2 | Exploratory Data Analysis

Graphs are useful tools in financial data analysis. Besides the time series plots, we can visualize the distribution of returns by examining either the histogram or empirical density function of the data.

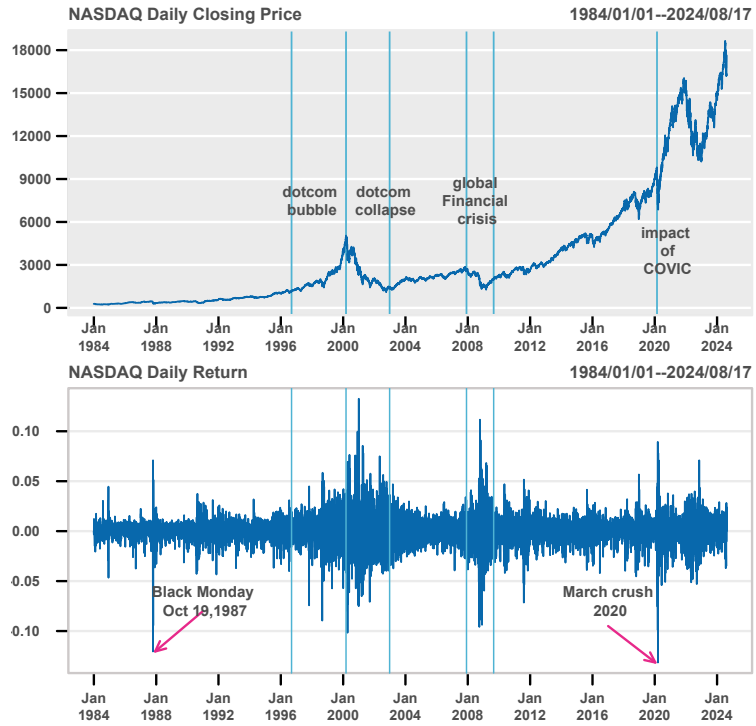


Figure 2.1: Daily adjusted closing prices and daily log returns of NASDAQ composite, starting from Jan 1, 1984 to Aug 17, 2024.

The top panel of Figure 2.1 is the time series plot of the daily adjusted closing indices of NASDAQ from Jan 1,1984 to Aug 17, 2024. We observe the dot-com bubble of 1997-2000 following by the dot-

com collapse of 2000-2002, and the 2007-2009 US financial crisis which became a global one. The NASDAQ index dropped more than 20% in the 2-month period from Jan 6 to Mar 6, 2009. The volatility clusterings of these periods are apparent in the daily return plot at the bottom. Also shown in the daily return plot is the stock market crush of Oct 19, 1987.

The initial reaction of the impact of COVID on financial markets was panic. The NASDAQ index lost 30.12% of its value between Feb 19 and March 23 of 2020. But the market abruptly surged to a new record and continued to sent prices to record highs.

Figure 2.2 shows the weekly and monthly log returns of NASDAQ for the same period. Although the profiles of the three plots are similar, the monthly return curve is a "smoothed" version of, for example, the daily return time series which exhibits higher volatile fluctuations.

Distributional Properties of Return Data

Each time series will be modeled as a sequence $Y_1, Y_2, \ldots$ of random variables each with a CDF, cumulative distribution function, F. F will vary between series but is assumed to be the same within each series implying the series is stationary. We will formally define stationarity later. The NASDAQ index price is nonstationary in the sense that the index movements over different time are different. For example, there was a slow and steady increase during 1982-1994, but the increase has become more sharply starting 2010. In

