SESSION 3: DATA DESCRIPTIVES

Session 3

Means and Moments

Data Descriptives: An Overview

- Data descriptives are designed to capture the features of data in metrics that can be used to summarize the data and make comparisons across time and samples.
- Broadly speaking, there are four groups of descriptive statistics:
  - *Measures of centrality* are meant to capture the central value of a variable, across a sample.
  - *Measures of dispersion* try to measure the spread of data around the central value, with greater dispersion referencing a larger spread.
  - *Measures of symmetry* capture the skew in the spread of the data around the central value
  - *Measures of extremes* measure the likelihood of extreme values (the fatness of distributional tails)

Measures of Centrality

1. Mean: Perhaps, the most common measure of centrality in a data is the average. It is computed by aggregating the numerical values across all observations and dividing by the number of observations:
2. Median: The median literally represents the mid-point in the data, where half of all of the observations are higher than it and half are lower.
3. Mode: When the data take on different values, the mode is the value that you observe most frequently across observations.

Comparing Central Tendency Measures

1. The average matters: The average is the most widely used measure of central tendency because other statistical measures are built around it; the standard deviation, for instance, is computed around the average.
2. But the median may be more meaningful: When the data is symmetric, i.e., outliers lie on both sides of the average in rough equivalence, the mean and the median will converge. When the outliers are more in one direction than in the other, the average will be skewed by the outliers, and the median may be more meaningful.
3. And the mode less so: The mode is the least used measure of central tendency, at least in finance and investing, because it is designed more for discrete or categorical data, than it is for continuous data.

Measures of Dispersion

- Range and variants: With the range, you look at the difference between the highest and lowest values for a variable, across a sample. In variants, you can look at the difference between the first and the third quartile of the data (interquartile range) or between the first and the ninth decile of the data.
- Standard deviation/Variance: With the standard deviation, you estimate the difference between each value of a variable and its mean and arrive at a measure of dispersion.
- Coefficient of variation: With the coefficient of variation, you divide the standard deviation of a data series by its mean, to provide a measure of comparison with data series of different levels.

The Range

- The range is the difference between the maximum and minimum values of a data variable in a sample.
  - Range = Maximum Value – Minimum Value
  - This range can be standardized by dividing by the average of the two values.
  - Standardized Range = (Max - Min)/ [(Max + Min)/2]
- It remains the simplest measure of dispersion for a data series, and the range requires no explicit assumptions about how the data is distributed.
- However, by using only the highest and lowest values, and ignoring the intermediate numbers, it wastes data and can give misleading measures of dispersion.
- In general, though, breaking data down into deciles or quartiles will give you useful information about dispersion.

Standard Deviation/ Variance & Coefficient of Variation

- The standard deviation is a measure of dispersion which uses all of the observations, computes the difference between each one and the mean, and summing up the squared differences:
  - Variance = $\frac{\sum_{j=1}^{j=n}(X_j - \mu)^2}{n}$ (where μ is the average across the n observations, and X_j is the value that of the j^{th} observation); The sum of the squared differences is divided by n, if your data comprises the entire population, or n-1, if it is a sample.
  - The standard deviation is the square root of the variance.
- When there is more divergence from the mean, the standard deviation will be higher, but it will be in the same units as the base data. Thus, if the base data is in dollars, the standard deviation will be in dollars, and if it is in percent, it will be in percent.
- Since standard deviations cannot be compared across two samples with different units or levels, you can compute a standardized version of the measure:
  - Coefficient of Variation = Std Deviation in Value/ Average Value

Standard Deviation and Standard Error

- The standard deviation measures the amount of variability, or dispersion, from the individual data values to the mean
- The standard error of the mean measures how far the sample mean (average) of the data is likely to be from the true population mean. It is computed as follows:
  - Standard Error = $\frac{Standard\ Deviation}{\sqrt{Number\ of\ observations\ in\ sample}}$
- As sample size increases, there will be no discernible effect on the former, but the latter will always decrease.
- When you are extrapolating from sample findings to the population, the standard errors become useful because they can be used to provide ranges for estimates. Thus, if your average is μ, and your standard error is SE, drawing on the central limit theorem, you can estimate the population mean:
  - With 67% confidence: $\mu \pm SE$
  - With 95% confidence: $\mu \pm 2 * SE$

Measures of Asymmetry

- When the data is symmetric, the deviations from the mean fall equally or roughly equally on either side of the mean.
- When the data is asymmetric, deviations on one side of the mean are much more pronounced than deviations on the other side. This deviation is measured with skewness.
  - If the deviations are more pronounced/extreme for the observations that have values higher than the average, the distribution is positively skewed.
  - If the deviations are more pronounced/extreme for the observations that have values lower than the average, the distribution is negatively skewed.
- When data is asymmetric, the average will be skewed in the same direction as the asymmetry, and in some cases the skew can be large enough to make it unrepresentative of the sample.

Boundedness, and consequences

- In some cases, data can also be bounded, where there are limits on how low a value can be (lower bound) or or high (upped bound). Those limits can come from either how the data variable is defined, or from how it is recorded.
- In others, data is unbounded, where it can take values, albeit very, very rarely, that approach infinity (plus or minus).
- When data is bounded on only one side and unbounded on the other, asymmetry almost always follows.

Measures of extreme values

- You can measure of how much, and how frequently, data takes extreme values, relative to its central value. That measure is called kurtosis.
- While variance and kurtosis are both affected by the presence (or absence) of extreme values, they measure different phenomenon.
  - You can have high variance and low kurtosis, low variance and high kurtosis or high variance and high kurtosis.
  - Distributions that have more frequent occurrences of extreme values are referred to as having fat tails or *leptokurtic*. Distributions that have less frequent occurrences of extreme values are referred to as *platykurtic*.