Intervalo de clase: Para agrupar los datos es necesario definir el limite inferior y superior de la clase. La diferencia entre los límites determina el intervalo.

Clase X (Estatura)	Frecuencia F N° Estudiantes
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8
Total	100

Número de clases: 5

En la clase 60-62 el límite inferior es: 60 y el límite superior es 62

Intervalo de clase: 62 – 60 = 2

Fuente: http://www.managementempresarial.com/diccionario/intervalo-de-clase.html

Concepto de probabilidad

Muchas veces podemos interpretar nuestros datos como el resultado de una secuencia de eventos al azar, por lo tanto es bueno para el análisis pensar en el proceso que los generó.

La probabilidad es una medida de la incerteza que tenemos del resultado de un experimento.

¿Se acuerdan de la variable aleatoria Uniforme?

El lanzamiento de los dados es un ejemplo:

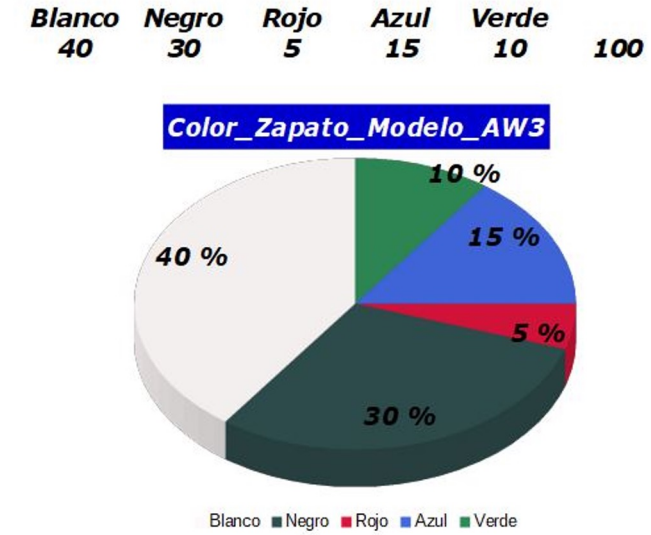
La probabilidad de que en un dado salga un 2 es ⅙ .
La probabilidad de que en un dado salga un 6 es ⅙ .

Si lanzo 1 dado y quiero saber la probabilidad de que salga un 2 o un 6, entonces la probabilidad es: ⅙ + ⅙ = ⅓

Concepto de probabilidad

Entonces, tomando esta idea podemos volver al gráfico de torta anterior:

Si son 100 zapatos y 40 son blancos:



¿Cuál es la probabilidad de que tomemos uno al azar y sea blanco?

Respuesta: 40/100 = 0.4

La probabilidad va entre 0 y 1 y el porcentaje va entre 0 y 100.

Esto se hace comúnmente con variables categóricas, ordinales y no ordinales. Está muy relacionado con lo visto anteriormente sobre frecuencias (relativas y absolutas).

Concepto de probabilidad

Es importante saber un poco de probabilidades, ya que en Machine Learning hay muchos algoritmos que trabajan con esta idea.

Además, como vimos, sirve para interpretar resultados de gráficos o tablas de datos.

Revisar:

1. **Probabilidad condicional**

2. **Teorema de Bayes**