A more realistic formulation is to assume that P_t^* follows a random walk so that $\Delta P_t^* = P_t^* - P_{t-1}^* = \epsilon_t$, which forms a sequence of independent and identically distributed random variables with mean zero and variance σ^2. In addition, $\{\epsilon_t\}$ is independent of $\{I_t\}$. In this case, $\text{Var}(\Delta P_t) = \sigma^2 + S^2/2$, but $\text{Cov}(\Delta P_t, \Delta P_{t-j})$ remains unchanged. Therefore,

$$\rho_1(\Delta P_t) = \frac{-S^2/4}{S^2/2 + \sigma^2} \leq 0.$$

The magnitude of the lag-1 autocorrelation of ΔP_t is reduced, but the negative effect remains when $S = P_a - P_b > 0$. In finance, it might be of interest to study the components of the bid–ask spread. Interested readers are referred to Campbell, Lo, and MacKinlay (1997) and the references therein.

The effect of bid–ask spread continues to exist in portfolio returns and in multivariate financial time series. Consider the bivariate case. Denote the bivariate order-type indicator by $I_t = (I_{1t}, I_{2t})'$, where I_{1t} is for the first security and I_{2t} for the second security. If I_{1t} and I_{2t} are contemporaneously positively correlated, then the bid–ask spreads can introduce negative lag-1 cross-correlations.

5.3 EMPIRICAL CHARACTERISTICS OF TRANSACTIONS DATA

Let t_i be the calendar time, measured in seconds from midnight, at which the ith transaction of an asset takes place. Associated with the transaction are several variables such as the transaction price, the transaction volume, the prevailing bid and ask quotes, and so on. The collection of t_i and the associated measurements are referred to as the *transactions data*. These data have several important characteristics that do not exist when the observations are aggregated over time. Some of the characteristics are given next.

1. *Unequally Spaced Time Intervals*. Transactions such as stock tradings on an exchange do not occur at equally spaced time intervals. As such, the observed transaction prices of an asset do not form an equally spaced time series. The time duration between trades becomes important and might contain useful information about market microstructure (e.g., trading intensity).
2. *Discrete-Valued Prices*. The price change of an asset from one transaction to the next only occurs in multiples of tick size. On the NYSE, the tick size was one-eighth of a dollar before June 24, 1997 and was one-sixteenth of a dollar before January 29, 2001. All NYSE and AMEX stocks started to trade in decimals on January 29, 2001. Therefore, the price is a discrete-valued variable in transactions data. In some markets, price change may also be subject to limit constraints set by regulators.
3. *Existence of a Daily Periodic or Diurnal Pattern*. Under the normal trading conditions, transaction activity can exhibit a periodic pattern. For instance, on the NYSE, transactions are "heavier" at the beginning and closing of the trading hours and "thinner" during lunch hour, resulting in a U-shape

transaction intensity. Consequently, time durations between transactions also exhibit a daily cyclical pattern.

4. *Multiple Transactions Within a Single Second*. It is possible that multiple transactions, even with different prices, occur at the same time. This is partly due to the fact that time is measured in seconds that may be too long a time scale in periods of heavy trading.

To demonstrate these characteristics, we consider first the IBM transactions data from November 1, 1990 to January 31, 1991. These data are from the Trades, Orders Reports, and Quotes (TORQ) dataset; see Hasbrouck (1992). There are 63 trading days and 60,328 transactions. To simplify the discussion, we ignore the price changes between trading days and focus on the transactions that occurred in the normal trading hours from 9:30 am to 4:00 pm Eastern time. It is well known that overnight stock returns differ substantially from intraday returns; see Stoll and Whaley (1990) and the references therein. Table 5.1 gives the frequencies in percentages of price change measured in the tick size of \$1/8 = \$0.125. From the table, we make the following observations:

1. About two-thirds of the intraday transactions were without price change.
2. The price changed in one tick approximately 29% of the intraday transactions.
3. Only 2.6% of the transactions were associated with two-tick price changes.
4. Only about 1.3% of the transactions resulted in price changes of three ticks or more.
5. The distribution of positive and negative price changes was approximately symmetric.

