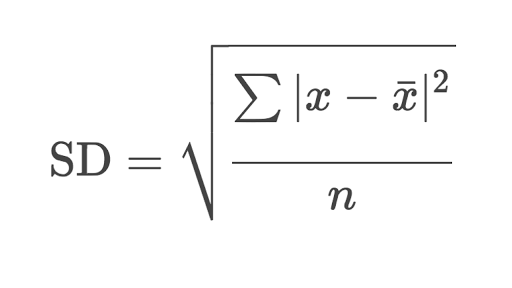
1. Anade Davis – Instructor, Data Science
   1. financial DS specialty
2. Data Source -> Data Flow
   1. Data Source -> Extract, Transform, Load -> Data Validation
   2. Data Source
      1. Structured Data:
         1. Text, Numbers
         2. 10-20% of data
      2. Unstructured Data:
         1. Emojis, Log Files (emails), Pictures, Videos
         2. 80-90% of data
      3. Data Formats
         1. Flat: text-based (csv, xml, JSON)
         2. Numerical / Quantitative / Parametric
            1. Discrete
            2. Continuous

Interval –

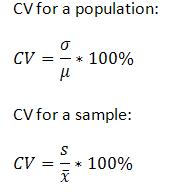
Ratio – Zero has “a meaningful zero and hence allow relative comparison of two measurements”

* + - 1. Categorical / Factors / Qualitative / Non-Parametric
         1. Ordinal – Ranked
         2. Nominal

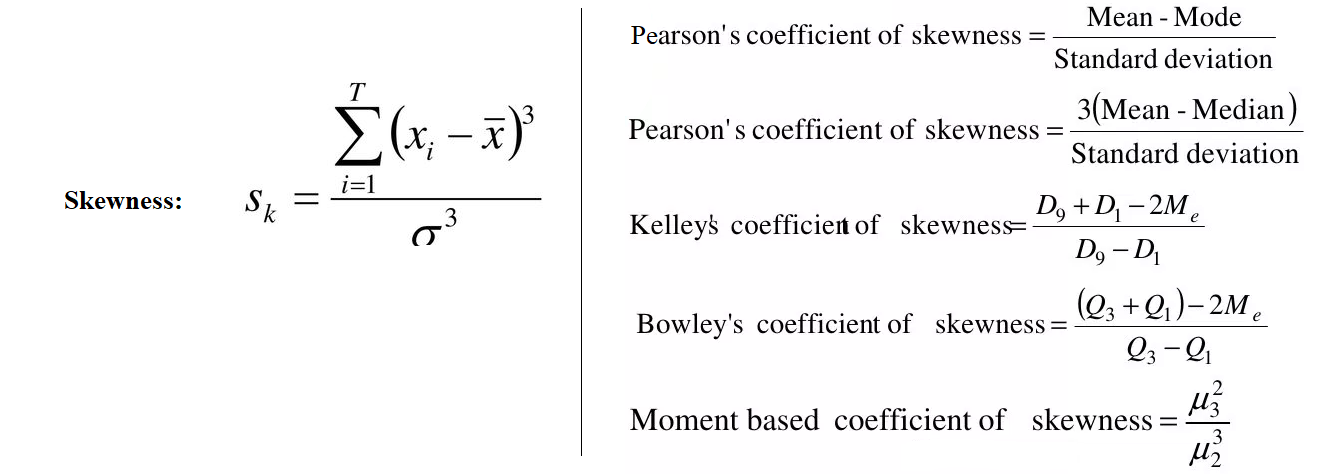
1. Distributions & Describing Them
   1. Bell Curve / Normal Distribution
      1. 68% in 1 std
      2. 95% of values in 2 std
      3. 99.7% in 3 std
   2. Std represents how one group differs from the mean value of the entire group
      1. Formula

