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

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

Creating the CMO Requested Tables and Histogram

Creating a Frequency Distribution with Excel

The following steps may be used to populate Table 1.

1. In the sheet `IMDB_BUSA_603`, select the column `Categorization`.
2. Via the ribbon under `Insert`, choose `PivotTable`. Choose `From Table/Range`. A pop-up window appears, where the value of `Table/Range` should be `IMDB_BUSA_603!$S$1:$S$381`. Select `New Worksheet`, then click `OK`. This will open a new sheet in the workbook.
3. In the box `PivotTable Fields`, drag and drop `Categorization` to `Rows`.

Creating a Frequency Distribution with Excel Cont.

4. In the box, PivotTable Fields, drag and drop Categorization to Σ Values. Click on the box of Categorization and select Value of Field Setting. select Count from the tab Summarize Values By field and click OK. Rename Row Labels to Category. Rename Count of Categorization to Frequency.
5. In the box, PivotTable Fields, drag and drop Categorization to Σ Values. Click on the box of Categorization and select Value of Field Setting. Select Count from the tab Summarize Value By field. Then under the tab Show Values As select Running Total In. Then click OK. Rename this column to Cumulative Frequency.
6. Rename the sheet to categor_frequency_dist.

Movie Category Frequency Distribution

Category	Frequency	Cumulative Frequency
Comedy	80	80
Documentary	32	112
Drama	121	233
Horror	52	285
Other	74	359
Sci-Fi	21	380
Grand Total	**380**	

Table 1: Movie Category Frequency Distribution[3]

[3] If one selected data using `IMDB_BUSA_603!$S:$S` instead of `IMDB_BUSA_603!$S$1:$S$381`, then a `(blank)` row would have been realized in the table. This row can be hidden by selecting the drop down box of `Category` and deselecting `(blank)`.

Another Excel Frequency Distribution

The following steps may be used to populate Table 2.

1. In the sheet `IMDB_BUSA_603`, select the column `startYear`.
2. Via the ribbon under `Insert`, choose `PivotTable`. Choose `From Table/Range`. A pop-up window appears, where the value of `Table/Range` should be `IMDB_BUSA_603!$M$1:$M$381`. Select `New Worksheet`, then click `OK`. This will open a new sheet in the workbook.
3. In the box `PivotTable Fields`, drag and drop `startYear` to `Rows`.
4. In the box, `PivotTable Fields`, drag and drop `startYear` to Σ `Values`. Click on the box of `startYear` and select `Value of Field Setting`; select `Count` and click `OK`.

Another Excel Frequency Distribution Cont.

5. Rename `Row Labels` to `Start Year`. Rename `Count of Categorization` to `Frequency`.
6. In the box, `PivotTable Fields`, drag and drop `Categorization` to Σ `Values`. Click on the box of `Categorization` and select `Value of Field Setting`. Select `Count` from the tab `Summarize Values By` field. Then under the tab `Show Values As` select `Running Total In`. Then click `OK`. Rename this column to `Cumulative Frequency`.
7. Rename the sheet to `startYear_frequency_dist`.

Movie Start Year Frequency Distribution

Start Year	Frequency	Cumulative Frequency
2017	54	54
2018	72	126
2019	144	270
2020	57	327
2021	53	380
Grand Total	**380**	

Table 2: Movie Start Year Frequency Distribution

If the Tables and Figures of this deck were to be presented in a business setting, they would be refined by further modifying labels, titles, axis names, font sizes, chart colors . . . in effort to make them "more appealing" to the audience.

Creating a Relative Frequency Distribution with Excel

The following steps may be used to populate Table 3.

1. Copy the sheet to `startYear_frequency_dist`. Rename the sheet to `startYear_r_frequency_distr`.
2. In the sheet `startYear_r_frequency_distr`, right-click on `Frequency`, select `Show Value As`. Select `% of Column Total`. Select column `B`. In the ribbon under `Home`, find the `General` drop down box above the `Number` option. In the drop down box, select `Number`. Replace `Frequency` with `Relative Frequency`.
3. Right-click on `Cumulative Frequency`, under `Show Value As` select `% Running Total In`. For the pop-up window that appears, select use `startYear` for `Base Field`. Click `OK`. In the ribbon under `Home`, find the `General` drop down box above the `Number` option.

