



2A. General Statistics

Introduction

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting, and presenting empirical data. The data are either described or used to make an inference about a population.

General Statistics Overview

Populations Versus Samples

The science of statistics can be focused on data from either a population or a sample. A population is an entire group about which specific information will be obtained. A population can be finite (e.g., the weight of every student in the 2nd-grade class at Hypothetical Elementary on 3/1/2022), or it can be very large, infinite, or continuing to emerge (e.g., the weight of every 2nd-grade student in the world), making it impractical or impossible to have values for all items in the population. If less than the entire population is in the data set, then the data set is a subset of the population and is referred to as a sample.

Descriptive and Inferential Statistics

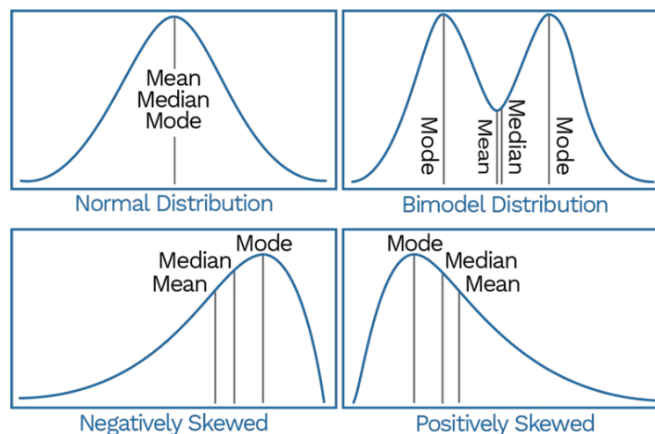
Descriptive statistics reduces a large number of values to a manageable few or presents the values in a way to facilitate visualization. Descriptive statistics are calculated using data (from a sample or from a population) and can be presented in graphical, tabular, or numerical ways. Examples of graphical and tabular summaries include histograms, probability plots, and frequency distributions. Examples of numerical descriptive statistics include mean, variance, and standard deviation.

Unlike descriptive statistics, inferential statistics uses a sample to generalize about a population. The sample must be representative of the population and the process used to select the sample must be repeatable. This may be referred to as random or probability sampling.

Measure of Central Tendency

A measure of central tendency is a method to represent a population or data set by a single point at the center of its data values. The mean (average value) is the most widely used statistic to describe the central tendency of a data set. The median (middle value) is the next most widely used statistic to express the central tendency of a data set. For symmetric data sets, the mean and median will be the same. Other statistics that are used to express the central tendency of a data set are midrange (halfway between the minimum and maximum values) and mode (the value occurring most often). A data set will have no mode if each value occurs only once in the data set, or the data set may be bimodal or multimodal if values occur with the same frequency (see figure 1).

Figure 1. Examples of Distributions and Skews



Measure of Dispersion

Similarly, there is often a desire to represent a population or a sample by a single number that expresses the data's dispersion (variability, scatter, or spread) around the center of the distribution. The range (the difference between highest and lowest values) is the simplest statistic to calculate. Due to the sensitivity of a range statistic to extreme values, it may be preferable to provide the maximum and minimum values or the interquartile range that focuses on the middle 50% of the values in the data set. A large interquartile range means that there is a large spread, greater variation, between these values.

Standard deviation is the most used measure of dispersion. It is appropriate to use standard deviation as a measure of dispersion when the mean is used to measure central tendency. Standard deviation compares each data point to the mean of all data points. Variance also measures the deviation of the data from the mean and is the square of the standard deviation.

Distributions

The frequency of each value can be displayed graphically with the values on the x-axis and the frequency of occurrence of each value on the y-axis. If the data are limited to distinct, separate values, known as a discrete random variable, the data will be displayed as the different values that are possible over the interval (e.g., quality control data from a balance with a readability to 0.01 gram would be displayed in bins of 0.01 g each). If the entries in the data set across the stated interval can take on any value with an infinite number of decimal places possible, a continuous random variable, then the distribution, or density curve, becomes smooth (e.g., quality control data for a quantitative method).

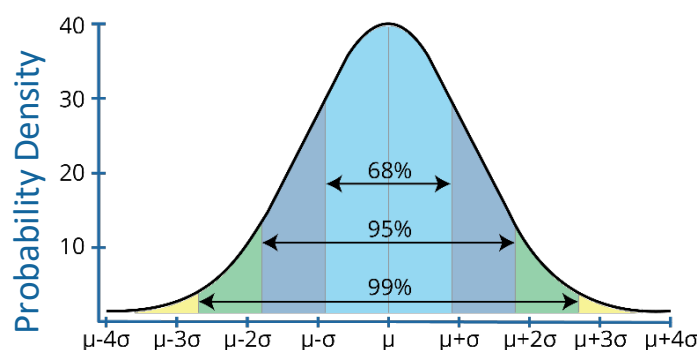


The Normal Distribution

Many people are familiar with normal distribution, also referred to as a Gaussian distribution or a bell curve. Data from many natural phenomena as well as measurement processes can be approximated by a normal distribution. This is a symmetric distribution where the majority of the data values center around the mean, and data values at the outside margins of the distribution are less likely to be observed. The mean, median, and mode are the same value in a normal distribution.

