DESCRIPTIVE ANALYSIS

- A sample of an attribute ordinarily comprises several measurements, which are best understood when organized in some way. This is an important aspect of statistics.
- The number of measurements in a sample is the sample size.
- There are multiple options to make the data more intelligible. Some of these are more convenient than others, depending on factors such as the sample size and the ultimate objectives of the study.

SAMPLE VISUALIZATION

FREQUENCY TABLE

Given some numerical information, if the interval of variation of the data is divided into class intervals—customarily of the same lengths—and all observations are assigned to their corresponding classes, the result is a count of relative frequency of the classes.

UNCF EXAMPLE FREQUENCY TABLE

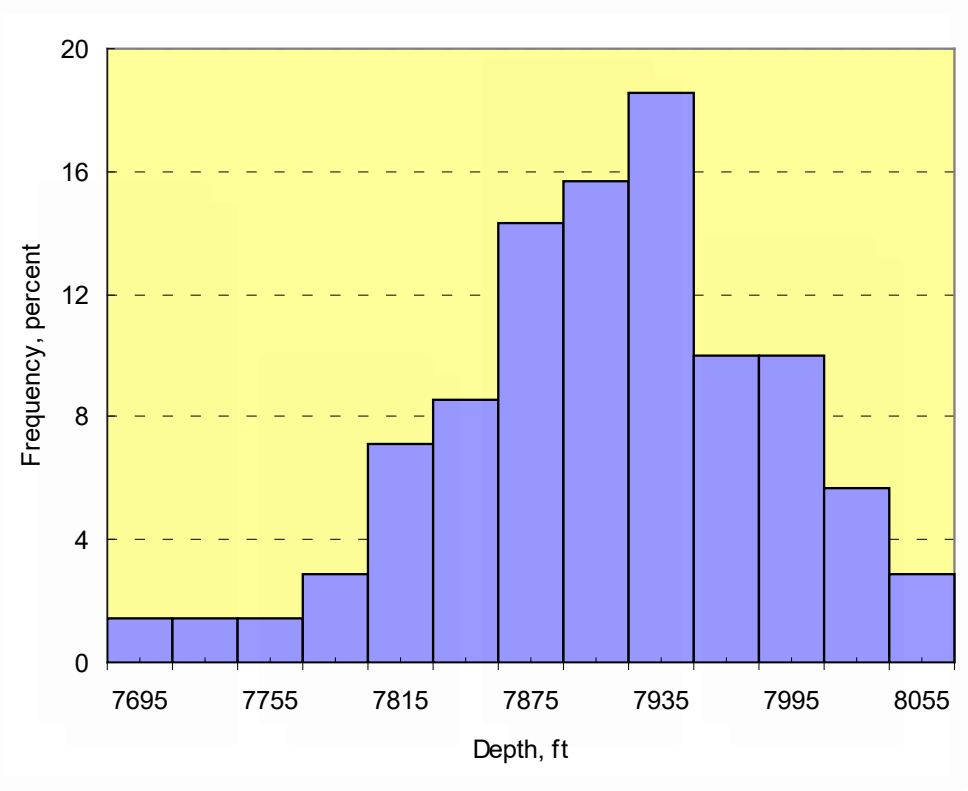
Class		Count	Frequency, %
1	7,680-7,710	1	1.43
2	7,710-7,740	1	1.43
3	7,740-7,770	1	1.43
4	7,770-7,800	2	2.86
5	7,800-7,830	5	7.14
6	7,830-7,860	6	8.57
7	7,860-7,890	10	14.29
8	7,890-7,920	11	15.71
9	7,920-7,950	13	18.57
10	7,950-7,980	7	10.00
11	7,980-8,010	7	10.00
12	8,010-8,040	4	5.71
13	8,040-8,070	2	2.86
Total		70	100.00

This example from a major oil company relates to depth in feet to an unconformity (UNCF) in an undisclosed area.

It will be used as a common reference to graphically illustrate other definitions in this chapter.