Movie Start Year Relative Frequency Distribution

In the drop down box, select `Number`. Rename the column to `Cumulative Relative Frequency`.

4 You have created the `startYear` relative frequency distribution.

Start Year	Relative Frequency	Cumulative Relative Frequency
2017	0.14	0.14
2018	0.18	0.33
2019	0.38	0.71
2020	0.15	0.86
2021	0.14	1.00
Grand Total	**1.00**	

Table 3: Movie Start Year Relative Frequency Distribution

Ratio Level of Measurement Variable Classes

The following steps may be used to populate Table 4. Not that `averageRating` has the ratio level of measurement.

1. In the sheet `IMDB_BUSA_603`, select the column `averageRating`.
2. Via the ribbon under `Insert`, choose `PivotTable`. Choose `From Table/Range`. A pop-up window appears, where the value of `Table/Range` should be `IMDB_BUSA_603!$Q$1:$Q$381`. Select `New Worksheet`, then click `OK`. This will open a new sheet in the workbook. Name this sheet `averageRating_frequency_dist`.
3. In the box `PivotTable Fields`, drag and drop `averageRating` to `Rows`.

Ratio Level of Measurement Variable Classes Cont.

4. Note there are more then 60 classes; having so many classes typically hinders the informative value of graphically displayed data. Thus we will now combine some classes. Right-click on any numeric class value. Select `Group`. In the `Grouping` pop-up window, set `Starting at` to `1`, `Ending at` to `10`, and `By` to `1`. Click on OK.

5. In the box, `PivotTable Fields`, drag and drop `averageRating` to Σ `Values`. Click on the box of `averageRating` and select `Value of Field Setting`; select `Count` and click `OK`. Rename `Row Labels` to `Average Rating`. Rename `Count of averageRating` to `Frequency`.

Ratio Level of Measurement Variable Classes Cont.

6. In the box, PivotTable Fields, drag and drop averageRating to Σ Values. Click on the box of averageRating, select Summarize Values By of Value of Field Setting, and then select Count. Select Running Total in under Show Values As. Then click OK. Rename this column to Cumulative Frequency.

Average Rating Frequency Distribution

Average Rating	Frequency	Cumulative Frequency
$[1, 2)$	2	2
$[2, 3)$	10	12
$[3, 4)$	21	33
$[4, 5)$	52	85
$[5, 6)$	81	166
$[6, 7)$	122	288
$[7, 8)$	67	355
$[8, 9)$	24	379
$[9, 10)$	1	380
Grand Total	**380**	

Table 4: Average Rating Frequency Distribution

Note that the values of the `Average Rating` column in the Excel workbook are slightly different than those provided above. In Table 4 we have identified when an integer yields a closed or open bound.

Histograms

Once we have a frequency distribution or a relative frequency distribution, we can put the information in graphical form by constructing a **histogram**.

Histograms based on frequency distributions are called **frequency histograms**, and histograms based on relative frequency distributions are called **relative frequency histograms**.

Histograms are related to bar graphs, and are appropriate for quantitative data. A histogram is constructed by drawing a rectangle for each class. The heights of the rectangles are equal to the frequencies or the relative frequencies, and the widths are equal to the class width.

The purpose of a histogram is to give a visual impression of the "shape" of a data set.

Creating a Frequency Histogram with Excel

The following steps may be used to create the CMO requested histogram.

1. In the sheet `IMDB_BUSA_603`, select the column `averageRating`.
2. Via the ribbon under `Insert`, choose `Recommended Charts`. Under `All Charts`, choose `Histogram`. Of the two diagrams that appear at the top of the pop-up window, choose `Histogram`. Click `OK`.
3. Right-click in the graph. Select `Move Chart`; ensure the `New sheet` option is chosen and name the new sheet `averageRating_Histogram`.
4. Click on the chart's title and replace the exiting title with `Average Rating Histogram`. You have created an `averageRating` histogram.