Consider next the number of transactions in a 5-minute time interval. Denote the series by x_t. That is, x_1 is the number of IBM transactions from 9:30 am to 9:35 am on November 1, 1990 Eastern time, x_2 is the number of transactions from 9:35 am to 9:40 am, and so on. The time gaps between trading days are ignored. Figure 5.1a shows the time plot of x_t, and Figure 5.1b the sample ACF of x_t for lags 1 to 260. Of particular interest is the cyclical pattern of the ACF with a periodicity of 78, which is the number of 5-minute intervals in a trading day. The number of transactions thus exhibits a daily pattern. To further illustrate the daily trading pattern, Figure 5.2 shows the average number of transactions within 5-minute time intervals over the 63 days. There are 78 such averages. The plot exhibits a "smiling" or U shape, indicating heavier trading at the opening and closing of the market and thinner trading during the lunch hours.

Table 5.1. Frequencies of Price Change in Multiples of Tick Size for IBM Stock from November 1, 1990 to January 31, 1991

Number (tick)	≤ -3	-2	-1	0	1	2	≥ 3
Percentage	0.66	1.33	14.53	67.06	14.53	1.27	0.63

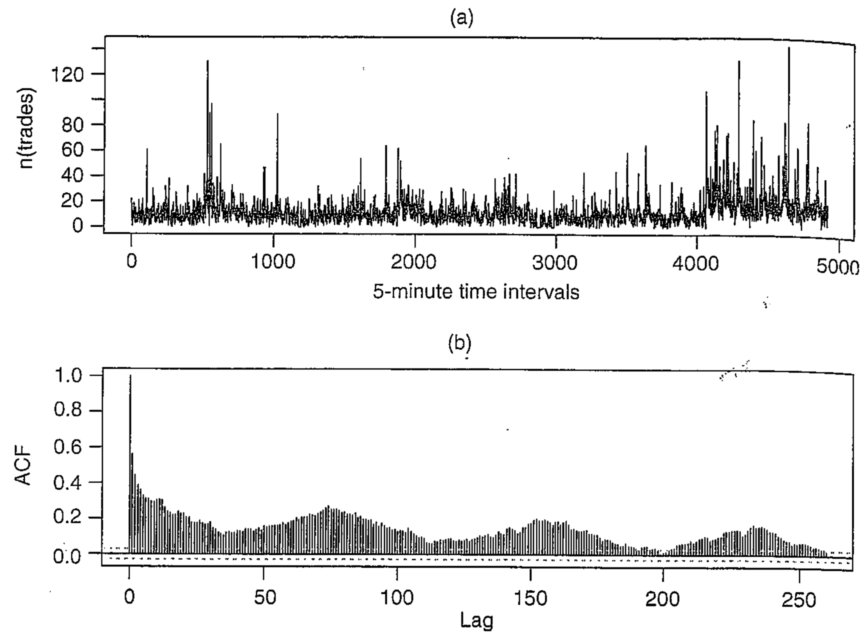

Figure 5.1. IBM intraday transactions data from 11/01/90 to 1/31/91: (a) the number of transactions in 5-minute time intervals and (b) the sample ACF of the series in part(a).

