2

SUMMARY STATISTICS

Perhaps one of the most important elements of conducting high-quality empirical research is to have a strong understanding of the data that are being used in the study. Similarly, for a reader of empirical research, to fully comprehend the results of the study and assess the applicability of these results beyond the scope of the study, it is important to have at least a cursory understanding of the data upon which the analyses presented in the article were performed. For these reasons, most empirical research papers present summaries of the data prior to discussing the main results. Frequently, the first table of a research paper presents such a summary.

In this chapter, we present the most commonly used approach in the empirical asset pricing literature to calculating and presenting summary statistics. Effective presentation of summary statistics represents a trade-off between showing enough results to give the reader a good sense of the important characteristics of the data and not presenting so much that the reader is overwhelmed. The optimal approach to presenting summary statistics depends greatly on the type of study being conducted. The approach presented in this chapter is most appropriate when the objective of the study is to understand a cross-sectional phenomenon of the entities (stocks, bonds, firms, etc.) being studied. The procedure, therefore, is geared toward understanding the cross-sectional distribution of the variables used in the study.

Empirical Asset Pricing: The Cross Section of Stock Returns, First Edition.
Turan G. Bali, Robert F. Engle, and Scott Murray.
© 2016 John Wiley & Sons, Inc. Published 2016 by John Wiley & Sons, Inc.

2.1 IMPLEMENTATION

The summary statistics procedure consists of two steps. In the first step, for each time period t, certain characteristics of the cross-sectional distribution of the given variable, X, are calculated. In the second step, the time-series properties of the periodic cross-sectional characteristics are calculated. In most cases, the time-series property of interest is the mean, in which case the final results that are presented represent the average cross section, where the average is taken over all periods t during the sample period.

2.1.1 Periodic Cross-Sectional Summary Statistics

The details of the first step are as follows. For each time period t, we calculate the cross-sectional mean, standard deviation, skewness, excess kurtosis, minimum value, median value, maximum value, and selected additional percentiles of the distribution of the values of X, where each of these statistics is calculated over all available values of X in period t. We let $Mean_t$ be the mean, SD_t denote the sample standard deviation, $Skew_t$ represent the sample skewness, $Kurt_t$ be the sample excess kurtosis, Min_t be the minimum value, $Median_t$ denote the median value, and Max_t represent the maximum value of X in period t. In addition, we will record the fifth, 25th, 75th, and 95th percentiles of X in month t, which we denote $P5_t$, $P25_t$, $P75_t$, and $P95_t$, respectively. Depending on the data and the objective of the study, it may be desirable to include additional percentiles of the distribution. For example, if the study focuses on extreme values of X, then it may be valuable to record the first, second, third, fourth, 96th, 97th, 98th, and 99th percentiles of the distribution as well. Alternatively, calculating the minimum, maximum, fifth percentile, and 95th percentile of the data may not be necessary if the data are reasonably well behaved. Exactly which statistics to record and present is a decision made by the researcher, who, presumably, has a much deeper understanding of the data than could possibly be presented in a research article. In addition to these statistics describing the time t cross-sectional distribution of X, we also record the number of entities for which a valid value of X is available in period t and denote this number n_t.

In Table 2.1, we present the annual summary statistics for market beta (β) from our methodologies sample. The results show that, for example, in 1988, the average β of the stocks in the sample is 0.46; the cross-sectional standard deviation of the values of β is 0.48; the sample skewness of β is 0.17; and the sample excess kurtosis of β is 2.80. Furthermore, the minimum, fifth percentile, 25th percentile, median, 75th percentile, 95th percentile, and maximum values of β in 1988 are −4.29, −0.20, 0.13, 0.40, 0.75, 1.31, and 3.28, respectively. Finally, there are 5690 stocks with a valid value of β in 1988.