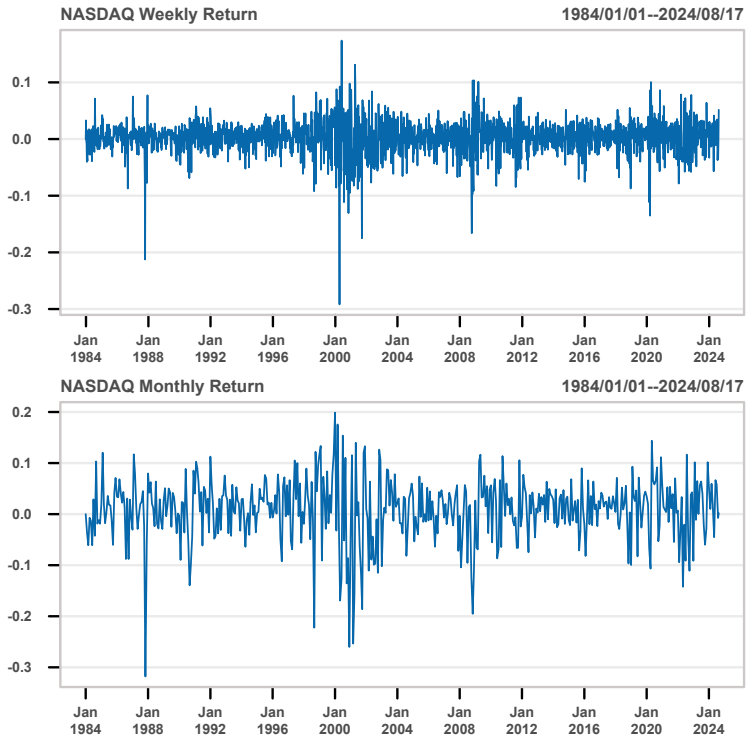


Figure 2.2: Weekly log returns and monthly log returns of NASDAQ composite, starting from Jan 1, 1984 to Aug 17, 2024.

contrast to the prices, the returns oscillate around a in constant level near 0. We will look into the marginal distribution of return series.

Marginal Distribution

By the marginal distribution of a stationary time series, we mean the distribution of Y_t given no knowledge of the other observations. When modeling a marginal distribution also called unconditional distribution, we disregard dependencies of the time series.

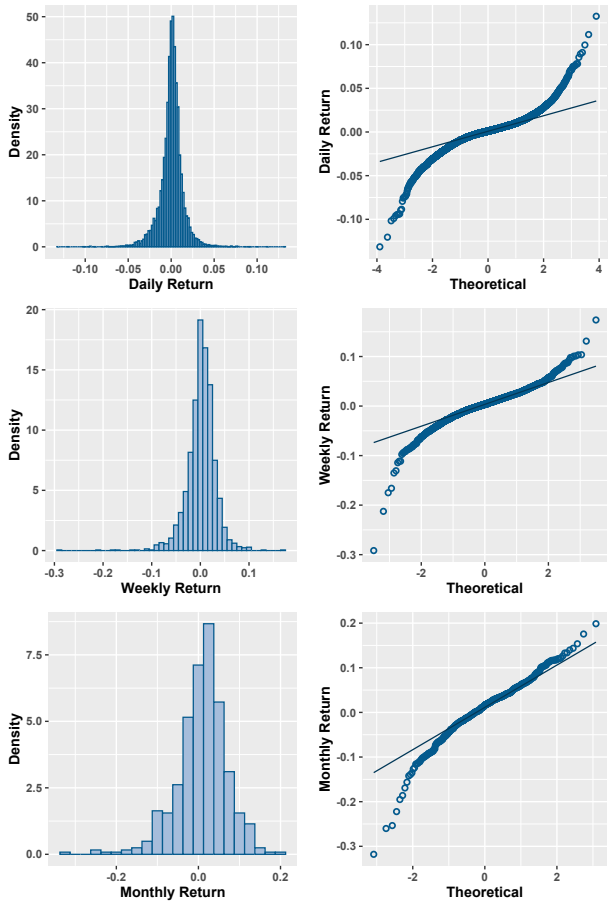


Figure 2.3: Histograms and normal probability plots of the daily, weekly and monthly log returns on the NASDAQ index form Jan 1, 1984 to Aug 17, 2024

Histogram Assume that the marginal F has a PDF, probability density function, f. The simplest estimator of PDF is histogram. The histogram is a fairly crude density estimator but it serves well for a preliminary study due to its simplicity.