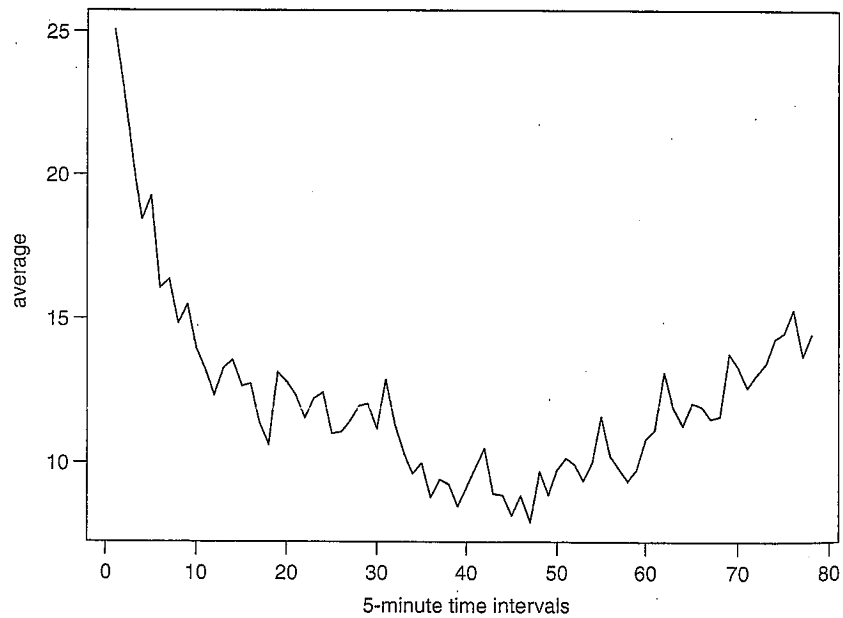

Figure 5.2. Time plot of the average number of transactions in 5-minute time intervals. There are 78 observations, averaging over the 63 trading days from 11/01/90 to 1/31/91 for IBM stock.

Since we focus on transactions that occurred during normal trading hours of a trading day, there are 59,838 time intervals in the data. These intervals are called the intraday *durations* between trades. For IBM stock, there were 6531 zero time intervals. That is, during the normal trading hours of the 63 trading days from November 1, 1990 to January 31, 1991, multiple transactions in a second occurred 6531 times, which is about 10.91%. Among these multiple transactions, 1002 of them had different prices, which is about 1.67% of the total number of intraday transactions. Therefore, multiple transactions (i.e., zero durations) may become an issue in statistical modeling of the time durations between trades.

Table 5.2 provides a two-way classification of price movements. Here price movements are classified into "up," "unchanged," and "down." We denote them by "+," "0," and "−," respectively. The table shows the price movements between two consecutive trades (i.e., from the $(i-1)$th to the ith transaction) in the sample. From the table, trade-by-trade data show that:

1. Consecutive price increases or decreases are relatively rare, which are about 441/59837 = 0.74% and 410/59837 = 0.69%, respectively.
2. There is a slight edge to move from "up" to "unchanged" rather than to "down"; see row 1 of the table.
3. There is a high tendency for the price to remain "unchanged."
4. The probabilities of moving from "down" to "up" or "unchanged" are about the same; see row 3.

The first observation mentioned before is a clear demonstration of bid–ask bounce, showing *price reversals* in intraday transactions data. To confirm this phenomenon, we consider a directional series D_i for price movements, where D_i assumes the value +1, 0, and −1 for up, unchanged, and down price movement, respectively, for the ith transaction. The ACF of $\{D_i\}$ has a single spike at lag 1 with value −0.389, which is highly significant for a sample size of 59,837 and confirms the price reversal in consecutive trades.

As a second illustration, we consider the transactions data of IBM stock in December 1999 obtained from the TAQ database. The normal trading hours are

Table 5.2. Two-Way Classification of Price Movements in Consecutive Intraday Trades for IBM Stock[a]

	ith Trade			
$(i-1)$th Trade	+	0	−	Margin
+	441	5498	3948	9887
0	4867	29779	5473	40119
−	4580	4841	410	9831
Margin	9888	40118	9831	59837

[a]The price movements are classified into "up," "unchanged," and "down." The data span is from 11/01/90 to 1/31/91.

from 9:30 am to 4:00 pm Eastern time, except for December 31 when the market closed at 1:00 pm. Comparing with the 1990–1991 data, two important changes have occurred. First, the number of intraday tradings has increased sixfold. There were 134,120 intraday tradings in December 1999 alone. The increased trading intensity also increased the chance of multiple transactions within a second. The percentage of trades with zero time duration doubled to 22.98%. At the extreme, there were 42 transactions within a given second that happened twice on December 3, 1999. Second, the tick size of price movement was $1/16 = $0.0625 instead of $1/8. The change in tick size should reduce the bid–ask spread. Figure 5.3 shows the daily number of transactions in the new sample. Figure 5.4a shows the time plot of time durations between trades, measured in seconds, and Figure 5.4b is the time plot of price changes in consecutive intraday trades, measured in multiples of the tick size of $1/16. As expected, Figures 5.3 and 5.4a show clearly the inverse relationship between the daily number of transactions and the time interval between trades. Figure 5.4b shows two unusual price movements for IBM stock on December 3, 1999. They were a drop of 63 ticks followed by an immediate jump of 64 ticks and a drop of 68 ticks followed immediately by a jump of 68 ticks. Unusual price movements like these occurred infrequently in intraday transactions.

