Reading2: Organizing, Visualizing, and Describing Data


```
key words: data types,frequency, distribution,contingency table,
visualization, skewness, kurtosis, correlation, measurement of central
tendency, location and dispersion
```


1. Organizing Data

Data Type

1. Numerical and Categorical Data

- Numerical data: counted or measured
  - Discrete: countable
  - Continuous: take any fractional value
- Categorical data: labels that can be classify a set of data into groups
  - Nominal: without order

- Ordinal: can be ranked in order respect to the specific characteristic

2. Time Series and Cross-sectional

- Time Series: observations taken periodically.时间变，种类不变
- Cross-sectional: comparable observations taken at one specific point in time.种类变，时间不变

The two types of data combined together form **panel data**.

3. Structured and Unstructured

Frequency Distribution

- Define the intervals: the range of values with upper and lower limits, all-inclusive, non-overlapping, numbers of interval
- Tally（一致,标签） the observations:observations should be assigned to their appropriate interval
- Count:
  - **absolute frequency**: actual number of observations
    - Mode frequency: the interval with greatest frequency
  - **relative frequency**: divide the absolute frequency of each return interval by the total number of observations (%)
  - **Cumulative absolute frequency** and **Cumulative relative frequency** can be calculated

Contingency Table

- 2-dimensional array, analyzing 2 variables as the same time, using nominal or ordinal data, with a finite number attributes
- **Joint frequency**: each cell shows the frequency which is related to two attributes simultaneously
- **Marginal frequency**: total frequency for the row or a column

- **Confusion Matrix**: one kind of contingency table with numbers of occurrences predicted and actually observed

2. Visualizing Data

Graph Type

- Histogram:
  - absolute frequency distribution
  - See where most of the observations are concentrated
  - a bar chart of continuous data
- Frequency polygon
- Cumulative (absolute/relative) frequency distribution chart
- Bar chart: illustrate **relative** sizes, degrees, or magnitudes
  - grouped bar chart/clustered bar chart: show 2 categories at once
  - stacked bar chart:
    - height of each bar: the cumulative frequency for a category
    - Colors within each bar: joint frequencies
- Tree map: visualizing **relative sizes** of categories with different colors or shades
- Word cloud: visualizing text, counting the uses/frequency of specific words which shows in **type size**
- Line charts: usually used for visualizing **time series** data
  - Multiple time series: when scales of time series are different, use left and right vertical axes
  - Bubble line chart: different sizes bubble represent the relative size of another variable
- Scatter plot: how *2 variables* tend to change *in relation to* each other
  - correlation coefficient: a measure of strength of a linear relationship
- Scatter plot matrix: analyze three variables
- Heat map: use color and shades to display data frequency, with data from contingency table

Choose from Visualization Types

- Relationship,Comparison,Distribution

- avoid from misrepresentation to mislead investors

3. Measures of Central Tendency

- identify the center, average to represent the typical or expected value in dataset
- **Arithmetic mean**: `Σobservation value/#of observations`, estimate the next observation, expected value of distribution
  - Eg:
    - **population mean**: given population only 1 mean
    - **sample mean**: to make inferences about the population mean
  - Properties:
    - all datasets only have one arithmetic mean
    - all data value are considered in arithmetic mean computation
    - all interval and ratio datasets have an arithmetic mean
    - The sum of *deviations* of each observation in the dataset from mean=0 `Σ(Xi-X_bar)=0`
  - **Outliers** have an influence in arithmetic mean, providing all-sided information.
    - it should be excluded from measure of central tendency
    - Instead, **trimmed mean** is used (1%= 0.5%lowest + 0.5%highest discarded) without outliers
    - **Winsorized mean**:substitute a value for <5th percentile and >95th percentile, if select 90% Winsor mean. Then calculate the revised dataset. Decrease the effects of outlier.
- **Weighted mean**: different observations have a disproportionate influence on mean.
  - $\sum_i^n W_i X_i$, where **Σ**Wi=1
  - usually used to calculate <u>portfolio return </u>, the *weight* of individual asset=market value of each asset/market value of entire portfolio
- **Median**: midpoint with sorted dataset
  - not so affected by extreme values
  - Calculation: with odd/even number of observations
- **Mode**: value occurs most frequently
  - datasets can have >=1 or no mode, Unimodal, Bimodal, Trimodal
- **Geometric Mean**: get investment returns over *multiple periods*, or used to compute *compound* growth rate.
  - Function:$G = (\prod_i X_i)^{\frac{1}{N}}$
    - the radical sign is non-negative.