Table 2.1 presents a detailed account of the cross-sectional distribution of β on a period-by-period basis. In this case, presenting the periodic summary statistics in detail is possible because our sample consists of only 25 periods, and we only present summary statistics for one variable, β. While it is certainly valuable to present all of these statistics, in most empirical asset pricing studies, the sample has many more

TABLE 2.1 Annual Summary Statistics for β

This table presents summary statistics for β for each year during the sample period. For each year t, we calculate the mean ($Mean_t$), standard deviation (SD_t), skewness ($Skew_t$), excess kurtosis ($Kurt_t$), minimum (Min_t), fifth percentile ($P5_t$), 25th percentile ($P25_t$), median ($Median_t$), 75th percentile ($P75_t$), 95th percentile ($P95_t$), and maximum (Max_t) values of the distribution of β across all stocks in the sample. The sample consists of all U.S.-based common stocks in the Center for Research in Security Prices (CRSP) database as of the end of the given year t and covers the years from 1988 through 2012. The column labeled n_t indicates the number of observations for which a value of β is available in the given year.

t	$Mean_t$	SD_t	$Skew_t$	$Kurt_t$	Min_t	$P5_t$	$P25_t$	$Median_t$	$P75_t$	$P95_t$	Max_t	n_t
1988	0.46	0.48	0.17	2.80	−4.29	−0.20	0.13	0.40	0.75	1.31	3.28	5690
1989	0.46	0.53	0.15	1.88	−3.51	−0.27	0.11	0.40	0.79	1.38	3.63	5519
1990	0.58	0.59	0.23	1.14	−3.15	−0.24	0.16	0.51	0.96	1.61	3.66	5409
1991	0.57	0.61	0.23	1.96	−3.28	−0.29	0.17	0.52	0.95	1.62	5.29	5303
1992	0.65	0.83	0.34	6.10	−5.21	−0.50	0.17	0.59	1.09	2.05	9.90	5389
1993	0.62	0.77	−0.10	4.29	−4.70	−0.56	0.20	0.57	1.04	1.90	7.59	5670
1994	0.70	0.71	−0.17	6.59	−6.92	−0.32	0.27	0.67	1.07	1.89	6.50	6148
1995	0.64	0.84	0.30	5.17	−6.32	−0.49	0.19	0.56	1.02	2.15	8.77	6288
1996	0.67	0.64	0.46	1.97	−4.32	−0.20	0.26	0.59	1.01	1.89	3.98	6586
1997	0.53	0.48	0.39	1.46	−2.36	−0.13	0.21	0.48	0.80	1.38	3.20	6867
1998	0.71	0.51	0.49	0.95	−1.80	0.01	0.34	0.67	1.03	1.62	3.75	6608
1999	0.41	0.50	1.39	4.81	−2.21	−0.18	0.11	0.32	0.61	1.33	3.77	6097
2000	0.70	0.72	1.27	1.33	−1.10	−0.06	0.19	0.49	1.01	2.23	3.76	5901
2001	0.76	0.73	1.29	2.13	−1.48	−0.05	0.25	0.60	1.07	2.25	4.21	5508
2002	0.67	0.55	0.70	0.69	−1.19	−0.04	0.25	0.62	0.97	1.73	2.99	5099
2003	0.72	0.56	0.40	0.49	−2.17	−0.04	0.29	0.68	1.06	1.72	3.04	4737
2004	1.03	0.70	0.43	0.24	−1.75	0.01	0.53	0.99	1.46	2.30	4.02	4574
2005	0.95	0.64	0.00	−0.17	−1.60	−0.06	0.46	0.99	1.39	1.96	3.69	4495
2006	1.02	0.70	0.08	0.17	−3.71	−0.02	0.48	1.00	1.51	2.18	3.75	4453
2007	0.87	0.54	−0.04	−0.20	−1.50	0.01	0.45	0.91	1.26	1.72	3.06	4332
2008	0.87	0.53	0.17	0.06	−1.49	0.03	0.48	0.87	1.22	1.74	3.45	4264
2009	1.10	0.72	0.51	0.62	−1.74	0.09	0.55	1.03	1.57	2.36	5.31	3977
2010	1.04	0.54	−0.06	−0.15	−0.85	0.10	0.68	1.05	1.41	1.90	2.95	3805
2011	1.07	0.54	−0.14	−0.37	−0.62	0.14	0.70	1.13	1.45	1.93	3.03	3682
2012	1.04	0.57	0.04	0.48	−2.33	0.11	0.66	1.05	1.40	1.99	3.43	3545