Focusing on trades recorded within regular trading hours, we have 61,149 trades out of 133,475 with no price change. This is about 45.8% and substantially lower than that between November 1990 and January 1991. It seems that reducing the tick size increased the chance of a price change. Table 5.3 gives the percentages of trades associated with a price change. The price movements remain approximately symmetric with respect to zero. Large price movements in intraday tradings are still relatively rare.

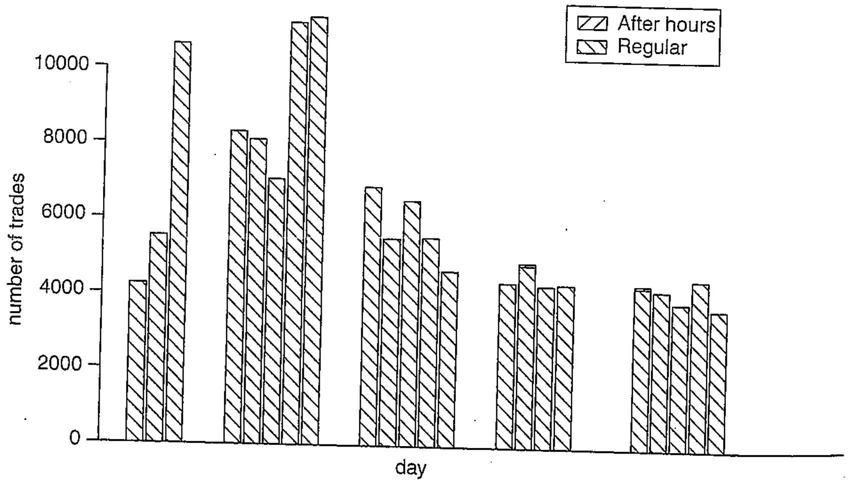

Figure 5.3. IBM transactions data for December 1999. The plot shows the number of transactions in each trading day with the after-hours portion denoting the number of trades with time stamp after 4:00 pm.

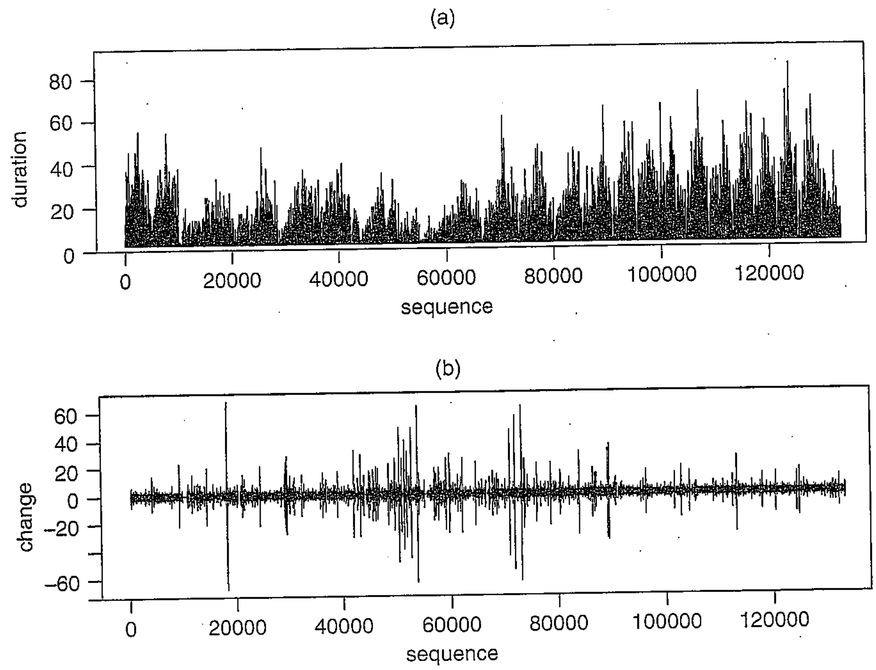

Figure 5.4. IBM transactions data for December 1999. (a) The time plot of time durations between trades. (b) The time plot of price changes in consecutive trades measured in multiples of the tick size of \$1/16. Only data during normal trading hours are included.