    - $1 + R_G = \sqrt[n]{\prod(1 + R_t)}$, where R_t is the return of period t. Then we can get R_G as the result at final.
    - By financial calculator: [y^x][n][$\frac{1}{x}$][=]
    - Geometric Mean ≤ Arithmetic Mean.
        - the difference between these two mean ↑, when the dispersion of observation ↑
        - When all observations are equal, arithmetic=geometric mean
- **Harmonic Mean**: used to calculate *average cost of shares* purchased over time.
    - Function: $\frac{N}{\sum_i^N \frac{1}{X_i}}$
    - after we get average cost per share, if the total money purchase of shares is given, the *total amount of shares purchased* can be computed.
    - harmonic mean< geometric mean< arithmetic mean

4. Measures of Location and Dispersion

Location: Quantile + Measures of Central Tendency

- **Quantile**: value <= a stated *proportion* of data with a distribution
    - Type:
        - Quartile: 4
            - $Q3 - Q1$: inter-quartile range
        - Quintile: 5
        - Decile: 10
        - Percentile: 100
    - Formula: Position of the observation $L_y = (n + 1)\frac{y}{100}$
        - y: at a given percentile
        - n: data points in ascending order
    - Visualization: Box and whisker plot
        - Box: inter-quartile range
        - Vertical line: entire range with largest/smallest values
- **Dispersion**: variability around the central tendency.
    - in finance, central tendency is reward, dispersion is a measure of risk.

- **Range**: provide extremely useful information
  - function: $Range = X_{max} - X_{min}$
- **Mean Absolute Deviation(MAD)**: use average the absolute value of deviation from the mean
  - function: $MAD = \frac{\sum_i^n |X_i - \bar{X}|}{n}$
- **Sample Variance(s^2)**: $s^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{n-1}$
  - use $n-1$ as denominator: *unbiased estimator* of population variance
  - **Sample Standard Variance**$s = \sqrt[2]{\frac{\sum_i^n (X_i - \bar{X})^2}{n-1}}$, whose *unit* is the same as the unit of data.
    - unbiased estimator of population standard deviation σ.
- **Relative Dispersion**: the amount of variability in a distribution relative to a benchmark.
  - Quantified as **Coefficient of Variation(CV)**, whose benchmark chosen as *mean* of distribution.
    - function: $CV = \frac{S_x}{X}$
    - The lower the better.
    - In Finance, CV measures the risk per unit of expected return .
  - Direct compare the dispersion between different sets of data.

- **Downside Risk**: only include outcomes< mean(or other benchmark)实际收益低于预期的风险，下行风险
  - Name: **Target Downside Deviation/Semi-deviation**
  - Function: $s_{target} = \sqrt[2]{\frac{\sum_{all\ X_i<B}^n (X_i - B)^2}{n-1}}$, where B is target
    - Be careful that the denominator we use $n-1$

5. Skewness, kurtosis, and Correlation

- *Symmetrical*: the degree of symmetry measures if the *deviations from the mean* are positive or negative.
- Skewness and Kurtosis are critical points for **risk management**.
  - greater positive kurtosis & more negative skewness in return distribution→ **increased** risk
- **Skewness**: show the non-symmetric distribution with positive/negative skew, result from the outliers occurrence .
  - Type:

        - Positively Skewed: outliers > mean, skewed right, long upper tail.
        - Negatively Skewed: outliers < mean, skewed left, long lower tail.
    - The relationship between Mean, Median, Mode:
        - Symmetrical distribution: `mean=median=mode`
        - For positive skewed, `mode < median < mean`
        - For negative skewed, `mean < median < mode`
        - Rule(easy to remember the order):
            - The effect of skew pays more attention on **mean**, other than median and mode.
            - mean is in the same direction of the skew.
            - Besides, median always stays between mean and mode, regardless the skewness direction.
    - **Sample Skewness**:
        - Function: $Sample\ Skewness = \frac{1}{n} * \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$
            - the denominator > 0, while the sign of **numerator** depend on the skewness direction.
            - if $|SampleSkewness| > 0.5$, treated as *significant*.
- **kurtosis**:
    - Definition: the degree of a distribution is **more or less peaked** than a *normal distribution*
    - Types:
        - Lepokurtic: more peaked
            - in investment, a *greater* likelihood of a large deviation from the expected return usually received as an increase in risk.
        - Platykurtic: flatter, less peaked
        - Mesokurtic: the same kurtosis as $Ndist$
    - **Excess Kurtosis**: *more or less kurtosis part* than the normal distribution.
        - Mesokurtic EK=0;
        - Lepokurtic EK>0;
        - Platykurtic EK<0;
    - **Sample Kurtosis**:
        - function: $Sample\ Kurtosis = \frac{1}{n} * \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4}$
    - $Excess\ Kurtosis = Sample\ Kurtosis - 3$
- **Correlation**
    - **Covariance**: how two variables move together
    - For Sample Covariance, the function: $S_{X,Y} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{n-1}$

        - the value is depend on the **units** of variables. Hence, the covariance cannot show the relative strength of the relationship, but only discloses the move direction whether the same or opposite.
    - **Correlation Coefficient**
        - Function: $\rho_{XY} = \frac{S_{XY}}{S_X S_Y}$
        - Properties:
            - range: $[-1, 1]$, without units, care about the strength and movement direction of the linear relationship between 2 variables.
    - **Visualization:Scatter Plots**
        - two variable relationship
        - it can reveal non-linear relationship, which cannot be shown by ρ.
    - Correlation != Causation.
    - **Spurious Correlation**伪关系
        - the relationship between two variables caused by association with a third variable or by chance.