If data from a process approaches a normal distribution with a population mean μ and population standard deviation σ , then the probability that the next value from that same process will be within one standard deviation of the mean is approximately 68%. The probability that the next value will be within two standard deviations of the mean is approximately 95%, and the probability that the next value will be within three standard deviations of the mean is approximately 99% (see figure 2). If data are not normally distributed, then statements regarding “the next value” can be made but the percentages may differ.

Figure 2. Normal Distribution (μ = mean, σ = standard deviation)



It is possible for two data sets to have the same mean or the same standard deviation; therefore, when describing a data set, it is preferable to provide both a descriptor of central tendency (e.g., mean) and of dispersion (e.g., standard deviation).

Related Primers

Statistical Sampling

Probability and Likelihood Ratios

Part of NIST SP 1500-28

<https://doi.org/10.6028/NIST.SP.1500-28>

Forensic Science Example

Table 1: Example calculations for ongoing quality control statistics of central tendency and dispersion using control data for a quantitative method for an attribute where there is a legal specification at “20”. Decimal places have been truncated.

Data Set	Describes	Sample Statistic	Calculation and/or Value for this data
20.000 19.997 19.998 19.999 19.997 20.003 20.003 20.001 20.000 20.002 19.998 19.999 20.000 20.003	Size	Sample: n	= 25
19.999 19.997 20.003 20.003 20.001 20.000 20.002 19.998 19.999 20.000 20.003	Central Tendency	Mean: $\bar{x}$	= 20.000
20.003 20.001 20.000 20.002 19.998 19.999 20.000 20.003	Central Tendency	Median	= 20.000
19.999 20.000 20.003 19.997 20.001 20.000 20.001 19.997 20.000 20.002 20.003 20.000	Central Tendency	Midrange	= 20.000
19.999 20.000 20.003	Central Tendency	Mode	= 20.003
19.997 20.001 20.000	Dispersion	Range	= 0.006
20.001 20.003 20.001 19.997 20.000 20.002 20.003 20.000	Dispersion	Interquartile Range	= 0.003
19.997 20.000 20.002 20.003 20.000	Dispersion	Standard Deviation: s	= 0.002
20.002 20.003 20.000	Dispersion	Variance: s^2	= 0.000004

Key Takeaways

- 1 A sample of one or more items is a subset of a population that contains all items under consideration.
- 2 Mean, median, mode, and midrange are measures of central tendency.
- 3 Standard deviation, variance, range, and interquartile range are measures of dispersion.
- 4 Both a descriptor of central tendency and of dispersion are required to adequately describe a data set.

Learn More

[Khan Academy](#) – search for statistics

ASQ's [Introduction to Statistical Concepts](#)

ANAB's [Uncertainty, Sampling and Data Analysis: Understanding Statistical Calculations](#) (course)

[ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics](#)



Glossary: Primer 2A General Statistics

Term	Definition	Reference	Primer Nos.
Descriptive statistics	Graphical, numerical or other summary depiction of observed values	ISO 3534-1:2006 Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability, International Organization for Standardization, Geneva, Switzerland	2A General Statistics
Interquartile range	IQR, n —the 75th percentile (0.75 quantile) minus the 25th percentile (0.25 quantile), for a data set. $IQR=Q3-Q1$	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics
Mean (arithmetic mean, average)	n —of a population, μ , average or expected value of a characteristic in a population – of a sample, $\bar{x}$, sum of the observed values in the sample divided by the sample size.	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics
Median	n —the 50th percentile in a population or sample	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics
Midrange	n —average of the minimum and maximum values in a sample	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics
Mode	The value occurring most frequently in a data set	ASQ Quality Glossary	2A General Statistics
Population	Totality of items under consideration	ISO 3534-1:2006 Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability, International Organization for Standardization, Geneva, Switzerland	2A General Statistics; 2C Statistical Sampling; 2D Population Statistics



Term	Definition	Reference	Primer Nos.
Population parameter	Summary measure of the values of some characteristic of a population	ISO 3534-2:2006 Statistics — Vocabulary and symbols — Part 2: Applied statistics, International Organization for Standardization, Geneva, Switzerland	2A General Statistics
Range	R, n—maximum value minus the minimum value in a sample	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics
Sample (General Statistics)	Subset of a population made up of one or more sampling units	ISO 3534-1:2006 Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability, International Organization for Standardization, Geneva, Switzerland	2A General Statistics; 2C Statistical Sampling; 2D Population Statistics
Sample statistic	Summary measure of the observed values of a sample	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics
Standard deviation	Of a population, σ , the square root of the average or expected value of the squared deviation of a variable from its mean; —of a sample, s , the square root of the sum of the squared deviations of the observed values in the sample from their mean divided by the sample size minus 1	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics; 2B Method Performance Statistics
Statistic	See sample statistic	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics
Statistical process control (SPC)	the use of statistical techniques to control a process or production method	ASQ Quality Glossary	2A General Statistics
Variance	σ^2 , s^2 , n —square of the standard deviation of the population or sample	ASTM E2586-19e1 Standard Practice for Calculating and Using Basic Statistics, ASTM International, West Conshohocken, PA, 2006	2A General Statistics