BUSA 603

Module 2 Supplement
Cross Tabulation Tables and Histograms

Brian Weikel

brian.weikel@franklin.edu

Business Analytics
Franklin University

Spring 2024

Cross Tabulation Table

A **cross tabulation** is a type of table for describing two variables. These variables may be categorical, discrete, or grouped continuous variables. These tables are frequently used to illustrate frequency distributions and relative frequency distributions. Such tables are also referred to as **contingency tables**.

A cross tabulation in Excel may be created using a `Pivot Table`.[1] The Python `pandas` function `crosstab` may be used to create contingency tables. R has a many packages to create contingency tables, such as `contingencytables`, `dplyr`, and `MASS`, as well as the R base function `table`.

As we saw in the Module 1 lecture notes and in a forthcoming module, cross tabulation tables are used frequently to present summary statistics in marketing analytics.

[1] Excel's `pivot table` functionality may also be used to summarize "large data" so visualizations may be created.

Qualitative and Quantitative Variables

For **qualitative data**, categories are used to divide data.

- Recall qualitative variables, also called categorical variables, classify individuals into groups or categories. For example, responses to yes/no questions on a survey, gender, and marital status, are categorical variables.
- Qualitative data includes ordinal and nominal levels of measurement.

Also recall that **quantitative variables** are numerical, and includes interval and ratio levels of measurement.

- A quantitative variable has the **ratio level of measurement** if zero represents the absence of the quantity, and ratios are meaningful.
- A quantitative variable has the **interval level of measurement** if zero does not represent the absence of the quantity, and ratios are not meaningful. Differences are meaningful, however.

Frequency Distribution Terminology

The **frequency** of a category is the number of times it occurs in the data set.

A **frequency distribution** is a table that presents the frequency for each category.

The **relative frequency** of a category is the frequency of the category divided by the sum of all the frequencies.

$$\text{Relative Frequency of Category} = \frac{\text{Frequency of Category}}{\text{Sum Frequencies Across all Categories}} \quad (1)$$

The frequency of a category is the number of category items. The relative frequency is the proportion of items in the category.

A **relative frequency distribution** is a table that presents category relative frequency. Frequently frequency is presented also.

Quantitative Variables as Classes

Frequency distributions for quantitative data are just like those for qualitative data, except that the data are divided into classes rather than categories.

- The **lower class limit** of a class is the smallest value that can appear in that class.
- The **upper class limit** of a class is the largest value that can appear in that class.
- The **class** width is the difference between consecutive lower class limits.

Guidelines for Creating Classes

FRANKLIN UNIVERSITY
Makes It Possible

Requirements for choosing classes.

- Every observation must fall into one of the classes.
- The classes must not overlap.
- The classes must be of equal width.
- There must be no gaps between classes. Even if there are no observations in a class, it must be included in the frequency distribution.

Guidelines for creating classes.

- For many data sets, the number of classes should be at least 5 but no more than 20.
- For very large data sets, a larger number of classes may be appropriate.

An Urgent Request from the CMO!!!

Suppose you are a working in the marketing department of Walt Disney Studios when an urgent request from the CMO arrives. She wants you and your colleagues to provide data for a presentation she and her vice presidents are creating.

To ensure she gets what she wants, she has provided templates for 4 tables, which are to be populated with IMBD data; the templates are on the next 4 slides. She is also requesting an Average Rating Histogram and a set of summary statistics for all movies and movies by category; the last 2 tables are templates for this request. For the weighted mean request, use number of votes as the weight. Finally, she is asking you to provide insights for each table and illustration.

Time is of the Essence!

The file `IMDB_BUSA_603.xlsx`, available in Canvas, may be used to address her request.

While You may use Excel, Python, R . . . to create the tables and histograms, the data used to create the tables and histograms must be made available in an Excel workbook, Google Sheets workbook, or set of delimited files that may be opened with Excel or Google Sheets.

As is the case for most unplanned business requests, this request needs to be done ASAP. Specifically, you and your colleagues have 1 hour to complete this task! Good luck!

Movie Category Frequency Distribution

Category	Frequency	Cumulative Frequency
Comedy		
Documentary		
Drama		
Horror		
Other		
Sci-Fi		
Grand Total		

Table 1: Movie Category Frequency Distribution Template

Movie Start Year Frequency Distribution

Start Year	Frequency	Cumulative Frequency
2017		
2018		
2019		
2020		
2021		
Grand Total		

Table 2: Movie Start Year Frequency Distribution Template

Movie Start Year Relative Frequency Distribution

Start Year	Relative Frequency	Cumulative Relative Frequency
2017		
2018		
2019		
2020		
2021		
Grand Total		

Table 3: Movie Start Year Relative Frequency Distribution Template

Average Rating Frequency Distribution

Average Rating	Frequency	Cumulative Frequency
$[1, 2)$		
$[2, 3)$		
$[3, 4)$		
$[4, 5)$		
$[5, 6)$		
$[6, 7)$		
$[7, 8)$		
$[8, 9)$		
$[9, 10)$		
Grand Total		

Table 4: Average Rating Frequency Distribution Template

Central Tendency Table to be Populated

Measure	Category	Value
Mean	All	
Weighted Mean	All	
Median	All	
Mode[2]	All	
Mean	Comedy	
Weighted Mean	Comedy	
Median	Comedy	
Mean	Documentary	
Weighted Mean	Documentary	
Median	Documentary	

[2] Though the mode is truly not a central tendency measure, it has been included in the table.

Central Tendency Table to be Populated Cont.

Measure	Category	Value
Mean	Drama	
Weighted Mean	Drama	
Median	Drama	
Mean	Horror	
Weighted Mean	Horror	
Median	Horror	
Mean	Sci-Fi	
Weighted Mean	Sci-Fi	
Median	Sci-Fi	
Mean	Other	
Weighted Mean	Other	
Median	Other	

Table 5: Average Rating Central Tendency Measures

Spread Table to be Populated

Measure	Category	Value
Range	All	
Variance	All	
Standard Deviation	All	
Range	Comedy	
Variance	Comedy	
Standard Deviation	Comedy	
Range	Documentary	
Variance	Documentary	
Standard Deviation	Documentary	

Spread Table to be Populated Continued

Measure	Category	Value
Range	Drama	
Variance	Drama	
Standard Deviation	Drama	
Range	Horror	
Variance	Horror	
Standard Deviation	Horror	
Range	Sci-Fi	
Variance	Sci-Fi	
Standard Deviation	Sci-Fi	
Range	Other	
Variance	Other	
Standard Deviation	Other	

Table 6: Average Rating Spread Measures