Introducción al análisis de variables categóricas

Distribución de Frecuencias y Tablas de Contingencia

Sebastián Muñoz-Tapia

Antropología UAH

2023-02-24

Contenido

Tipos de análisis

- Pueden distinguirse al considerar:

Cantidad de variables:

- univariado, bivariados, multivariados

Si es:

- un análisis descriptivo
- un análisis inferencial

Tipo de variables:

- si utilizan variables categóricas, cuantitativas o ambas.

Si es :

- inferencial o utiliza machine learning/ inteligencia artificial

Según cantidad de variables

- *Análisis univariado*:
  - Se enfoca en examinar una sola variable.
  - Por ejemplo, si queremos analizar la cantidad de personas por género en una población, estaríamos realizando un análisis univariado.
- *Análisis bivariado*:
  - Examina la relación entre dos variables.
  - Por ejemplo, si queremos examinar cómo la edad de una persona afecta su nivel de educación, estaríamos realizando un análisis bivariado.
- *Análisis multivariado*:
  - Examina de la relación entre tres o más variables.
  - Por ejemplo, si queremos examinar cómo la edad, el género y la educación afectan la probabilidad de que una persona vote, estaríamos realizando un análisis multivariado.

Según tipos de variables

- *Análisis de variables cuantitativas*:
  - Examina variables numéricas, como la edad, el ingreso o la estatura.
- *Análisis de variables cualitativas*:
  - Examina variables categóricas, como el género, la etnia o la religión.
- *Análisis de variables cuantitativas y cualitativas*:
  - Examina la relación entre variables cuantitativas y cualitatitvas.
  - Por ejemplo, si existe una diferencia de salarios entre hombres y mujeres.

Según si es descriptivo o inferencial

- *Estadística descriptiva*:
  - Describe, resume variables que refieren directamente al universo.
  - Por ejemplo, el CENSO.
- *Estadística inferencial*:
  - Examina variables que refieren a una muestra del universo.
  - Realiza **inferencias** a través de pruebas de hipótesis de significación estadística.
  - Por ejemplo, una encuesta.

De lo inferencial al machine learning/ inteligencia artificial

- *Análisis inferencial*:
  - Examina variables que refieren a una muestra del universo.
  - Realiza pruebas de hipótesis de significación estadística.
- *Machine Learning/Inteligencia Artificial*:
  - Examina variables que refieren a una muestra o del universo.
  - Realiza procesos de aprendizaje de los algoritmos.
  - Por ejemplo, Algoritmos de recomendación de consumos culturales (Spotify, Netflix, etcétera).

Ejemplos

Análisis univariado:

- *Distribución de frecuencias*:
  - Se utiliza para examinar la frecuencia con la que ocurre cada valor en una **variable cualitativa** o una **cuantitativa en rangos**. Por ejemplo, si queremos saber cuántas personas de una muestra tienen cierta religión, podemos hacer una distribución de frecuencias de la variable "religión" para identificar cuántas personas están en cada rango de edad.
- *Promedios, mediana y moda*:
  - Se utilizan para resumir la información de una variable **cuantitativa**. El promedio indica el valor promedio de una variable, la mediana indica el valor central de una variable y la moda indica el valor más común de una variable.

Ejemplos, continuación...

Análisis bivariado:

- *Tablas de contingencia*:
  - Examina la relación entre dos **variables categóricas**. Por ejemplo, para entender cómo el género y la preferencia política están relacionados, podemos hacer una tabla de contingencia que muestre cuántas personas de cada género prefieren cada partido político.
- *Correlaciones*:
  - Examina la relación entre dos **variables cuantitativas**. Por ejemplo, para saber si hay una relación entre la altura y el peso, podemos calcular la correlación entre estas dos variables y ver si hay una relación positiva, negativa o no hay relación entre ellas.

Ejemplos, continuación...

Análisis multivariado:

- *Regresiones multiples*:
  - Examina la relación entre una *variable dependiente* y *varias variables independientes cuantitativas*. Por ejemplo, si queremos saber cómo la edad, el género y el nivel educativo afectan el salario, podemos hacer una regresión que nos permita analizar cómo estas tres variables están relacionadas con el salario.