USGS

HISTOGRAM

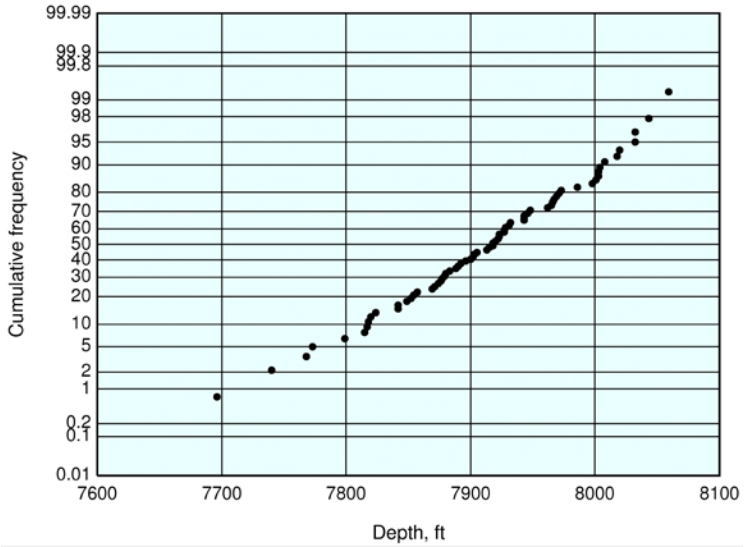


A histogram is a graphical representation of a frequency table.

USGS

CUMULATIVE FREQUENCY

Summaries based on frequency tables depend on the selection of the class interval and origin.



Given a sample of size n, this drawback is eliminated by displaying each observation z_i versus the proportion of the sample that is not larger than z_i.

Each proportion is a multiple of $100/n$. The vertical axis is divided in n intervals and the data are displayed at the center of the corresponding interval.

Customarily, the vertical axis is scaled so that data from a normal distribution (page 49) display as a straight line.

USGS

SUMMARY STATISTICS

SUMMARY STATISTICS

Summary statistics are complementary or alternative descriptors to histograms and cumulative distributions.

A statistic is a synoptic value that is calculated from a sample of observations, which is usually, but not necessarily, an estimator of some population parameter.

Generally, summary statistics are subdivided into three categories:

- Measures of location or centrality
- Measures of spread or dispersion
- Measures of shape

USGS

MEASURES OF LOCATION

Measures of location give an idea about the central tendency of the data. They are:

- mean
- median
- mode

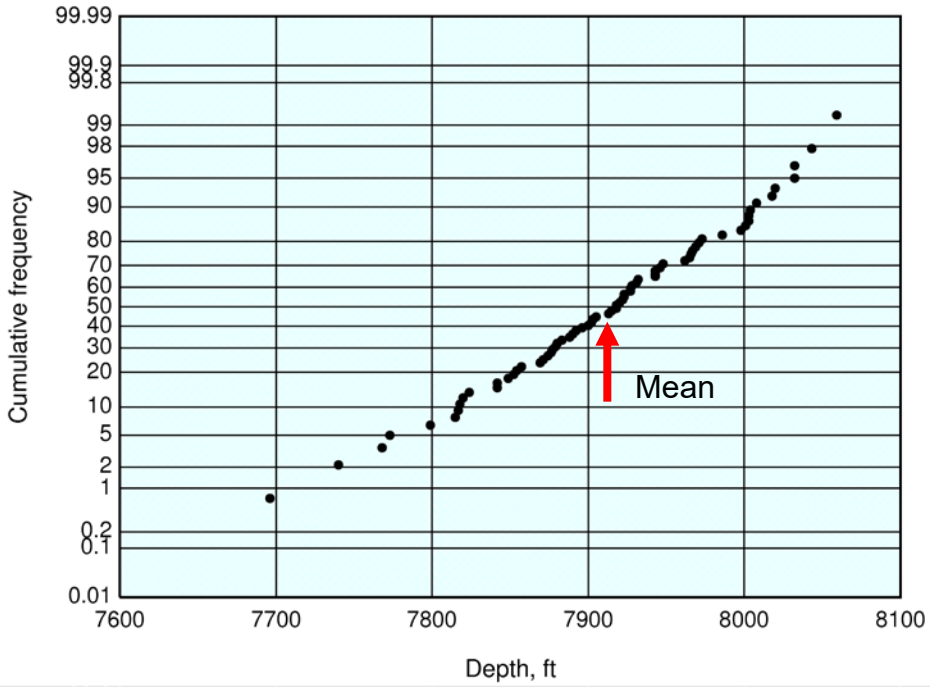
USGS

MEAN

The arithmetic mean or simply the mean, $\hat{m}$, of a sample of size n is the additive average of all the observations, z_i:

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} z_i .$$

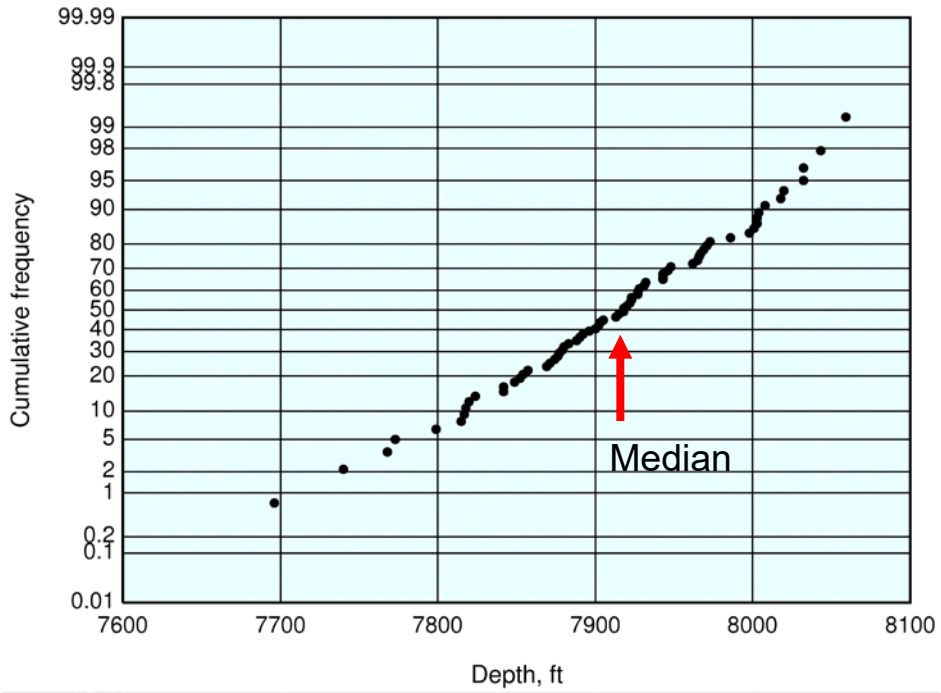
The mean of the UNCF sample is 7,912.2 ft.



MEDIAN

The median, Q_2, of a sample is the value that evenly splits the number of observations z_i into a lower half of smaller observations and an upper half of larger measurements.

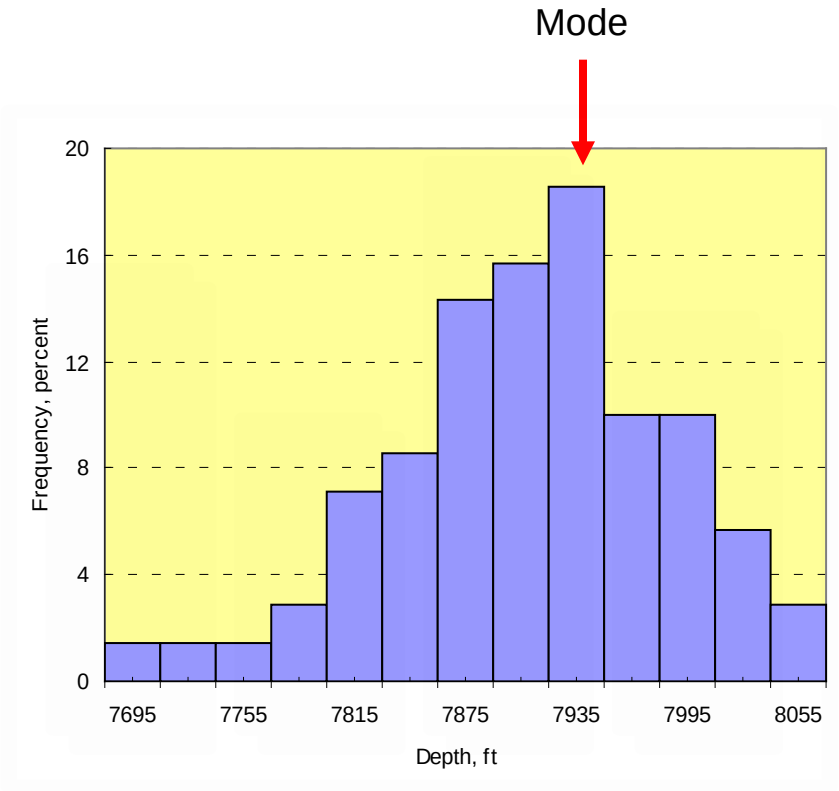
If z_i is sorted by increasing values, then