Average Rating Histogram

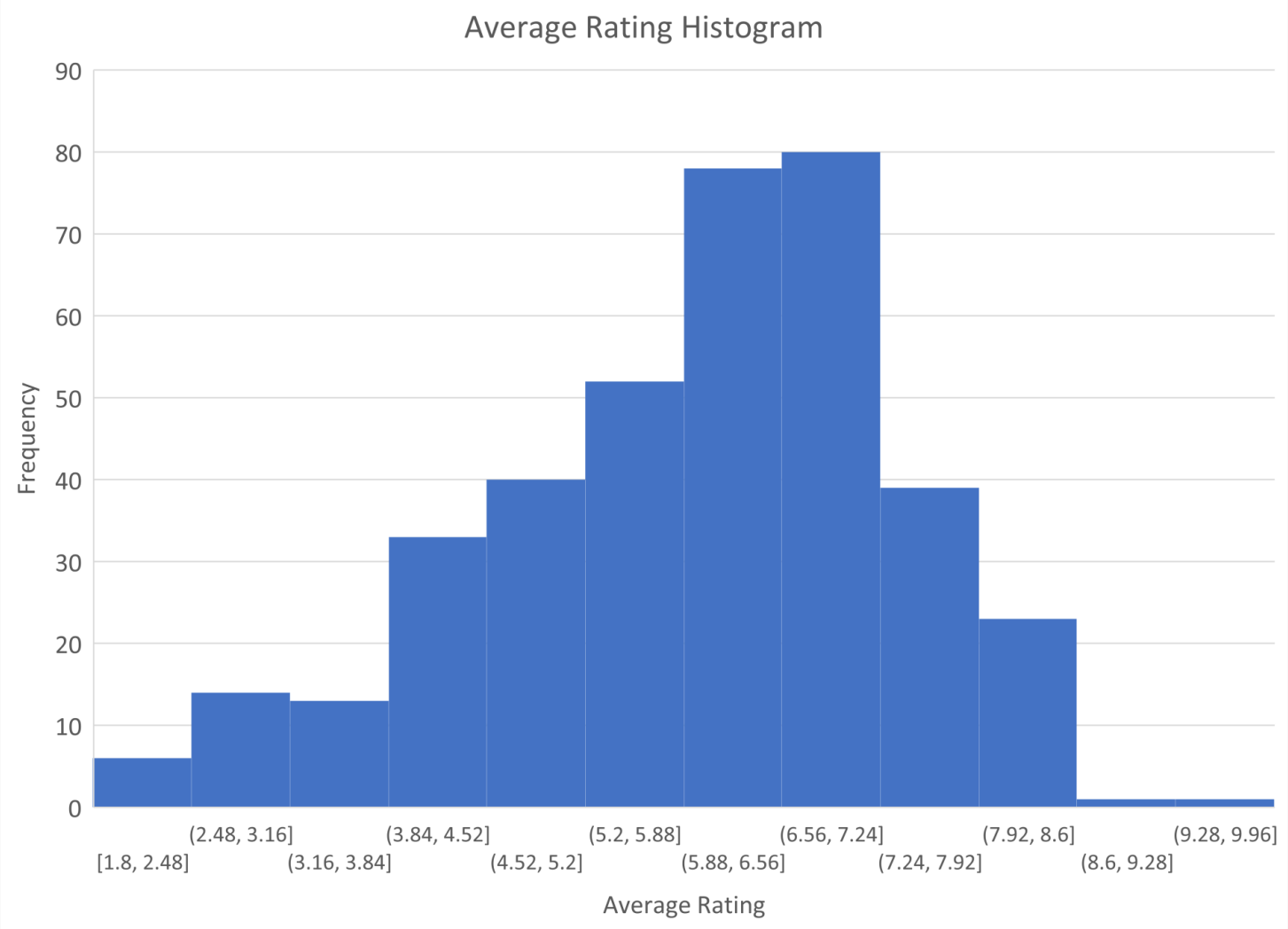

Figure 1: Average Rating Histogram

Shapes of Histograms

The following applies to frequency histograms and relative frequency histograms.

