Statistical Concepts and Market Returns

By Dias Kulzhanov

Construction of a Frequency Distribution

- **1** Sort the data in ascending order.
- **2** Calculate the range of the data, defined as Range = Maximum value − Minimum value.
- **3** Decide on the number of intervals in the frequency distribution, *k*.
- **4** Determine interval width as Range/*k*.
- **5** Determine the intervals by successively adding the interval width to the minimum value, to determine the ending points of intervals, stopping after reaching an interval that includes the maximum value.
- **6** Count the number of observations falling in each interval.
- **7** Construct a table of the intervals listed from smallest to largest that shows the number of observations falling in each interval

Histogram/Frequency polygon

- A **histogram** is a bar chart of data that have been grouped into a frequency distribution
- A **frequency polygon** is a graph of frequency distributions obtained by drawing straight lines joining successive points representing the class frequencies.

Measures of central tendency(mean)

- **Population Mean Formula.** The **population mean**, μ, is the arithmetic mean value of a population. For a finite population, the population mean is

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (2)$$

- **Sample Mean Formula.** The **sample mean** or average, $\bar{X}$ (read "X-bar"), is the arithmetic mean value of a sample:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (3)$$

Weighted and Harmonic mean

- A portfolio's return is a weighted mean return computed from the returns on the individual assets, where the weight applied to each asset's return is the fraction of the portfolio invested in that asset.

- **Weighted Mean Formula.** The **weighted mean** $\bar{X}_w$ (read "X-bar sub-w"), for a set of observations $X_1, X_2, \ldots, X_n$ with corresponding weights of $w_1, w_2, \ldots, w_n$ is computed as

$$\bar{X}_w = \sum_{i=1}^{n} w_i X_i \qquad (4)$$

- **Harmonic Mean Formula.** The harmonic mean of a set of observations $X_1, X_2, \ldots, X_n$ is

$$\bar{X}_H = n \Big/ \sum_{i=1}^{n} (1/X_i) \qquad (7)$$

with $X_i > 0$ for $i = 1,2,\ldots,n$

Geometric mean

- The geometric mean is especially important in reporting compound growth rates for time series data

■ **Geometric Mean Formula.** The **geometric mean**, G, of a set of observations $X_1, X_2, \ldots, X_n$ is

$$G = \sqrt[n]{X_1 X_2 X_3 \ldots X_n} \quad \text{with } X_i \geq 0 \text{ for } i = 1,2,\ldots,n. \tag{5}$$

■ **Geometric Mean Return Formula.** Given a time series of holding period returns R_t, $t = 1, 2, \ldots, T$, the geometric mean return over the time period spanned by the returns R_1 through R_T is

$$R_G = \left[\prod_{t=1}^{T}(1 + R_t)\right]^{\frac{1}{T}} - 1 \tag{6}$$

Median, quartiles, quintiles, deciles, and percentiles

- Quartiles divide the distribution into quarters.
- Quintiles into fifths.
- Deciles into tenths
- Percentiles into hundredths.

$$L_y = (n+1)\frac{y}{100} \qquad (8)$$

where y is the percentage point at which we are dividing the distribution and L_y is the location (L) of the percentile (P_y) in the array sorted in ascending order. The value of L_y may or may not be a whole number. In general, as the sample size increases, the percentile location calculation becomes more accurate; in small samples it may be quite approximate.

Population variance/standard deviation

- **Mean Absolute Deviation Formula.** The **mean absolute deviation** (MAD) for a sample is

$$\text{MAD} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n} \qquad (10)$$

- **Population Variance Formula.** The population variance is

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \qquad (11)$$

- **Population Standard Deviation Formula.** The **population standard deviation**, defined as the positive square root of the population variance, is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \qquad (12)$$

Sample variance/*sample standard deviation*

- **Sample Variance Formula.** The **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \qquad (13)$$

- **Sample Standard Deviation Formula.** The **sample standard deviation**, s, is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad (14)$$

Semivariance

- The semivariance is the average squared deviation below the mean.
- **Target semivariance** is the average squared deviation below a target level.

i. Calculate the sample mean.

ii. Identify the observations that are smaller than or equal to the mean (discarding observations greater than the mean).

iii. Compute the sum of the squared negative deviations from the mean (using the observations that are smaller than or equal to the mean).

iv. Divide the sum of the squared negative deviations from Step iii by the *total* sample size minus 1: $n - 1$. A formula for semivariance approximating the unbiased estimator is

$$\sum_{\text{for all } X_i \le \bar{X}} \left(X_i - \bar{X}\right)^2 \Big/ (n-1)$$

Chebyshev's inequality

- **Definition of Chebyshev's Inequality.** According to Chebyshev's inequality, for any distribution with finite variance, the proportion of the observations within k standard deviations of the arithmetic mean is at least $1 - 1/k^2$ for all $k > 1$.

Table 23 Proportions from Chebyshev's Inequality

k	Interval around the Sample Mean	Proportion (%)
1.25	$\bar{X} \pm 1.25s$	36
1.50	$\bar{X} \pm 1.50s$	56
2.00	$\bar{X} \pm 2s$	75
2.50	$\bar{X} \pm 2.50s$	84
3.00	$\bar{X} \pm 3s$	89
4.00	$\bar{X} \pm 4s$	94

Note: Standard deviation is denoted as s.

Coefficient of variation

- The coefficient of variation, CV, is the ratio of the standard deviation of a set of observations to their mean value.

■ **Coefficient of Variation Formula.** The **coefficient of variation**, CV, is the ratio of the standard deviation of a set of observations to their mean value:[36]

$$\mathrm{CV} = s/\bar{X} \qquad (15)$$

Sharpe ratio

■ **Sharpe Ratio Formula.** The **Sharpe ratio** for a portfolio p, based on historical returns, is defined as

$$S_h = \frac{\bar{R}_p - \bar{R}_F}{s_p} \qquad (16)$$

where $\bar{R}_p$ is the mean return to the portfolio, $\bar{R}_F$ is the mean return to a risk-free asset, and s_p is the standard deviation of return on the portfolio.[37]

Skewness

- Skew describes the degree to which a distribution is not symmetric about its mean.
- A return distribution with positive skewness has frequent small losses and a few extreme gains. A return distribution with negative skewness has frequent small gains and a few extreme losses.

■ **Sample Skewness Formula. Sample skewness** (also called sample relative skewness), S_K, is

$$S_K = \left[\frac{n}{(n-1)(n-2)}\right] \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3} \qquad (17)$$

cubed deviation, $S_K \approx \left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$. As a frame of reference, for a sample size of 100 or larger taken from a normal distribution, a skewness coefficient of ±0.5 would be considered unusually large.

Kurtosis

- Kurtosis measures the peakedness of a distribution and provides information about the probability of extreme outcomes. A distribution that is more peaked than the normal distribution is called leptokurtic; a distribution that is less peaked than the normal distribution is called platykurtic.

■ **Sample Excess Kurtosis Formula.** The **sample excess kurtosis** is

$$K_E = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{s^4} \right) - \frac{3(n-1)^2}{(n-2)(n-3)} \qquad (18)$$

In Equation 18, **sample kurtosis** is the first term. Note that as n becomes large, Equation 18 approximately equals $\frac{n^2}{n^3} \frac{\sum (X - \bar{X})^4}{s^4} - \frac{3n^2}{n^2} = \frac{1}{n} \frac{\sum (X - \bar{X})^4}{s^4} - 3$. For a sample of 100 or larger taken from a normal distribution, a sample excess kurtosis of 1.0 or larger would be considered unusually large.