The definition of density function $f(y)$ when exists is ,

$$f(y) := \frac{\partial}{\partial y} F(y) = \lim_{b \to 0} \frac{F(y+b) - F(y-b)}{2b}$$

Histogram estimation is to divide the line into a set of J equal-sized bins with bin width $2b$ and estimates the density with a given bin:

$$\hat{f}_H(y) = \frac{\#Y_i \in (a_j, a_{j+1}]}{2bn}, \qquad y \in (a_j, a_{j+1}]. \tag{2.1}$$

where $(a_j, a_{j+1}]$ defines the boundaries of the jth bin.

The appearance of histogram is sensitive to the bin-width. There are several automatic bin/binwidth selection available. R's `hist()` offers three such methods, those of Sturges (1926), Scott(1979) and Freedman & Diaconis (1981) with `Stuges` as the default.

		NASDAG Return Example		
Method	Number of Bins	Daily $n = 10,240$	Weekly $n = 2,121$	Monthly $n = 489$
Sturges	$\log_2 n + 1$	15	13	10
Scott	$\frac{\text{range} \cdot n^{1/3}}{3.5s}$	119	59	19
FD	$\frac{\text{range} \cdot n^{1/3}}{2 \cdot \text{IQR}}$	240	101	32

Among the three methods, Sturges's bin-width is most sensitive to the outliners, while the bin-width of Freedman and Diaconis is most robust to the outliers.

The return data are usually long-tailed, outliers with large sample size are expected. The default Sturges method will not be suitable.

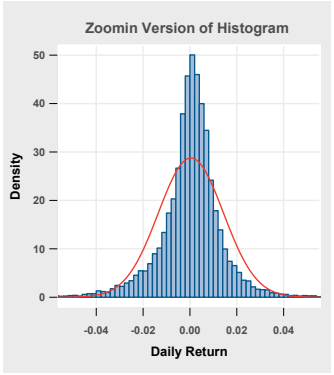
The histograms in Figure 2.3 are done by using Scott's formula.

```
hist(x, breaks = "scott", freq = F ) ##  freq = T gives counts
```

When the data set is large such as the daily returns of our example, the histogram is difficult to see due to the outliers that cause the x-axis to expand. We can zooms in on the higher-probability region to exam our data.

```
hist(x, breaks = "scott", freq = F, xlim = c(-0.055, 0.055 ))
zs = seq(-0.055, 0.055, 0.001)
lines(zs, dnorm(zs, mean(x), sd(x)), col = "red")
```

Often, we would compare the distribution of our data with the normal distribution by superimposing a normal curve as shown here.



As expected, the distribution of daily returns is deviated from a normal distribution, a feature also shown clearly in the normal probability plot in Figure 2.3.

Kernel Density Estimation The histogram is a simple but fairly crude density estimator. A typical histogram looks more like a big city skyline than a density function. A simple better alternative is to $\hat{f}_H(y)$ of (2.1) is to estimate $f(y)$ at each point y (instead of each interval) as

$$\hat{f}(y) = \frac{\#Y_i \in (y-b, y+b]}{2bn} \tag{2.2}$$

This is a kernel estimator defined next in equation (2.3) with a kernel function K being a uniform density on $[-1, 1)$.

A kernel function K is a probability density function that is symmetric about 0. The Gaussian kernel, *i.e.*, the standard normal density function is a common choice for K and works better than the uniform kernel. The kernel density estimator based on $Y_1, \ldots, Y_n$ is

$$\hat{f}(y) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{b}\right), \tag{2.3}$$

where b is the bandwidth or smoothing parameter.

Eg 2.1. The histogram shown here is a data set from Simonoff (1996). This data set are the 3-month CD rates for 69 Long Island banks given in the August 23, 1989, issue of *Newsday*.