A histogram is **skewed** if one side, or tail, is longer than the other. A histogram with a long right-hand tail is said to be **skewed to the right**, or **positively skewed**. A histogram with a long left-hand tail is said to be **skewed to the left**, or **negatively skewed**.

A histogram is **symmetric** if its right half is a mirror image of its left half. Very few histograms are perfectly symmetric, but many are approximately symmetric.

- A symmetric histogram with a peak in the middle is referred to as a **bell-shaped histogram**.
- A histogram in which all the classes have equal frequencies is said to be **uniformly distributed**.

Mode

A peak, or high point, of a histogram is referred to as a **mode**. The mode, if one exists, is thus the most frequently occurring value.

A histogram is **unimodal** if it has only one mode, and **bimodal** if it has two clearly distinct modes.

While some textbooks refer to the mode as a measure of **central tendency**, I recommend not using it as such. The mean and median are preferred measures of central tendency.

Misleading Histograms

We know that the width of all histogram intervals, at times referred to as **bins**, should be the same.

Suppose a data set contains many observations that fall into a relatively narrow part of the range, whereas others are widely dispersed. We might be tempted to construct a frequency distribution with narrow intervals where the bulk of the observations are and broader ones elsewhere.

Even if we remember that it is the areas, rather than the heights, of the rectangles of the histogram that must be proportional to the frequencies, it is still never a desirable option to construct such a histogram with different bin widths because it may easily deceive or distort the findings. We included this section simply to point out potential errors that we might find in histograms.

Creating the Central Tendency Table in Excel

Firstly, in your Excel workbook `IMDB_BUSA_603` create a sheet `Central_Tendency` by clicking the $\bigoplus$ at the bottom of the workbook.[4]

- In cell `A1` type `Measure`. In cell `B1` type Category. In cell `C1` type `Value`. For each typed value, place it in bold font.
- Then type entries for `Measures` and `Category` in cells `A2` through `B23` to produce Table 5 as illustrated above.

[4] All numerical summaries and illustrations presented in these lecture notes are available in the Excel workbook `IMDB_BUSA_603_Module_2`.

Calculating in Excel Central Tendency Measures

We will place the requested summary statistics in the tab `Central_Tendency`.

1. In cell `C2` type `=ROUND(AVERAGE(IMDB_BUSA_603!$Q$2:$Q$381),2)` to calculate the mean using all observations. By using the `ROUND` command we are maintaining 2 decimal points for the statistic.
2. In cell `C3` type `=ROUND(SUMPRODUCT(IMDB_BUSA_603!$Q$2:$Q$381, IMDB_BUSA_603!$R$2:$R$381) /SUM(IMDB_BUSA_603!$R$2:$R$381),2)` to calculate the weighted mean using all observations, where `numVotes` is the weight variable.
3. In cell `C4` type `=ROUND(MEDIAN(IMDB_BUSA_603!$Q$2:$Q$381),2)` to calculate the median using all observations.
4. In cell `C5` type `=MODE.SNGL(IMDB_BUSA_603!$Q$2:$Q$381)` to calculate the mode using all observations.

Calculating in Excel Group-Level Means . . .

5. In cell C6 type `=ROUND(AVERAGEIF(IMDB_BUSA_603!$S$2:$S$381,B6, IMDB_BUSA_603!$Q$2:$Q$381),2)` to calculate the mean for Comedy observations. The cell value for B6 of the formula resolves to ‘‘Comedy’’.
6. In cell C7 type `=ROUND(SUMPRODUCT(IF(IMDB_BUSA_603!$S$2:$S$381=B7, IMDB_BUSA_603!$Q$2:$Q$381*IMDB_BUSA_603!$R$2:$R$381)) /SUMIF(IMDB_BUSA_603!$S$2:$S$381,B7, IMDB_BUSA_603!$R$2:$R$381),2)` to calculate the weighted mean for Comedy observations where numVotes is the weight variable.
7. In cell C8 type `=MEDIAN(IF(IMDB_BUSA_603!$S$2:$S$381=B8, IMDB_BUSA_603!$Q$2:$Q$381))` to calculate the median for Comedy observations.
8. In cells C9 through C23 complete steps 5, 6, and 7 above for the other values of Category in the table.

