Introduction to Statistics - Young Researchers Fellowship Program

Lecture 2 - More on descriptive statistics & statistical data visualization

Daniel Sánchez Pazmiño

Laboratorio de Investigación para el Desarrollo del Ecuador

September 2024

Recap

- So far, we covered univariate descriptive statistics:
    - Measures of central tendency
    - Measures of dispersion
    - Measures of position
    - Measures of distributional shape
- We must also look at descriptive statistics in other contexts:
    - Categorical data descriptive stats
    - Bivariate descriptive stats (measures of association)
    - Statistical data visualization: boxplots, histograms, scatter plots, etc.

Categorical data descriptive statistics

Describing categorical data

- Our univariate descriptive statistics applied quite well to numerical data.
- However, for categorical data, would we be able to calculate a mean?
  - No, because categories are not numbers.
- There are specific descriptive stats, some of them which mirror numerical data stats, which should be reviewed for categorical data.
  - The frequency of each category
  - Frequency tables
  - Relative frequencies

Frequency of ocurrence

- The frequency of ocurrence of a category is the number of times it appears in the dataset.

$$f = \sum_{i=1}^{n} I(x_i = c)$$

where f is the frequency of category c, n is the number of observations, and I is the indicator function. - $I(x_i = c)$ is 1 if $x_i = c$ and 0 otherwise.

- This can be called the *absolute frequency* of a category.

Frequency of ocurrence

- Notice that if a variable in a dataset is categorical, it may have two or more categories within itself.
  - `sex` may have two categories: `male` and `female`
  - `ethnicity` may have multiple categories: `mestizo`, `afroecuadorian`, `indigenous`, etc.
- Each category of a categorical variable would have its own frequency of ocurrence.

Relative frequency

- The relative frequency of a category is the proportion of times it appears in the dataset.

$$rf = \frac{f}{n}$$

where rf is the relative frequency of category c, f is the frequency of category c, and n is the number of observations.

- This is given to you in *proportion* form.
  - For example, if the relative frequency of `male` is 0.6, then 60% of the dataset falls under the `male` category.
  - Proportions are always between 0 and 1.
  - Find a percentage by multiplying by 100, however, it is recommended to keep it in proportion form for easier calculations.

Frequency tables

- A frequency table is a table that shows the frequency of each category in a categorical variable.
- It is a way to summarize the distribution of a categorical variable.
- For example, consider the SUPERCIAS dataset. We can calculate the frequency of each category in the `region` variable.

Var1	Freq
COSTA	105744
GALÁPAGOS	1340
ORIENTE	7257
SIERRA	95277

Frequency tables

- A frequency table can be presented with both the absolute frequency and the relative frequency.
- The relative frequency is calculated by dividing the absolute frequency by the total number of observations.
- The relative frequency is a proportion, so it is always between 0 and 1.

Frequencies with R

- We can use the `table()` function in R to calculate the frequency of ocurrence of each category in a categorical variable (i.e. a table of frequencies).
  - Works similarly to the numerical data `table()` function.
- Alternatively, use `count` from `dplyr` to calculate the frequency of ocurrence of each category in a categorical variable.
  - This is a shorthand for `group_by()` and `summarize()` for a variable which isn't numerical.
- We may extract a specific category frequency by subsetting the table or using `pull()` from `dplyr`.

Example: SUPERCIAS dataset

- The code for the previous frequency table is as follows:

```
supercias$region %>%
    table()
```

.

```
    COSTA GALÁPAGOS   ORIENTE    SIERRA
   105744      1340      7257     95277
```

Example: SUPERCIAS dataset

- A tidyverse workflow for the frequency table is as follows:

```
## Relative frequencies

supercias %>%
    count(region)
```

```
# A tibble: 4 x 2
  region          n
  <chr>       <int>
1 COSTA      105744
2 GALÁPAGOS    1340
3 ORIENTE      7257
4 SIERRA      95277
```

R implementation for relative frequencies

- For a relative frequency table, we may add an additional column to the frequency table with `mutate()`.
    - This column will be the relative frequency of each category.
- A base R implementation would be passing the `table()` call to `prop.table()`.

Example: SUPERCIAS dataset

- The code for the relative frequency table is as follows:

```
supercias$region %>%
    table() %>%
    prop.table()
```

```
.
      COSTA   GALÁPAGOS     ORIENTE      SIERRA
0.504460495 0.006392581 0.034620119 0.454526806
```

Example: SUPERCIAS dataset

- A tidyverse workflow for the relative frequency table is as follows:

```
supercias %>%
    count(region) %>%
    mutate(relative_frequency = n / sum(n))
```

```
# A tibble: 4 x 3
  region          n relative_frequency
  <chr>       <int>              <dbl>
1 COSTA      105744            0.504
2 GALÁPAGOS    1340            0.00639
3 ORIENTE      7257            0.0346
4 SIERRA      95277            0.455
```

- Note how the denominator, n, is the sum of the frequencies, `sum(n)`.

Dichotomous variables

- A dichotomous variable is a categorical variable with only two categories, which in some cases can be represented as 0 and 1.
    - These are also called binary or dummy variables.
- For example, `sex` can be represented as `male` and `female`, which can be coded as 0 and 1, respectively.
    - It's important to read the variables dictionary in a dataset to understand the coding of dichotomous variables.

Dichotomous variables

- The reason why dichotomous variables are important is that they can be used in statistical models.
    - It is beneficial to understand the category of interest as a 1 and the other category as a 0.
    - We will talk more about these in other lectures and the Econometrics module.
- For now, know that if you take the mean of a dichotomous variable, you are calculating the proportion of the category of interest in the dataset.

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

where $\bar{x}$ is the mean of the dichotomous variable, n is the number of observations, and x_i is the value of the dichotomous variable for observation i.

Dealing with dichotomous variables in R

- If a variable is dichotomous, we may want to recode it to its original values for better interpretation.
  - For example, `1` and `0` can be recoded to `male` and `female`, respectively.
- This can be done using `case_when()` from `dplyr` in a `mutate()` call.
- This would also allow you to do the reverse, recoding a categorical variable to a dichotomous variable.

Example: Dichotomous variable in SUPERCIAS dataset

- Other solutions exist for recoding dichotomous variables, such as `recode()` from `dplyr` or `if_else()` from `dplyr`.
- However, R allows for the use of factors, which are a much more effective way to deal with categorical variables for statistical models.
  - These mantain the categories and their levels (order, if applicable or a numerical value) at the same time.
- We can convert a dichotomous variable to a factor using `as.factor()`.
  - This is a base R solution, the `forcats` package from the tidyverse also has a `as_factor()` function.

Example: Dichotomous variable in SUPERCIAS dataset

Descriptive statistics for bivariate data

Cross-tabulation

```
  - Depending on the context, you may want to modify the `margin`
  - For example, `margin = 1` would give you the relative frequenc
  - `margin = 2` would give you the relative frequency of each col
  - The default is `margin = NULL`, which gives you the relative f
```