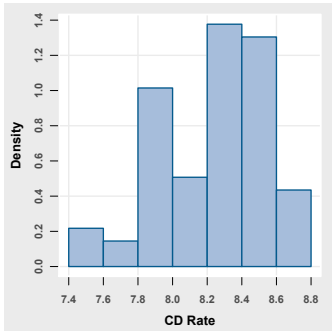


Figure 2.4 presents uniform and Gaussian kernel estimate for these data. Both estimates are with bandwidth $b = .08$. The uniform kernel estimate is more local in nature comparing to the histogram but is hardly a reasonable estimate of a smooth density. The problem is that the uniform density is discontinuous, so is the kernel density estimate based on it.

The Gaussian kernel estimate is appealingly smooth, and a trimodal form, with modes at 7.5%, 8.0%, and 8.5%, is apparent. The curves along the bottom of the plot illustrate the additive form in (2.3) the

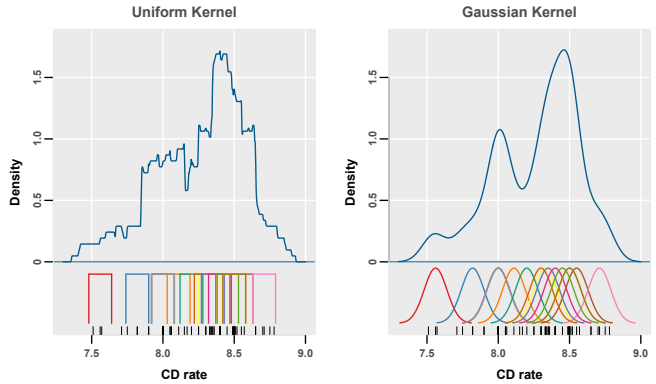


Figure 2.4: Kernel density estimates of the CD rate data using Gaussian kernel.

density estimate at any point is an average of the Gaussian densities centered at each observation.

Choosing b requires one to make a tradeoff between bias and variance. Appropriate values of b depend on both the sample size n and the true density which is unknown. Luckily, as a result of automatic selection research, modern statistical softwares such as R can select the bandwidth automatically. Density estimates computed using automatic bandwidth selectors should be checked visually and adjusted if necessary.

The R function `density()` computes and plot kernel density estimates with the Gaussian kernel as the default kernel. The tuning parameter `adjust` in the R `density()` function is the multiplier of the default bandwidth, i.e., $b =$ `adjust` $\times$ `bw`.

```
den = density(x) ## default bw = "nrd0", adjust = 1
plot(den, main = "", xlab = "CD Rate")
```


Compare to the normal curve The red blue curve in Figure 2.5 is the Gaussian kernel density estimate. Also shown in the figure are normal density curves whose mean and standard deviation are estmated by (1) the sample mean and standard deviation of the data, `mean()` and `sd()`; and (2) the median and the median absolute deviation of the data, `median()` and `mad()` . The second method is the robust estimation of the location and scale parameter. Though the robust normal curve is closer to the kernel density estimate in the center region, but deviates at the tails. This coincides with the vast empirical evidence that the returns have thick tailed distributions and fit better with t- distributions if the data were modelled parametrically.

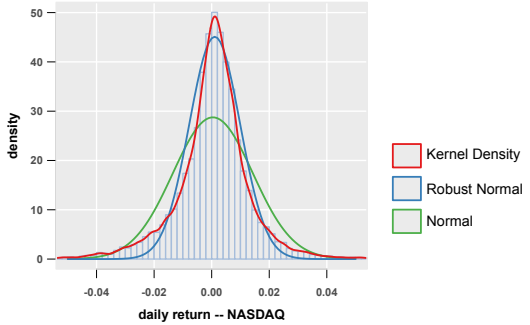


Figure 2.5: The kernel density estimates of the NASDAQ daily returns compared with normal densities.