$$Q_2 = \begin{cases} z_{(n+1)/2}, \text{ if } n \text{ is odd}, \\ 0.5 \cdot \left(z_{n/2} + z_{(n/2)+1} \right), \text{ if } n \text{ is even}. \end{cases}$$

The median of the UNCF sample is 7,918 ft.

MODE



The mode of a sample is the most probable or frequent value, or, equivalently, the center point of the class containing the most observations.

For the UNCF sample, the center of the class with the most observations is 7,935 ft.

ROBUSTNESS

Robustness denotes the ability of statistical methods to work well not only under ideal conditions but also in the presence of data problems, mild to moderate departures from assumptions, or both.

For example, in the presence of large errors, the median is a more robust statistic than the mean.

MEASURES OF SPREAD

Measures of spread provide an idea of the dispersion of the data. The most common measures are:

- variance
- standard deviation
- extreme values
- quantiles
- interquartile range

USGS

VARIANCE

The variance, $\hat{\sigma}^2$, is the average squared dispersion around the mean:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - m)^2 = \frac{1}{n}\left[\sum_{i=1}^{n} z_i^{\,2} - n \cdot m^2\right],$$

expressions that are commonly restricted to estimate variances of finite populations.

When dealing with samples, the denominator is often changed to *n* - 1.

Because this is a quadratic measure, it is less robust than most other measures of spread.

The variance of the UNCF sample is 5,474 sq ft.