periods than the 25 periods in the methodology sample. Presenting results such as those in Table 2.1 when there are a large number of periods will not only make it difficult to display the periodic summary statistics but will also make it difficult for the reader to get a general understanding of the characteristics of the data. These issues are magnified when, as in most studies, showing summary statistics for several variables is desirable. Thus, while there are certainly interesting patterns to be observed by presenting such a detailed account of each variable, doing so is usually not necessary to inform a reader about the most salient characteristics of the data, and thus most articles present statistics that are substantially more summarized than the results in Table 2.1. We proceed now to describe how to further summarize the periodic cross-sectional summary statistics.

2.1.2 Average Cross-Sectional Summary Statistics

The second step in the summary statistics procedure is to calculate the time-series averages of the periodic cross-sectional values. For example, the average cross-sectional mean of the variable X, which we denote *Mean* (no subscript), is found by taking the time-series average of the values of $Mean_t$ over all periods t in the sample. Similarly, we calculate the times-series means of the other cross-sectional summary statistics.

For most studies, it is these time-series average values that are presented in the research article. These values describe the average cross section in the sample. This is appropriate when the objective of the study is to examine a cross-sectional phenomenon, as is the case for the analyses in this book. Table 2.2 presents the time-series averages of the annual cross-sectional summary statistics for β. The numbers in the table, therefore, represent the cross-sectional distribution of β for the average year in the methodologies sample. As can be seen, in the average year, the mean value of β is 0.75 and the median value of β is 0.71. Consistent with the mean being slightly greater than the median, in the average year, the skewness of the distribution of β of 0.34 is slightly positive. The cross-sectional distribution of β, in the average year, is leptokurtic because the average excess kurtosis of 1.78 is positive. The average cross-sectional standard deviation of β is 0.62. Finally, in the average year, there are 5198 stocks for which there is a valid value of β.

TABLE 2.2 Average Cross-Sectional Summary Statistics for β

This table presents the time-series averages of the annual cross-sectional summary statistics for β. The table presents the average mean (*Mean*), standard deviation (*SD*), skewness (*Skew*), excess kurtosis (*Kurt*), minimum (*Min*), fifth percentile (*P*5), 25th percentile (*P*25), median (*Median*), 75th percentile (*P*75), 95th percentile (*P*95), and maximum (*Max*) values of the distribution of β, where the average is taken across all years in the sample. The column labeled *n* indicates the average number of observations for which a value of β is available.

Mean	*SD*	*Skew*	*Kurt*	*Min*	*P*5	*P*25	*Median*	*P*75	*P*95	*Max*	*n*
0.75	0.62	0.34	1.78	−2.78	−0.13	0.33	0.71	1.12	1.85	4.40	5198

2.2 PRESENTATION AND INTERPRETATION

In most studies, there are many variables for which summary statistics should be presented. It is usually optimal to present the summary statistics for all variables in a single table. While each paper will present summary statistics in a slightly different manner, the approach we take in this book is to compile a table in which each row (with the exception of the header row) presents summary statistics for one of the variables.

Table 2.3 gives an example of how we present summary statistics throughout this text. The first column indicates the variable whose summary statistics are presented in the given row. The subsequent columns present the time-series averages of the cross-sectional summary statistics.

The objectives in analyzing the summary statistics are twofold. First, the summary statistics are intended to give a basic overview of the cross-sectional properties of the variables that will be used in the study. This is useful for understanding the types of entities that comprise the sample. Second, the summary statistics can be used to identify any potential issues that may arise when using these variables in statistical analyses. We exemplify how the summary statistics can be used for each of these objectives in the following two paragraphs using the methodology sample and the results in Table 2.3.

