Financial Enterprise Risk Management
Source Author: Paul Sweeting (2017)

Chapter 9: Some Useful Statistics (Background Only)

Overview of This Reading

Even though the syllabus says this chapter "background only," it's very important to be able to describe these statistics and be able to perform calculations using them

These concepts in this chapter are the basic ingredients for more advanced correlation and copula topics covered later in readings like ERM-101 and ERM-125

Key topics for the exam include:

- Basic univariate statistical measures
    - Location measures (mean, median, mode)
    - Spread measures (variance, range)
    - Skew measures
    - Kurtosis
- Measures of correlation between two variables
    - Pearson's rho
    - Spearman's rho
    - Kendall's tau
    - Tail correlation

Location Measures

Location measure – gives an indication of the point around which observations are based

A **mean** is the most commonly used measure of central tendency in modeling (based on first moment of distribution)

Sample mean of a set of observations:

$$\overline{X} = \frac{1}{T} \sum_{t=1}^{T} X_t$$

Population mean (μ) is usually <u>not</u> observable, but is calculated the <u>same</u> as $\overline{X}$

$$\mu = \frac{1}{T} \sum_{t=1}^{T} X_t$$

Median – measure of the mid-point of a distribution (50th percentile)

- Useful for analyzing simulated data

Mode – the most common observation

- Discrete distributions:

    1. Count the number of each observation

    2. Mode = observation with highest count

- Continuous distributions: mode = maximum of density function = point at which the first derivative = 0

    ○ First derivative is sometimes called the "gradient"

Spread Measures

Spread measures indicate how far away an observation may fall from a location measure

- Can help establish confidence intervals

Variance is the most common spread measure (based on 2nd moment of distribution)

Population variance:

$$\sigma^2 = \frac{1}{T} \sum_{t=1}^{T} (X_t - \mu)^2$$

- Appropriate if the dataset represents all possible observations (not likely for risk management)

- Not a good estimate of the true population variance since some observations exist in the future and cannot be known

    ○ Biased downwards for finite samples (i.e. it tends to underestimate the true variance)

    ○ Bias <u>increases</u> as the sample size <u>falls</u>

Sample variance is adjusted to reduce the bias:

$$s^2 = \frac{1}{T-1} \sum_{t=1}^{T} (X_t - \overline{X})^2$$

Range = difference between the largest and smallest value in dataset

- Alternative to variance for measuring spread

- May capture information about the effect of potential extreme events

- Straightforward to calculate, but can't be used for parametric distributions or if observations are unbounded (i.e. observations can be zero to ∞)

- For unbounded distributions, one solution is to use an inter-quartile range (e.g. 75th percentile − 25th percentile)

Skew Measures

Skew – a measure of a distribution's asymmetry (based on 3rd moment)

- Skew = 0 for perfectly symmetric distributions (e.g. normal distribution)

- **Negative skew** means left tail > right tail (positive skew is the opposite)

   ○ This assumes the observations are increasing (i.e. worst values on the left, best values on the right)

- Mean and variance do not capture skew

Many risk distribution are negatively skewed

- Probability of a large loss > probability large gain

Problems with ignoring skew or assuming skew = 0:

1. Underestimates risk $\Rightarrow$ may result in lower-than-expected profits

2. May result in profitable projects being rejected if there is a desire for large profits

Population skew:

$$\omega = \frac{1}{T}\left(\frac{\sum_{t=1}^{T}(X_t - \mu)^3}{\sigma^3}\right)$$

- Will be biased if full distribution is not available (usually it's not available)

Sample skew mitigates the bias:

$$w = \left(\frac{T}{(T-1)(T-2)}\right)\left(\frac{\sum_{t=1}^{T}\left(X_t - \overline{X}\right)^3}{s^3}\right)$$

Kurtosis

Kurtosis indicates the likelihood of extreme observations relative to those that would be <u>expected</u> with the normal distribution

- In other words, it measures how fat the tails are $\Rightarrow$ higher kurtosis = fatter tails

- Based on the 4th moment of the distribution

Types of distributions based on the kurtosis value

1. **Mesokurtic:** kurtosis = 3 (true of the normal distribution)

2. **Platykurtic:** kurtosis < 3 $\Rightarrow$ distribution has thin tails relative to the normal distribution ("negative excess kurtosis")

3. **Leptokurtic:** kurtosis > 3 $\Rightarrow$ has fat tails relative to the normal distribution ("positive excess kurtosis")

   - If this is not accounted for, the probability of extreme events will be underestimated

   - Many risk distributions are leptokurtic

Memory tip: leptokurtic means the kurtosis "leptover" 3! :)

Population measure of <u>excess</u> kurtosis:

$$\kappa = \frac{1}{T} \frac{\sum_{t=1}^{T} \left(X_t - \mu \right)^4}{\sigma^4} - 3$$