- *Análisis factorial*:
  - Examina la *relación entre varias variables cuantitativas*. Por ejemplo, si queremos saber cómo se relacionan los diferentes tipos de actividades culturales que las personas realizan, podemos hacer un análisis factorial que nos permita identificar los factores subyacentes que explican la relación entre estas actividades.

Ejemplos, continuación...

Análisis descriptivo:

- *Total de estudiantes de una universidad*: Examinar el puntaje de todos los estudiantes de antropología de la Universidad Alberto Hurtado.

Análisis inferencial:

- *Muestra de estudiantes de una universidad*: Examinar el puntaje de una muestra de los estudiantes de antropología de la Universidad Alberto Hurtado.

Preguntas?...

- Si tiene que analizar la relación entre nivel socioeconómico y votación en las últimas elecciones...
    - ¿Cuáles serían sus variables?
    - ¿Qué tipos de variables son?
    - ¿Qué tipos de análisis univariados realizaría?
    - ¿Qué tipo de análisis bivariados realizaría?
- Si tiene que analizar la relación entre edad e ingresos salariales...
    - ¿Cuáles serían sus variables?
    - ¿Qué tipos de variables son?
    - ¿Qué tipos de análisis univariados realizaría?
    - ¿Qué tipo de análisis bivariados realizaría?

Distribución de Frecuencias

Introducción

- La distribución de frecuencias es una herramienta estadística que nos permite conocer cómo se distribuyen los datos en una muestra o población.
- En esta presentación, veremos los principales conceptos y técnicas para la construcción e interpretación de una distribución de frecuencias.

¿Qué es la distribución de frecuencias?

- La distribución de frecuencias es una técnica estadística que nos permite conocer cómo se distribuyen los datos en una muestra o población.
- La distribución de frecuencias puede ser construida para variables cuantitativas y cualitativas.
- Para trabajar con cuantitativas se deben recodificar.
- En una distribución de frecuencias, se agrupan los datos en clases o intervalos, y se cuenta la frecuencia de cada clase.

Tabla de contingencia para religión y raza

	Blanca	Negra	Otra	Sum
Protestante	50,13	72,95	19,90	50,71
Católica	24,50	6,65	47,10	23,96
Ninguna	17,24	12,34	16,61	16,47
Cristiana	2,90	4,53	3,80	3,22
Judía	2,27	0,32	0,41	1,81
Budismo	0,40	0,32	3,70	0,69
Inter o no confesional	0,48	0,93	0,10	0,51
Musulmana/Islam	0,17	1,12	2,16	0,49
Cristiana ortodoxa	0,56	0,06	0,05	0,44
Hinduismo	0,05	0,03	3,19	0,33
Otra religión oriental	0,12	0,06	0,51	0,15
Nativa americana	0,04	0,00	0,82	0,11
Otra	1,08	0,58	1,49	1,05
No sabe	0,06	0,10	0,15	0,07

Cálculo de las frecuencias absolutas y relativas

- La frecuencia absoluta de una clase es el número de observaciones que caen en cada clase.
- La frecuencia relativa de una clase es la proporción de observaciones que caen en esa clase respecto al total de observaciones.
- La frecuencia relativa puede ser expresada como un porcentaje.

```
## Installing package into 'C:/Users/sebastian/AppData/Local/R/win-library/4.2
## (as 'lib' is unspecified)

## Warning: unable to access index for repository http://www.stats.ox.ac.uk/pu
##   cannot open URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/4

## package 'DataExplorer' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##     C:\Program Files\KMSpico\temp\RtmpUjGZ8t\downloaded_packages

##
## DataExplorer installed

## Warning in pacman::p_load(tidyverse, openxlsx, readxl, readr, janitor, forc
## DataExplorer
```

Distribución de frecuencias de Relgión

Religion	Frecuencia	%	% Acumulado
Protestante	10846	50.71	50.71
Católica	5124	23.96	74.66
Ninguna	3523	16.47	91.13
Cristiana	689	3.22	94.35
Judía	388	1.81	96.17

Principales elementos de una distribución de frecuencias

- Clases o intervalos: son los rangos de valores en los que se divide el conjunto de datos.
- Frecuencia absoluta: es el número de observaciones en cada clase.
- Frecuencia relativa: es la proporción de observaciones en cada clase respecto al total de observaciones.
- Frecuencia acumulada: es la suma de las frecuencias absolutas hasta una determinada clase.
- Frecuencia relativa acumulada: es la proporción de observaciones acumuladas hasta una determinada clase respecto al total de observaciones.