The mean column in Table 2.3 can roughly be interpreted as indicating that the average stock in our sample has a β of 0.75, a market capitalization of just over \$2 billion, and a book-to-market ratio of 0.71. More precisely, the table indicates that in the average month, the cross-sectional means of the given variables are as indicated in the table, but we frequently adopt the simpler language used in the previous sentence. The average value of *Size*, which is the natural log of *MktCap*, is 5.04, and the average one-year-ahead excess return is 12.40%.

Table 2.3 shows that for β, the mean and the median are quite similar and, consistent with this, the skewness is quite small in magnitude and values of β are reasonably symmetric about the mean. The distribution of β is also slightly leptokurtic as the excess kurtosis of its cross-sectional distribution in the average year is 1.78.

The results for *MktCap* show that the distribution of market capitalization is highly positively skewed. This is driven by a small number of observations that have very large values of *MktCap*. The summary statistics therefore indicate that the sample is comprised predominantly of low-market capitalization stocks along with a few stocks that have very high market capitalizations. The median stock in the sample has a market capitalization of \$188 million, which is much smaller than the mean of more than \$2 billion. It is also worth noting that the smallest value of *MktCap* of 0, which means that the stock has market capitalization of less than \$0.5 million, is less than 0.02 standard deviations from the median and less than 0.1 standard deviations from the mean. This indicates that a very large portion of the variability of *MktCap* comes from extremely large values, consistent with the high positive skewness. The distribution of *MktCap* presents potential issues for statistical analyses, such as regression, that rely on the magnitude of the variables used, as the data points corresponding to the very large values may exert undesirably strong influence on the results of such analyses. Therefore, most empirical studies use *Size*, defined as the natural log of *MktCap*, in such analyses. Table 2.3 shows that the distribution of *Size* is much more symmetric than that of *MktCap*, as the average skewness is only 0.32. Furthermore, the excess kurtosis of −0.07 indicates that tails of the distribution of *Size* are, in the average year, very similar to those of a normal distribution. *Size*, therefore, appears much better suited for use in statistical analyses than *MktCap*.

TABLE 2.3 Summary Statistics for β, *MktCap*, and *BM*

This table presents summary statistics for our sample. The sample covers the years t from 1988 through 2012, inclusive, and includes all U.S.-based common stocks in the CRSP database. Each year, the mean (*Mean*), standard deviation (*SD*), skewness (*Skew*), excess kurtosis (*Kurt*), minimum (*Min*), fifth percentile (5%), 25th percentile (25%), median (*Median*), 75th percentile (75%), 95th percentile (95%), and maximum (*Max*) values of the cross-sectional distribution of each variable are calculated. The table presents the time-series means for each cross-sectional value. The column labeled n indicates the average number of stocks for which the given variable is available. β is the beta of a stock calculated from a regression of the excess stock returns on the excess market returns using all available daily data during year t. *MktCap* is the market capitalization of the stock calculated on the last trading day of year t and recorded in $millions. *Size* is the natural log of *MktCap*. *BM* is the ratio of the book value of equity to the market value of equity. r_{t+1} is the one-year-ahead excess stock return.

	Mean	*SD*	*Skew*	*Kurt*	*Min*	5%	25%	*Median*	75%	95%	*Max*	*n*
β	0.75	0.62	0.34	1.78	−2.78	−0.13	0.33	0.71	1.12	1.85	4.40	5198
MktCap	2030	10,230	14.20	282.85	0	9	48	188	802	7524	287,033	5550
Size	5.04	2.07	0.32	−0.07	−1.19	1.89	3.56	4.91	6.39	8.70	12.33	5550
BM	0.71	2.90	−9.49	1,226.68	−124.31	0.05	0.29	0.57	0.97	2.11	44.87	4273
r_{t+1}	12.40	80.83	5.94	125.33	−97.86	−67.46	−26.87	0.90	31.84	124.54	1,841.43	5381

As for the book-to-market ratio (*BM*), Table 2.3 shows that while the vast majority of the *BM* values fall between 0.05 (the fifth percentile) and 2.11 (the 95th percentile), the tails of the distribution are extremely long, as the kurtosis of *BM* is greater than 1226. Interestingly, despite the fact that the mean is greater than the median, in the average month, the distribution of *BM* is negatively skewed, as the average cross-sectional skewness of *BM* is −9.49.