Table 5.3. Percentages of Intraday Transactions Associated with a Price Change for IBM Stock Traded in December 1999[a]

Size	1	2	3	4	5	6	7	>7
				Upward Movements				
Percentage	18.03	5.80	1.79	0.66	0.25	0.15	0.09	0.32
				Downward Movements				
Percentage	18.24	5.57	1.79	0.71	0.24	0.17	0.10	0.31

[a]The percentage of transactions without price change is 45.8% and the total number of transactions recorded within regular trading hours is 133,475. The size is measured in multiples of tick size \$1/16.

Remark. The recordkeeping of high-frequency data is often not as good as that of observations taken at lower frequencies. Data cleaning becomes a necessity in high-frequency data analysis. For transactions data, missing observations may happen in many ways, and the accuracy of the exact transaction time might be questionable for some trades. For example, recorded trading times may be beyond 4:00 pm Eastern time even before the opening of after-hours tradings. How to handle these observations deserves a careful study. A proper method of data cleaning requires a deep understanding of the way in which the market operates. As such, it is important to specify clearly and precisely the methods used in data cleaning. These methods must be taken into consideration in making inference. □

Again, let t_i be the calendar time, measured in seconds from midnight, when the ith transaction took place. Let P_{t_i} be the transaction price. The price change from the $(i-1)$th to the ith trade is $y_i \equiv \Delta P_{t_i} = P_{t_i} - P_{t_{i-1}}$ and the time duration is $\Delta t_i = t_i - t_{i-1}$. Here it is understood that the subscript i in Δt_i and y_i denotes the time sequence of transactions, not the calendar time. In what follows, we consider models for y_i and Δt_i both individually and jointly.

5.4 MODELS FOR PRICE CHANGES

The discreteness and concentration on "no change" make it difficult to model the intraday price changes. Campbell, Lo, and MacKinlay (1997) discuss several econometric models that have been proposed in the literature. Here we mention two models that have the advantage of employing explanatory variables to study the intraday price movements. The first model is the ordered probit model used by Hauseman, Lo, and MacKinlay (1992) to study the price movements in transactions data. The second model has been considered recently by McCulloch and Tsay (2000) and is a simplified version of the model proposed by Rydberg and Shephard (2003); see also Ghysels (2000).

5.4.1 Ordered Probit Model

Let y_i^* be the unobservable price change of the asset under study (i.e., $y_i^* = P_{t_i}^* - P_{t_{i-1}}^*$), where P_t^* is the *virtual* price of the asset at time t. The ordered probit model assumes that y_i^* is a continuous random variable and follows the model

$$y_i^* = \boldsymbol{x}_i \boldsymbol{\beta} + \epsilon_i, \tag{5.15}$$

where $\boldsymbol{x}_i$ is a p-dimensional row vector of explanatory variables available at time t_{i-1}, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, $E(\epsilon_i|\boldsymbol{x}_i) = 0$, $\text{Var}(\epsilon_i|\boldsymbol{x}_i) = \sigma_i^2$, and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. The conditional variance σ_i^2 is assumed to be a positive function of the explanatory variable $\boldsymbol{w}_i$—that is,

$$\sigma_i^2 = g(\boldsymbol{w}_i), \tag{5.16}$$

where $g(.)$ is a positive function. For financial transactions data, $\boldsymbol{w}_i$ may contain the time interval $t_i - t_{i-1}$ and some conditional heteroscedastic variables. Typically, one also assumes that the conditional distribution of ϵ_i given $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ is Gaussian.

Suppose that the observed price change y_i may assume k possible values. In theory, k can be infinity, but countable. In practice, k is finite and may involve combining several categories into a single value. For example, we have $k = 7$ in Table 5.1, where the first value "−3 ticks" means that the price change is −3 ticks