Sample Cumulative Distribution Function and Quantiles

The sample CDF Given a random sample $Y_1 \ldots, Y_n$ from a probability distribution with CDF F. The sample or empirical CDF $F_n(y)$ is

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i \le y\},$$

where $I\{\cdot\}$ is the indicator function so that $I\{Y_i \le y\}$ is 1 if $Y_i \le y$ and is 0 otherwise.

The sample quantiles The order statistics $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$. The q sample quantile, $0 < q < 1$ is $Y_{[qn]}$ where $[qn]$ is qn rounded to an integer. The following result is the central limit theorem for sample quantiles.

Result 2.1. *Let $Y_1, \ldots, Y_n$ be an i.i.d. sample with a CDF F. Suppose that F has a density f that is continuous and positive at $F^{-1}(q)$, $0 < q < 1$. Then for large n, the qth sample quantile is approximately normally distributed with mean equal to the population quantile $F^{-1}(q)$ and variance equal to*

$$\frac{q(1-q)}{n[f\{F^{-1}(q)\}]^2}. \tag{2.4}$$

In practice, f and $F^{-1}(q)$ are unknown. If we would like to construct a confidence interval for a population quantile, however, f can be estimated by the kernel density estimator and $F^{-1}(q)$ can be estimated by the qth sample quantile. The R function for the sample quantile is `quantile()`. E.g. the first and third quartiles of the NASDAQ daily returns.

```
quantile(x, c(0.25,0.75)) ## the 1st and 3rd quartiles

##          25%          75%
## -0.004755604  0.006454180
```

Normal probability plots Normal probability plots are useful for checking the assumption of normality. If the assumption is true,

then the the sample quantile will be approximately equal to the normal population quantile $\mu + \sigma\Phi^{-1}(q)$, where Φ is the CDF of $N(0,1)$. One version of the normal probability plot is a plot of Y_i versus $\Phi^{-1}\{(i-0.5)/n\}$. These are the $(i-0.5)/n$ sample and population quantiles respectively. If the data are from normal distribution, the plot will be linear. Any nonlinear pattern of a normal probability plot indicates a deviation from normal distribution. The R commands for normal probability plots:

```
qqnorm(x) ## normal probability plot
qqline(x) ## reference line
```

Quantile-quantile plots Normal probability plots are special cases of quantile-quantile plots, also known as QQ plots. A QQ plot is simply a plot of sample quantiles against quantiles of a second sample or theoretical distribution.

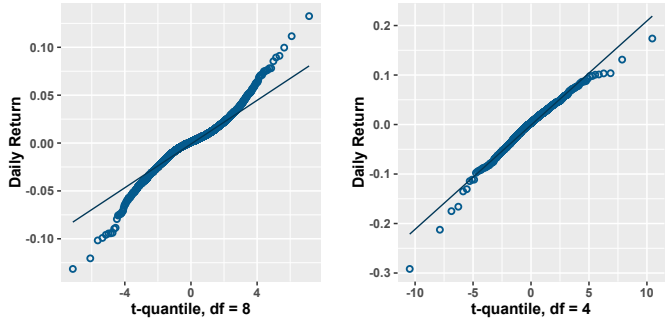


Figure 2.6: The t probability plots of the daily returns on the NASDAQ index from Jan 1, 1982 to Jan 19, 2021.

The normal probability of NASDAQ's daily returns in Figure 2.3 indicates the distribution is heavy tailed. Figure 2.6 presents the Q-Q probability of he NASDAQ daily returns and t-quantiles with de-

grees of freedom 4, and 8. The t distribution with 4 degrees of freedom fits better from the plots.

Normality tests It is sometimes difficult to judge from just a plot whether or not deviation from normality is systemic, or due to sampling variation. Statistical tests of normality test the null hypothesis that the sample comes from a normal distribution.

The Shapiro-Wilk test is based on the same concept as the normal probability plot: The sample order statistics and the expected normal order statistics are compared. Under normality the correlation between sample order statistics and the expected normal order statistics should be close to 1. The null hypothesis of normality is rejected if the correlation coefficient is too small.