Finally, the table indicates that the average one-year-ahead excess return (r_{t+1}) of the stocks in the methodology sample is 12.40% per year. The cross-sectional distribution of r_{t+1} is highly skewed and leptokurtic, with an average skewness of 5.94 and excess kurtosis of more than 125. This is driven by the fact that the minimum possible return is −100%, whereas there is no upper bound on the value that r_{t+1} can take. Table 2.3 shows that, in the average year, the maximum r_{t+1} is more than 1841%, with more than 5% of stock realizing excess returns greater than 100%.

There is one more aspect of the return data that is worth mentioning because it is not apparent in the presentation of the summary statistics. The latest data in the version of the CRSP database used to construct the methodology sample are from 2012. However, when t corresponds to year 2012, then r_{t+1} corresponds to excess returns from 2013. Unfortunately, return data for 2013 are not available. Thus, the summary statistics for r_{t+1} reported in Table 2.3 actually cover returns for the 24 years from 1989 through 2012, whereas the summary statistics for the other variables cover the 25 years from 1988 through 2012. While this detail of the summary statistics is not usually discussed in a research article because it rarely has a meaningful impact on the interpretation of the results, it is something that should be clearly understood by the researcher.

Although Table 2.3 is certainly expository, there are many characteristics of the data that are not captured in the highly summarized results. The most important drawback of summarizing the data in such a manner is that it does not indicate any time-series variation in the variables used in the study. For example, referring back to Table 2.1, it is evident that the mean and median values of β increase quite substantially over time. This feature of the data is not in any way captured in the summary presented in Table 2.3. Additionally, given that the values of market capitalization (*MktCap*) have not been adjusted for inflation, it is reasonable to assume that value of *MktCap* will exhibit generally increasing pattern over time as well. This is confirmed in unreported results. Furthermore, values of *MktCap* are likely to drop when the stock market experiences a large loss and increase as the stock market realizes gains. The opposite would be true for the book-to-market ratio (*BM*) as the market capitalization is the denominator of this variable, although in this case the increase in values of *BM* may be delayed due to the timing of the calculation of *BM*, which is discussed in detail in Chapter 10.

None of these characteristics of the data are captured in the summary statistics as presented in Table 2.3. In most cases, these details are not very important when interpreting and drawing conclusions from the results of subsequent analyses in the article. However, as a researcher, it is important to be aware of such patterns and to assess whether these patterns may have a significant impact on the main conclusions of the study. In many cases, this is done by subsample analyses aimed at examining whether the main conclusions hold in both early periods of the study as well as in late periods. Frequently, it is also worthwhile to investigate whether the main results hold in periods of normal economic conditions as well as periods of deteriorating or poor economic conditions. This is especially the case if the summary statistics for the focal variables of the study are substantially different for these subperiods.

2.3 SUMMARY

In summary, the main objective of presenting summary statistics is to give the reader a sufficient but succinct understanding of the data being used and the characteristics of the entities that comprise the sample. In addition, the summary statistics can be used to identify and remedy any potential issues with using statistical analysis on the data. The approach that we have discussed presents the distribution of the given variables in the average cross section. While the results presented in the summary statistics table may be sufficient for a reader, they are likely not sufficient for the researcher. It is difficult to conduct high-quality research without having an in-depth understanding of the data. A good researcher will understand any potential issues with the data that are not evident in the summary statistics and address these issues in the statistical analyses presented in the research article.