Average Rating Central Tendency Measures

Measure	Category	Value
Mean	All	5.95
Weighted Mean	All	7.3
Median	All	6.1
Mode	All	6
Mean	Comedy	6.03
Weighted Mean	Comedy	7.18
Median	Comedy	6.1
Mean	Documentary	7.27
Weighted Mean	Documentary	7.44
Median	Documentary	7.15

Average Rating Central Tendency Measures Cont.

Measure	Category	Value
Mean	Drama	6.14
Weighted Mean	Drama	7.69
Median	Drama	6.3
Mean	Horror	4.74
Weighted Mean	Horror	6.87
Median	Horror	4.75
Mean	Sci-Fi	5.79
Weighted Mean	Sci-Fi	7.33
Median	Sci-Fi	6
Mean	Other	5.86
Weighted Mean	Other	7.21
Median	Other	5.7

Table 5: Average Rating Central Tendency Measures

Calculating in Excel Spread Measures

We will place the requested spread summary statistics in the tab `Spread`, where the tab is created in a manner analagous to that used to create `Central_Tendency`.

Recall that the IMDb data is a sample of the population of movies. Thus we will calculate sample measures of spread.

1. In cell `C2` type `=ROUND(MAX(IMDB_BUSA_603!$Q$2:$Q$381)-MIN(IMDB_BUSA_603!$Q$2:$Q$381),3)` to calculate the range using all observations. The invoked `ROUND` command means we are maintaining 3 decimal points for the statistic.
2. In cell `C3` type `=ROUND(VAR.S(IMDB_BUSA_603!$Q$2:$Q$381),3)` to calculate the variance using all observations.
3. In cell `C4` type `=ROUND(STDEV.S(IMDB_215!$Q$2:$Q$381),3)` to calculate the standard deviation using all observations. In lieu of this formula, one may have use `=ROUND(SQRT(C3),3)` to calculate the measure.

Calculating in Excel Group-Level Ranges . . .

4. In cell C5 type =ROUND(MAXIFS(IMDB_BUSA_603!$Q$2:$Q$381, IMDB_BUSA_603!$S$2:$S$381,B6)-MINIFS(IMDB_BUSA_603!$Q$2:$Q$381, IMDB_BUSA_603!$S$2:$S$381,B5),3) to calculate the range for Comedy observations. The cell value for B5 of the formula resolves to ‘‘Comedy’’.
5. In cell C6 type =ROUND(((SUMPRODUCT(IF(IMDB_BUSA_603!$S$2:$S$381=B6, IMDB_BUSA_603!$Q$2:$Q$381*IMDB_BUSA_603!$Q$2:$Q$381))) -COUNTIF(IMDB_BUSA_603!$S$2:$S$381,B6)*AVERAGEIF(IMDB_BUSA_603!$S$2:$S$381,B6, IMDB_BUSA_603!$Q$2:$Q$381)^2) /COUNTIF(IMDB_BUSA_603!$S$2:$S$381,B6),3) to calculate the variance for Comedy observations.
6. In cell C7 type =ROUND(SQRT(C6),3) to calculate the standard deviation for Comedy observations.

Average Rating Spread Measures

7. In cells `C8` through `C22` complete steps 4, 5, and 6 above for the other values of `Category` in the table.

Measure	Category	Value
Range	All	7.6
Variance	All	1.991
Standard Deviation	All	1.411
Range	Comedy	6.4
Variance	Comedy	1.645
Standard Deviation	Comedy	1.283
Range	Documentary	2.4
Variance	Documentary	0.469
Standard Deviation	Documentary	0.685

Average Rating Spread Measures Continued

Measure	Category	Value
Range	Drama	6.7
Variance	Drama	1.366
Standard Deviation	Drama	1.169
Range	Horror	5.8
Variance	Horror	1.953
Standard Deviation	Horror	1.397
Range	Sci-Fi	6.2
Variance	Sci-Fi	1.958
Standard Deviation	Sci-Fi	1.399
Range	Other	7.1
Variance	Other	2.194
Standard Deviation	Other	1.481

Table 6: Average Rating Spread Measures