Ejemplos

Edad de los habitantes de una comunidad indígena

- En este caso, las clases podrían ser los rangos de edad (por ejemplo, 0-10 años, 11-20 años, 21-30 años, etc.), y las frecuencias absolutas permitirían conocer cuántas personas hay en cada rango de edad. Además, se podría calcular la frecuencia relativa para conocer la proporción de personas en cada rango de edad.

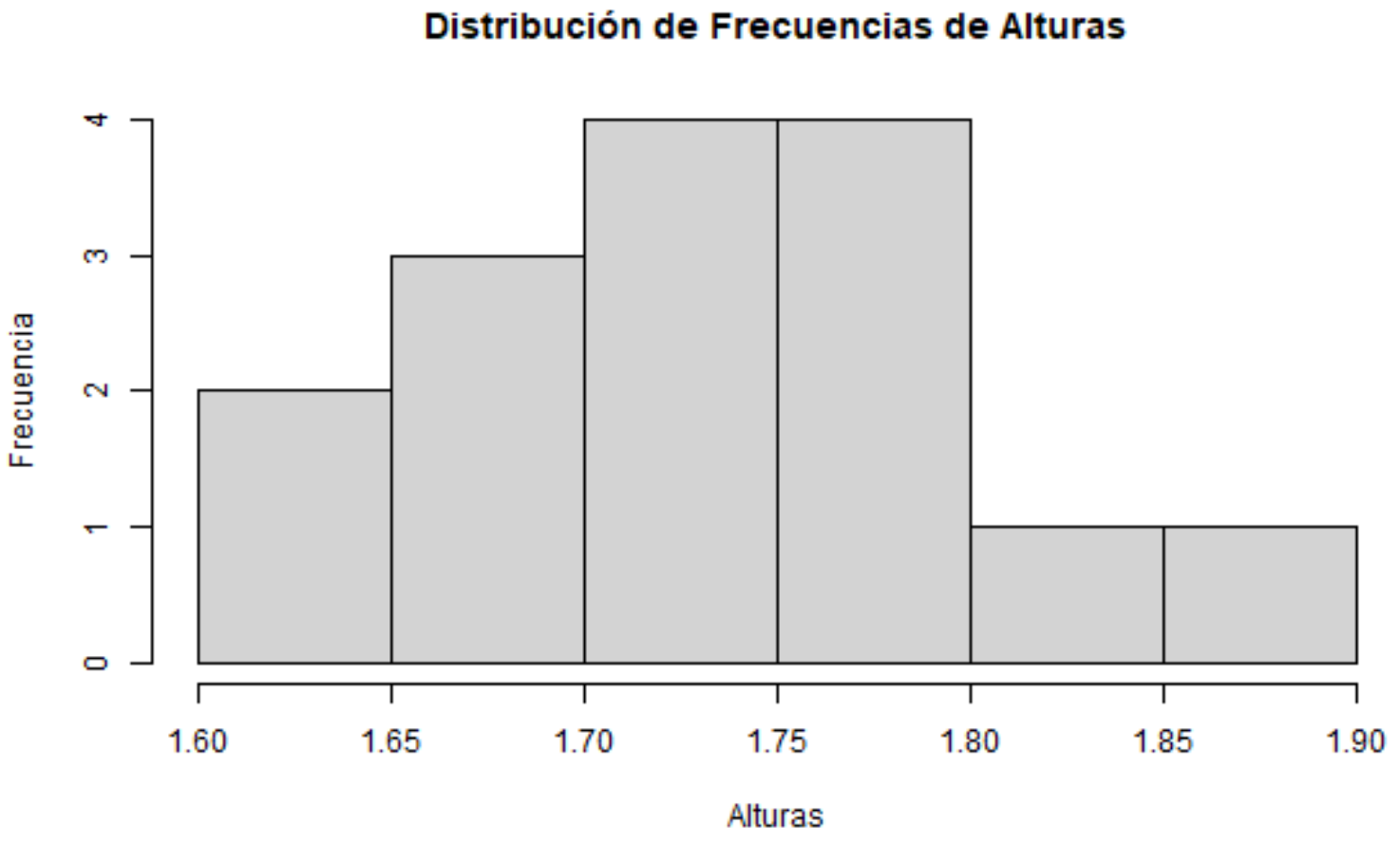