- Measures the kurtosis against the normal distribution (hence subtract 3)

Sample excess kurtosis corrects for bias if the full population data isn't available:

$$k = \left(\frac{T(T+1)}{(T-1)(T-2)(T-3)} \right) \left(\frac{\sum_{t=1}^{T} \left(X_t - \overline{X} \right)^4}{s^4} \right) - \frac{3(T-1)^2}{(T-2)(T-3)}$$

Correlation Measures

Correlation is important in ERM because it measures the diversification benefits of aggregating risks

Interpretation of correlation between 2 variables

1. Strong positive correlation

   - The risk of two events occurring simultaneously is high

2. Low correlation

   - The risks can diversify one another

3. Strongly negative correlation

   - Indicates an incentive to <u>increase</u> the level of one risk taken in order to offset the second

Three Measures of Correlation

1. **Pearson's rho**: $\rho_{X,Y}$ (a.k.a. "linear correlation coefficient")

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

$$\sigma_{X,Y} = \frac{1}{T} \sum_{t=1}^{T} (X_t - \mu_X)(Y_t - \mu_Y) = \text{population covariance}$$

- Attractive, widely used, easy to calculate

- Only a valid if the data series are <u>jointly elliptical</u> (i.e. related to the multiple variate normal distribution)

- If not jointly elliptical, $\rho_{X,Y} = 0$ does not necessarily mean independence

- For sample populations, use:

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

$$s_{X,Y} = \frac{1}{T-1} \sum_{t=1}^{T} \left(X_t - \overline{X}\right)\left(Y_t - \overline{Y}\right)$$

2. **Spearman's rho:** $_s\rho$ (a.k.a. Spearman's rank correlation coefficient)

   For sample populations:

$$_s r_{X,Y} = 1 - 6 \cdot \left(\frac{\sum_{t=1}^{T}(V_t - W_t)^2}{T(T^2 - 1)} \right)$$

$$V_t, W_t = \text{rankings of } X_t \text{ and } Y_t, \text{ respectively}$$

- Equals Pearson's rho if data are uniformly distributed

- Differences from Pearson's rho:
   - Ranks can be in ascending or descending order (since rank differences are squared)
   - Spearmean's rho is independent of the statistical distribution

3. **Kendall's tau:** (τ) – compares <u>pairs</u> of data points
   - Suppose we have 2 observations: (X_1, Y_1) and (X_2, Y_2)

   - If $X_2 - X_1$ and $Y_2 - Y_1$ have the same sign, these pairs are **concordant**; else they are **discordant**

   - For T observations, the total number of possible pairs is

$$T\left(\frac{T-1}{2}\right)$$

- **Sample Kendall's tau** normalizes all concordant pairs (p_c) and discordant pairs (p_d) by the total number of pairings

$$t_{X,Y} = \frac{2(p_c - p_d)}{T(T - 1)}$$

- Spearman's rho ($_s\rho$) and Kendall's tau (τ) are related in the following way:

$$\frac{3}{2}\tau - \frac{1}{2} \leq {_s\rho} \leq \frac{1}{2} + \tau - \frac{1}{2}\tau^2 \qquad\qquad \text{if } \tau \geq 0$$

$$\frac{1}{2} + \tau + \frac{1}{2}\tau^2 \leq {_s\rho} \leq \frac{3}{2}\tau + \frac{1}{2} \qquad\qquad \text{if } \tau < 0$$

Key comparisons among the 3 correlation measures above:

- Pearson's rho is calculated directly from the data series, while Spearman's rho and Kendall's tau are rank measures
    - **Rank measure** – a statistic calculated from the position (rank) of the observations

- Pearson's rho is only valid if the data series are jointly elliptical, but rank measures are always valid since they do not depend on the distribution's shape

- Rank measures are usually combined with copulas[5]
    - Kendall's tau has simple relationships with a number of copula functions

- Limitations of the 3 correlation measures:
    - Each describes <u>only one</u> aspect of the variables' relationship

    - Copulas can describe correlation relationships more accurately than any of the above measures alone

Tail Correlation

The 3 correlation measures described above imply that X and Y always have the same relationship

In extreme situations (tail events), variables' relationships can change

Tail correlation looks at the variables' relationship only in the tail (e.g. lowest and highest 10% of observations)

Key problems: Determining where the "tail" begins is subjective and can cause instability in parameterization

---

[5] Copulas will be covered in much more detail later in readings like ERM-101 and ERM-125. We will also see how the basic correlation measures in this chapter relate to various copula methods. Copulas identify relationships among each individual variable's distribution to create a multivariate joint distribution. Among other advantages, a copula can describe how correlations change as the value of variables change—something that is very useful for risk modeling where extreme tail events can cause variables' relationships with each other to change.