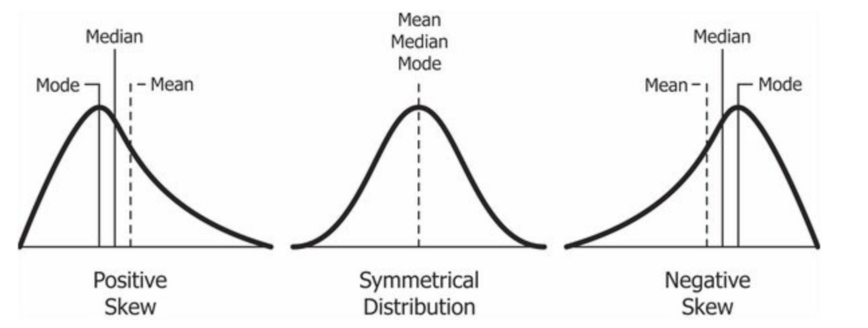


* + 1. Unbiased Estimate: s/sqrt(n-1)
  1. Coefficient of Variation (CV) or Relative Standard Dev. (RSD)
     1. Wiki: “The coefficient of variation should be computed only for data measured on a ratio scale, that is, scales that have a meaningful zero and hence allow relative comparison of two measurements… Kelvin temperature has a meaningful zero, the complete absence of thermal energy, and thus is a ratio scale. In plain language, it is meaningful to say that 20 Kelvin is twice as hot as 10 Kelvin, but only in this scale with a true absolute zero.”
     2. Advantages
        1. “The coefficient of variation is useful because the **standard deviation of data must always be understood in the context of the mean of the data***.* In contrast, the actual value of the **CV is independent of the unit in which the measurement has been taken**, so it is a dimensionless number. **For comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation.**
     3. Disadvantages
        1. “When the mean value is close to zero, the coefficient of variation will approach infinity and is therefore sensitive to small changes in the mean. This is often the case if the values do not originate from a ratio scale.”
        2. “Unlike the standard deviation, it cannot be used directly to construct confidence intervals for the mean.”
        3. “CVs are not an ideal index of the certainty of measurement when the number of replicates varies across samples because CV is invariant to the number of replicates while the certainty of the mean improves with increasing replicates. In this case, standard error in percent is suggested to be superior”
     4. Formula

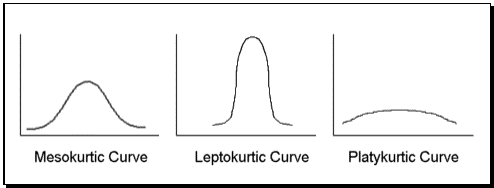


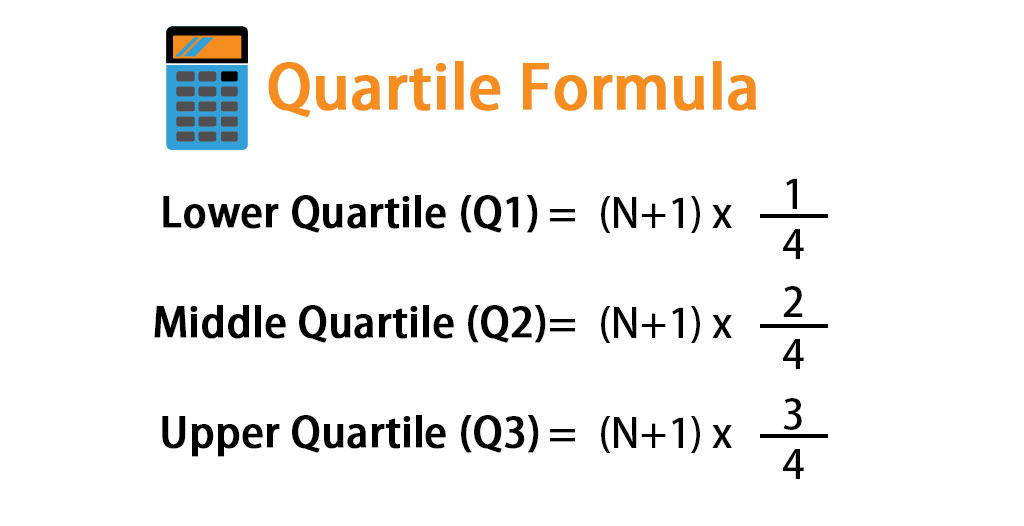
* + 1. Applications
       1. “The coefficient of variation is also common in applied probability fields such as renewal theory, queueing theory, and reliability theory. In these fields, the exponential distribution is often more important than the normal distribution. The standard deviation of an exponential distribution is equal to its mean, so its coefficient of variation is equal to 1.
       2. **Distributions with CV < 1 (such as an Erlang distribution) are considered low-variance, while those with CV > 1 (such as a hyper-exponential distribution) are considered high-variance**[citation needed]. Some formulas in these fields are expressed using the squared coefficient of variation, often abbreviated SCV. In modeling, a variation of the CV is the CV(RMSD). Essentially the CV(RMSD) replaces the standard deviation term with the Root Mean Square Deviation (RMSD). While many natural processes indeed show a correlation between the average value and the amount of variation around it, accurate sensor devices need to be designed in such a way that the coefficient of variation is close to zero, i.e., yielding a constant absolute error over their working range.
       3. In actuarial science, the CV is known as unitized risk.[14]
       4. In Industrial Solids Processing, CV is particularly important to measure the degree of homogeneity of a powder mixture. Comparing the calculated CV to a specification will allow to define if a sufficient degree of mixing has been reached.”
       5. In finance, lower volatility on mean returns indicates lower risk
  1. Skewness
     1. “In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined.
     2. For a unimodal distribution (one highest value), **negative skew commonly indicates that the tail is on the left side of the distribution**, and **positive skew indicates that the tail is on the right**. In cases where one tail is long but the other tail is fat, skewness does not obey a simple rule. For example, a zero value means that the tails on both sides of the mean balance out overall; this is the case for a symmetric distribution, but can also be true for an asymmetric distribution where one tail is long and thin, and the other is short but fat.”