Idioma materno de una población

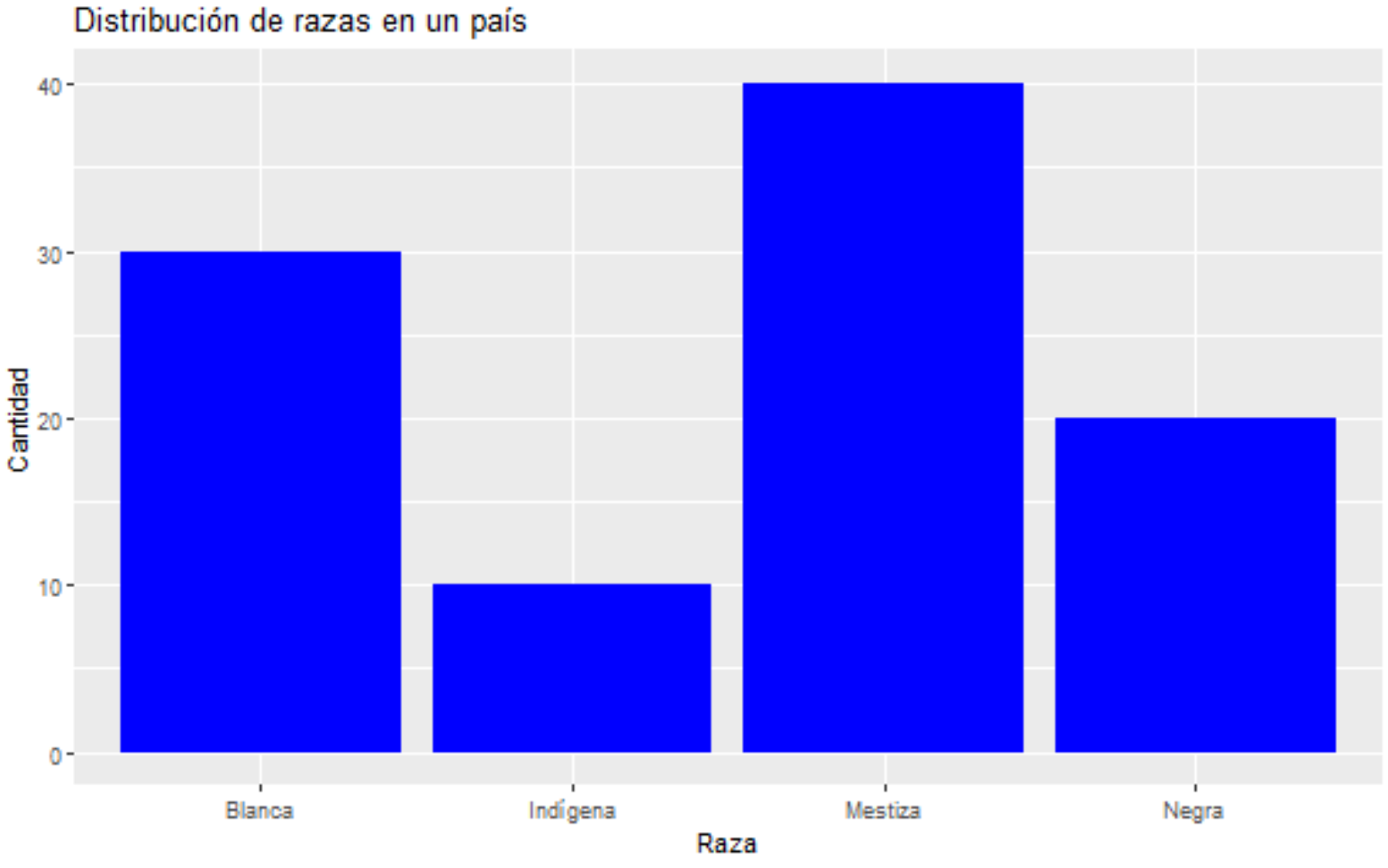
- En este caso, las categorías podrían ser los distintos idiomas hablados por los integrantes de la población (por ejemplo, español, mapudungun, aymara, etc.). Las frecuencias absolutas permitirían conocer cuántas personas tienen cada idioma materno y las frecuencias relativas la proporción de la población que habla cada idioma materno..

