Sample Variance

we have presented statistics that describe the central tendencies of a data set, we are also interested in ones that describe the spread or variability of the data values. A statistic that could be used for this purpose would be one that measures the average value of the squares of the distances between the data values and the sample mean. This is accomplished by the sample variance.

The sample variance, call it s^2, of the data set $x_1, \ldots, x_n$ is defined by

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Example Problem

Find the sample variances of the data sets A and B given below.

$$A : 3, 4, 6, 7, 10 \qquad B : -20, 5, 15, 24$$

An algebraic identity

The following algebraic identity is often useful for computing the sample variance:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Proof:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

Sample variance

The computation of the sample variance can also be eased by noting that if

$$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

then $\bar{y} = a + b\bar{x}$, and

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

That is, if s_y^2 and s_x^2 are the respective sample variances, then

$$s_y^2 = b^2 s_x^2$$

In other words, adding a constant to each data value does not change the sample variance; whereas multiplying each data value by a constant results in a new sample variance that is equal to the old one multiplied by the square of the constant.

Example Problem

The following data give the worldwide number of fatal airline accidents of commercially scheduled air transports in the years from 1985 to 1993.

Year	Accidents
1985	22
1986	22
1987	26
1988	28
1989	27
1990	25
1991	30
1992	29
1993	24

Find sample variance.

Sample Standard Deviation

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

The quantity s, defined by

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

Sample Percentile

The sample $100p$ percentile is that data value such that $100p$ percent of the data are less than or equal to it and $100(1-p)$ percent are greater than or equal to it. If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these two values.

To determine the sample $100p$ percentile of a data set of size n, we need to determine the data values such that

1. At least np of the values are less than or equal to it.
2. At least $n(1-p)$ of the values are greater than or equal to it.

The sample 25 percentile is called the first quartile; the sample 50 percentile is called the sample median or the second quartile; the sample 75 percentile is called the third quartile.

Example Problem

Table lists the populations of the 25 most populous U.S. cities for the year 1994. For this data set, find (a) the sample 10 percentile and (b) the sample 80 percentile.

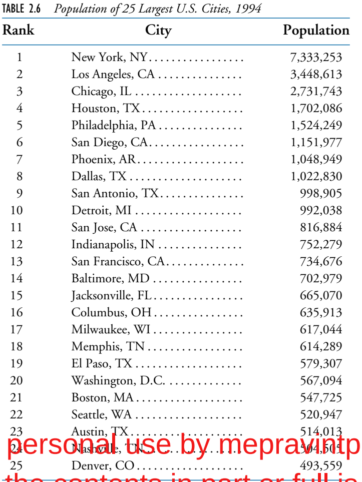

TABLE 2.6 *Population of 25 Largest U.S. Cities, 1994*

Rank	City	Population
1	New York, NY	7,333,253
2	Los Angeles, CA	3,448,613
3	Chicago, IL	2,731,743
4	Houston, TX	1,702,086
5	Philadelphia, PA	1,524,249
6	San Diego, CA	1,151,977
7	Phoenix, AR	1,048,949
8	Dallas, TX	1,022,830
9	San Antonio, TX	998,905
10	Detroit, MI	992,038
11	San Jose, CA	816,884
12	Indianapolis, IN	752,279
13	San Francisco, CA	734,676
14	Baltimore, MD	702,979
15	Jacksonville, FL	665,070
16	Columbus, OH	635,913
17	Milwaukee, WI	617,044
18	Memphis, TN	614,289
19	El Paso, TX	579,307
20	Washington, D.C.	567,094
21	Boston, MA	547,725
22	Seattle, WA	520,947
23	Austin, TX	514,013
24	Nashville, TN	504,505
25	Denver, CO	493,559

Example Problem with box plot

Noise is measured in decibels, denoted as dB. One decibel is about the level of the weakest sound that can be heard in a quiet surrounding by someone with good hearing; a whisper measures about 30 dB; a human voice in normal conversation is about 70 dB; a loud radio is about 100 dB. Ear discomfort usually occurs at a noise level of about 120 dB.

The following data give noise levels measured at 36 different times directly outside of Grand Central Station in Manhattan.

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85 69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

Determine the quartiles.

Example Problem with box plot

Noise is measured in decibels, denoted as dB. One decibel is about the level of the weakest sound that can be heard in a quiet surrounding by someone with good hearing; a whisper measures about 30 dB; a human voice in normal conversation is about 70 dB; a loud radio is about 100 dB. Ear discomfort usually occurs at a noise level of about 120 dB.

The following data give noise levels measured at 36 different times directly outside of Grand Central Station in Manhattan.

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85 69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

Determine the quartiles.

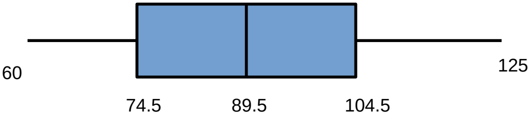

Figure: Box plot

Example Problem

Table lists the populations of the 25 most populous U.S. cities for the year 1994. For this data set, find (a) the sample 10 percentile and (b) the sample 80 percentile.

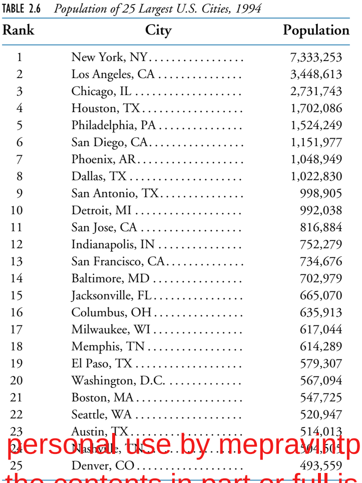

TABLE 2.6 *Population of 25 Largest U.S. Cities, 1994*

Rank	City	Population
1	New York, NY	7,333,253
2	Los Angeles, CA	3,448,613
3	Chicago, IL	2,731,743
4	Houston, TX	1,702,086
5	Philadelphia, PA	1,524,249
6	San Diego, CA	1,151,977
7	Phoenix, AR	1,048,949
8	Dallas, TX	1,022,830
9	San Antonio, TX	998,905
10	Detroit, MI	992,038
11	San Jose, CA	816,884
12	Indianapolis, IN	752,279
13	San Francisco, CA	734,676
14	Baltimore, MD	702,979
15	Jacksonville, FL	665,070
16	Columbus, OH	635,913
17	Milwaukee, WI	617,044
18	Memphis, TN	614,289
19	El Paso, TX	579,307
20	Washington, D.C.	567,094
21	Boston, MA	547,725
22	Seattle, WA	520,947
23	Austin, TX	514,013
24	Nashville, TN	504,505
25	Denver, CO	493,559

Chebyshev's Inequality

Statement: Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set. Assuming that $s > 0$,
Chebyshev's inequality states that for any value of $k \geq 1$, greater than $100(1 - 1/k^2)$ percent of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$.

Chebyshev's Inequality

Statement: Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set. Assuming that $s > 0$,
Chebyshev's inequality states that for any value of $k \geq 1$, greater than $100(1 - 1/k^2)$ percent of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$.
Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, \ldots, x_n$, where $s > 0$. Let

$$S_k = \{i, 1 \leq i \leq n : |x_i - \bar{x}| < ks\}$$

and let $N(S_k)$ be the number of elements in the set S_k. Then, for any $k \geq 1$,

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$