There are many tests for normality comparing different features of the sample and normal distribution. The Jarque-Bera test uses the sample skewness and kurtosis coefficients. Other tests of normality in common use are the Anderson-Darling, Cramér-von Mises, and Kolmogorov-Smirnov tests. These tests compare the sample CDF to the normal CDF with mean equal to the sample mean and variance equal to the sample variance. The Kolmogorov-Smirnov test statistic is the maximum absolute difference between these two functions, while the Anderson-Darling and Cramér-von Mises tests are based on a weighted integral of the squared difference.

A recent comparison found that the Shapiro-Wilk test (probably the simplest) is as powerful as several others for both short-tailed and long-tailed symmetric alternatives, and is the most powerful test for

asymmetric alternatives.

the Shapiro-Wilk test can be implemented using the `shapiro.test()` function.

```
rt.week = weeklyReturn(Ad(IXIC), type = "log", leading = F)[-1]
shapiro.test(as.vector(rt.week))

##
##  Shapiro-Wilk normality test
##
## data:  as.vector(rt.week)
## W = 0.92918, p-value < 2.2e-16

rt.month = monthlyReturn(Ad(IXIC), type = "log", leading = F)[-1]
shapiro.test(as.vector(rt.month))

##
##  Shapiro-Wilk normality test
##
## data:  as.vector(rt.month)
## W = 0.95539, p-value = 5.551e-11

rt.quarter = quarterlyReturn(Ad(IXIC), type = "log", leading = F)[-1]
shapiro.test(as.vector(rt.quarter))

##
##  Shapiro-Wilk normality test
##
## data:  as.vector(rt.quarter)
## W = 0.94328, p-value = 4.115e-06
```

The calculation of Shapiro-Wilk test requires the sample size to be between 3 and 5000. Also, `shapiro.test()` does not take data of xts class. The first value of returns is removed because setting `leading = F` in quantmod's `weeklyReturn()` gives the first value `NA`.

When the sample size is large, it's important to look at normal plots to see if the deviation from normality is big enough to really matter. For financial time series, non-normality in tails is often large enough to be of significant importance.

Stylized Facts of Financial Time Series

Stock return, exchange rate, interest rate and other financial time series have stylized facts that are different from other time series. A good candidate for modelling financial time series should represent the properties of the stochastic processes and be able to replicate these stylized facts appropriately. We summarize these stylized features below.

Stationarity The prices of an asset recorded over times are often not stationary due to, for example, the steady expansion of economy, the increase of productivity resulting from technology innovation, and economic recessions or financial crisis. However their returns, typically fluctuates around a constant level, suggesting a constant mean over time. See Figures 2.1 and 2.2.

Heavy tails The probability distribution of return rt often exhibits heavier tails than those of a normal distribution. Figures 2.3, 2.5 and 2.6 provide the normal plots and density curves for graphical checking of normality. The Shapiro -Wilk test also reject the null of normality in returns. Nevertheless asset return is assumed typically to have at least two finite moments, although it is debatable how many moments actually exist for a given asset.

The quantile-quantile or Q-Q plots in Figure 2.6 suggests that a degrees-of-freedom-4 t distribution is a good fit to the NASDAQ daily return data. The t-distribution with 4 degrees of freedom is an example of heavy tailed distribution having finite second moments.

Asymmetry The distribution of return is often negatively skewed, as seen in the weekly and monthly returns of the NASDAQ, reflecting the fact that the downturns of financial markets are often much steeper than the recoveries. Investors tend to react more strongly to negative news than to positive news.

Volatility clustering Large price changes occur in clusters. See the return plots in Figures 2.1 and 2.2. Indeed, large price changes tend to be followed by large price changes, and periods of tranquility alternate with periods of high volatility. The volatility clusterings are apparent during the dotcom bubble and collapse, the global financial crisis 2008-2010.

Aggregational Gaussianity When the time horizon increases, the central limit law sets in and the distribution of the returns over a long time-horizon, such as a month or a quarter, tends toward a normal distribution. See Figures 2.1 and 2.2.