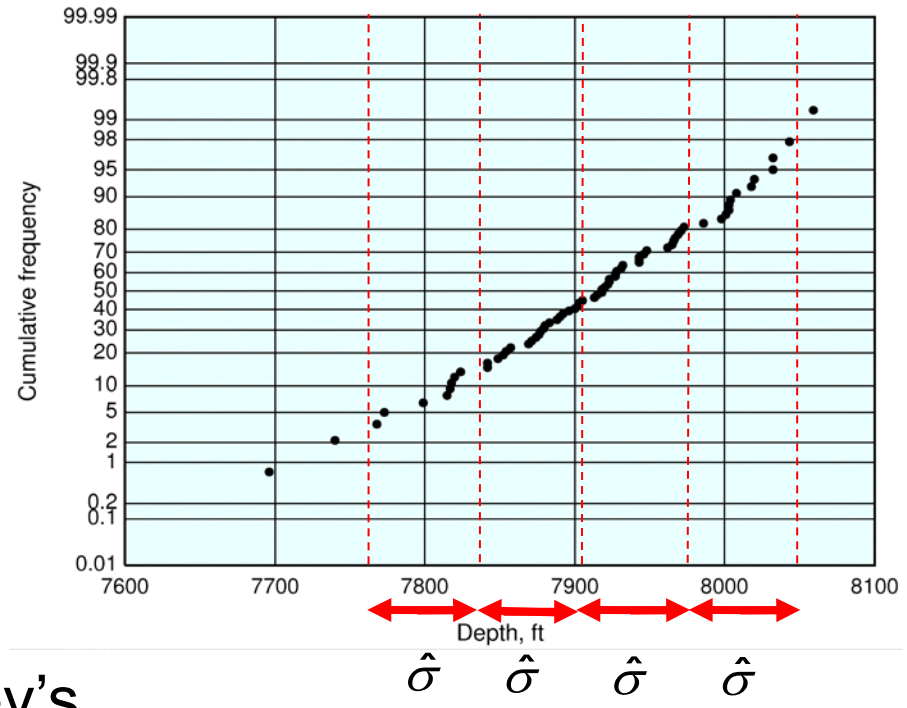
USGS

STANDARD DEVIATION

The standard deviation is the positive square root of the variance.

It has the advantage of being in the same units as the attribute.

The standard deviation of the UNCF sample is 74.5 ft.



According to Chebyshev's theorem, for any sample and $t > 1$, the proportion of data that deviates from the mean $\hat{m}$ at least $t \cdot \hat{\sigma}$ is at most t^{-2}:

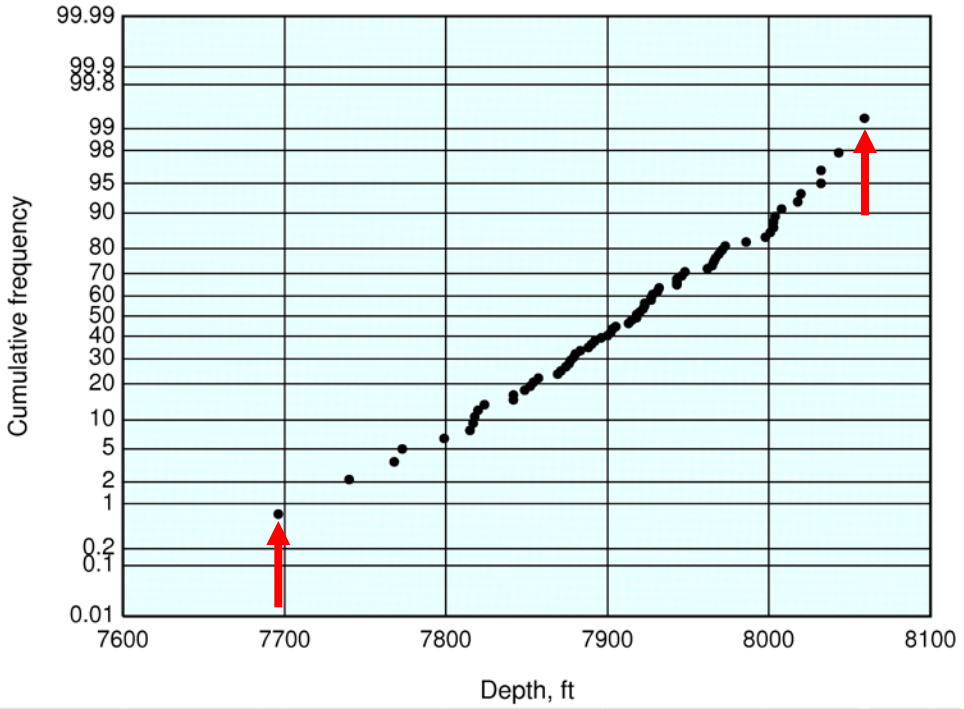
$$\text{Prop}\left(\left|X - \hat{m}\right| \geq t \cdot \hat{\sigma}\right) \leq \frac{1}{t^2}$$

EXTREME VALUES

The extreme values are the minimum and the maximum.

For the UNCF sample, the minimum value is 7,696 ft and the maximum value is 8,059 ft.

This measure is not particularly robust, especially for small samples.



QUANTILES

The idea of the median splitting the ranked sample into two equal-size halves can be generalized to any number of partitions with equal numbers of observations. The partition boundaries are called quantiles or fractiles. The names for the most common quantiles are:

- Median, for 2 partitions
- Quartiles, for 4 partitions
- Deciles, for 10 partitions
- Percentiles, for 100 partitions

The number of boundaries is always one less than the number of partitions.