* 1. Kurtosis
     1. “Along with skewness, kurtosis is an important descriptive statistic of data distribution. However, the two concepts must not be confused with each other. **Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.**
     2. In finance, kurtosis is used as a measure of financial risk. *A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns.* On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.”
     3. An excess kurtosis is a metric that compares the kurtosis of a distribution against the kurtosis of a normal distribution. The kurtosis of a normal distribution equals 3. Therefore, the excess kurtosis is found using the formula below: **Excess Kurtosis = Kurtosis – 3**
     4. Types
        1. Mesokurtic: (Excess) Kurtosis is near zero so the distribution is normal
        2. Leptokurtic: Positive excess kurtosis indicates large outliers. In finance, this indicates greater risk because there is a propensity for extreme values
        3. Platykurtic: Negative excess kurtosis indicates a distribution with flat tails (that is there are a small number of outliers). This is a lower risk for investors because it is less likely for extreme values.



* 1. IQR: Q3 – Q1
     1. 
     2. Lower Quartile is the median of lower half of data
     3. Upper Quartile is the median of the upper half
     4. Middle Quartile is the Median of the data
  2. Variance
     1. Variance = std.dev.^2
     2. “variance is the expectation of the squared deviation of a random variable from its mean”
     3. “The variance is the square of the standard deviation, the second central moment of a distribution, and the covariance of the random variable with itself, and it is often represented by σ ^2 s^2, or Var ( X )”

1. Deductive Approach
   1. Based on predetermined structures to analyze the data. Usually used when the researcher has general knowledge of the expected results.
   2. Frequentist Stats
2. Inductive Approach
   1. Is not based on predetermined structures or prior knowledge (little knowledge of subject or expected outcome)
   2. Bayesian Stats
3. Bias
   1. Selection: The sample population does not reflect the true population
   2. Non-Response Bias: In studies, there is usually a response bias of respondents
   3. Social Desirability Bias: “In social science research, social-desirability bias is a type of response bias that is the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. It can take the form of over-reporting "good behavior" or under-reporting "bad", or undesirable behavior.”
4. Hypothesis Testing
   1. Formulate Null & Alternative
      1. Null assumes there is no relationship between control and experimental group/variable(s)
   2. Select Significance Level
   3. Collect Data
   4. Compute p-value: the p-value is the probability the data would be at least as extreme as those observed if the null hypothesis were true
   5. If p <= alpha, then reject null & accept Alt. Hypothesis
5. Types of Sampling
   1. Simple Random Sample
   2. Systemic Sampling: sample every k’th value in sample
      1. Systemic Random Sample: k = N/n (N= pop size, n = sample size, k=sampled value)
   3. Stratified random Sample
      1. One approach is proportionate stratification. With proportionate stratification, the sample size of each stratum is proportionate to the population size of the stratum. Strata sample sizes are determined by the following equation :
         1. n\_h = (N\_h/N) \* n
         2. N = pop size, N\_h size of population of stratum, n is total sample size
      2. Stratified random sampling divides a population into subgroups. Random samples are taken in the same proportion to the population from each of the groups or strata. The members in each stratum (singular for strata) formed have similar attributes and characteristics.
   4. Cluster Sampling = Samples w/in a cluster?
      1. <https://www.scribbr.com/methodology/cluster-sampling/>
6. Types of Error
   1. Type I: False Positive or the rejection of a true null hypothesis for a false acceptance of the alternative hypothesis.
   2. Type II: False Negative or the failure to reject the null hypothesis when the alternative hypothesis is true (a guilty person is not convicted)