Representación gráfica de la distribución de frecuencias

- La distribución de frecuencias puede ser representada gráficamente con un histograma, un polígono de frecuencias o un gráfico de barras.
- Estos gráficos nos permiten visualizar la distribución de los datos y las características más relevantes de la distribución.

```r
alturas <- c(1.70, 1.78, 1.75, 1.65, 1.68, 1.72, 1.85, 1.62, 1.80, 1.76,
hist(alturas, breaks = 5, main = "Distribución de Frecuencias de Alturas
```




Distribución de razas en un país
Cantidad
40
30
20
10
0
Blanca
Indígena
Mestiza
Negra
Raza

¿Qué son las tablas de contingencia?

- En la investigación antropológica, a menudo se utilizan las tablas de contingencia para analizar la relación entre **dos o más variables categóricas**.
- Las tablas de contingencia presentan los datos en una tabla de dos o más dimensiones, con las categorías de una variable en una dimensión y las categorías de la otra variable en la otra dimensión.
- Las tablas de contingencia pueden ser analizadas utilizando técnicas estadísticas como el **chi-cuadrado** y el **test exacto de Fisher** para determinar si la relación entre las variables es significativa o no.
- Son útiles en antropología para analizar las relaciones entre variables categóricas, como la relación entre la etnia y la religión o la relación entre la etnia y la preferencia política.

Organización de variables

- En una tabla de contingencia, la **variable independiente** se coloca en la parte superior y la **variable dependiente** se coloca en el lateral izquierdo.
- Por ejemplo, si se está analizando la relación entre la etnia y la religión, se colocaría la etnia en la parte superior (variable independiente) y la religión (variable dependiente) en el lateral izquierdo.

Organización de las proporciones

- La decisión sobre dónde colocar las proporciones en una tabla de contingencia depende del objetivo de la investigación y de las hipótesis que se quieren probar. Las proporciones se pueden calcular por filas, por columnas o por toda la tabla.

Proporciones por filas

- Permite observar las **proporciones** de *cada categoría de la variable dependiente* dentro de cada *categoría de la variable independiente*.
- Este tipo de análisis es útil para comparar la frecuencia relativa de la variable dependiente para cada categoría de la variable independiente.
- Por ejemplo, si se desea analizar la relación entre la etnia y la religión, se puede colocar la proporción de cada religión dentro de cada etnia

Proporciones por columnas

- Permite observar las **proporciones** de *cada categoría de la variable independiente* dentro de *cada categoría de la variable dependiente*.
- Este tipo de análisis es útil para comparar la frecuencia relativa de la variable independiente para cada categoría de la variable dependiente.
- Por ejemplo, si se desea analizar la relación entre la religión y la etnia, se puede colocar la proporción de cada etnia dentro de cada religión.
- Este tipo de análisis es el más utilizado (y quizás claro) para observar cómo la variable dependiente se relaciona con la variable independiente.
- Si el porcentaje del total en cada categoría es superior al **5%** se suele poner señalar que hay influencia.

Proporciones por totales

- Permite observar las **proporciones** de *cada categoría de la variable dependiente* y de la *variable independiente* en **conjunto**.
- Este tipo de análisis es útil para obtener una **visión general** de la relación entre las dos variables.
- Por ejemplo, si se desea analizar la relación entre la edad y el género, se puede colocar la proporción de cada religión y cada etnia en toda la tabla.

En resumen

- Los posibles tipos de análisis se pueden distinguir al considerar la **cantidad de variables** (univariado, bivariados, multivariados) y si **utilizan variables categóricas, cuantitativas o ambas**.
- La **distribución de frecuencias** permite organizar los datos en clases o intervalos y contar la cantidad de observaciones en cada clase, lo que permite obtener información sobre la frecuencia y la proporción de valores dentro de cada intervalo.
- **Las tablas de contingencia** son una herramienta útil para analizar la relación entre dos variables en la investigación antropológica.
    - En ellas, es importante identificar la *variable independiente* y la *variable dependiente* para poder interpretar correctamente los resultados.
    - El uso de *porcentajes en columnas* permite visualizar cómo la variable dependiente se ve afectada por la variable independiente.