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 $k \geq 1$,

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

Chebyshev's Inequality:Proof

$$
\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \sum_{i \in S_k}(x_i-\bar{x})^2 - \sum_{i \notin S_k}(x_i-\bar{x})^2 \\
&\geq \sum_{i \notin S_k}(x_i-\bar{x})^2 \\
&\geq \sum_{i \notin S_k} k^2 s^2 \\
&= k^2 s^2 (n - N(S_k))
\end{aligned}
$$

where the first inequality follows because all terms being summed are non–negative, and the second follows since $(x_1-\bar{x})^2 \geq k^2s^2$ when $i \notin S_k$. Dividing both sides of the preceding inequality by nk^2s^2 yields that

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

Example Problem

Table lists the 10 top-selling passenger cars in the United States in 1999. A simple calculation gives that the sample mean and sample standard deviation of

TABLE 2.7 *Top 10 Selling Cars for 1999*

	1999	
1.	Toyota Camry	448,162
2.	Honda Accord	404,192
3.	Ford Taurus	368,327
4.	Honda Civic	318,308
5.	Chevy Cavalier	272,122
6.	Ford Escort	260,486
7.	Toyota Corolla	249,128
8.	Pontiac Grand Am	234,936
9.	Chevy Malibu	218,540
10.	Saturn S series	207,977

Example Problem

Table lists the 10 top-selling passenger cars in the United States in 1999. A simple calculation gives that the sample mean and sample standard deviation of

TABLE 2.7 *Top 10 Selling Cars for 1999*

	1999	
1.	Toyota Camry	448,162
2.	Honda Accord	404,192
3.	Ford Taurus	368,327
4.	Honda Civic	318,308
5.	Chevy Cavalier	272,122
6.	Ford Escort	260,486
7.	Toyota Corolla	249,128
8.	Pontiac Grand Am	234,936
9.	Chevy Malibu	218,540
10.	Saturn S series	207,977

A simple calculation gives that the sample mean and sample standard deviation of these data are $\bar{x} = 298,217.8$ and $s = 124,542.9$

Suppose now that we are interested in the fraction of data values that exceed the sample mean by at least k sample standard deviations, where k is positive. That is, suppose that $\bar{x}$ and s are the sample mean and the sample standard deviation of the data set $x_1, x_2, \ldots, x_n$. Then, with

$$N(k) = \text{number of } i : x_i - \bar{x} \geq ks$$

$$\begin{aligned} \frac{N(k)}{n} &\leq \frac{\text{number of } i : x_i - \bar{x} \geq ks}{n} \\ &\leq \frac{1}{k^2} \text{ by Chebyshev's inequality} \end{aligned}$$

However, we can make a stronger statement, as is shown in the one-sided version of Chebyshevs inequality.

One Sided Chevbyshev Inequality

Statement: For $k > 0$,

$$\frac{N(k)}{n} \leq \frac{1}{1+k^2}$$

Proof: Let $y_i = x_i - \bar{x}, i = 1, \ldots, n$.For any $b > 0$,we have that

$$\begin{aligned}\sum_{i=1}^{n}(y_i+b)^2 &\geq \sum_{i:y_i \geq k_s}(y_i+b)^2 \\ &\geq \sum_{i:y_i \geq k_s}(ks+b)^2 \\ &\geq N(k)(ks+b)^2\end{aligned}$$

where the first inequality follows because $(y_i + b)^2 \geq 0$, and the second because both ks and b are positive.

However,

$$\begin{aligned}\sum_{i=1}^{n}(y_i+b)^2 &= \sum_{i=1}^{n}(y_i^2+2by_i+b^2)\\ &= \sum_{i=1}^{n}y_i^2+2b\sum_{i=1}^{n}y_i+nb^2\\ &= (n-1)s^2+nb^2\end{aligned}$$

where the final equation used that
$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$
Therefore, we obtain from equation in previous slide

$$N(k) \leq \frac{(n-1)s^2+nb^2}{(ks+b)^2)}$$

Implying that,

$$\frac{N(k)}{n} \leq \frac{s^2 + b^2}{(ks + b)^2}$$

Because the preceding is valid for all $b > 0$, we can set $b = \frac{s}{k}$ (which is the value of b that minimizes the right-hand side of the preceding) to obtain that

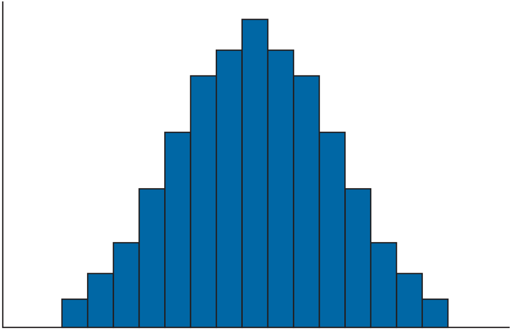
$$\frac{N(k)}{n} \leq \frac{s^2 + \frac{s^2}{k^2}}{(ks + \frac{s}{k})^2}$$

Multiplying the numerator and the denominator of the right side of the preceding by $\frac{k^2}{s^2}$ gives

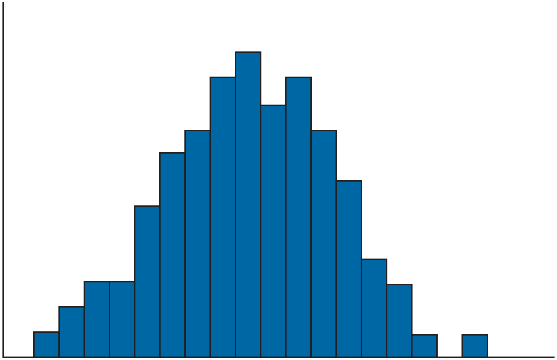
$$\frac{N(k)}{n} \leq \frac{k^2 + 1}{(k^1 + 1)^2} = \frac{1}{k^2 + 1}$$

Thus, for instance, where the usual Chebyshev inequality shows that at most 25 percent of data values are at least 2 standard deviations greater than the sample mean, the one-sided Chebyshev inequality lowers the bound to at most 20 percent.

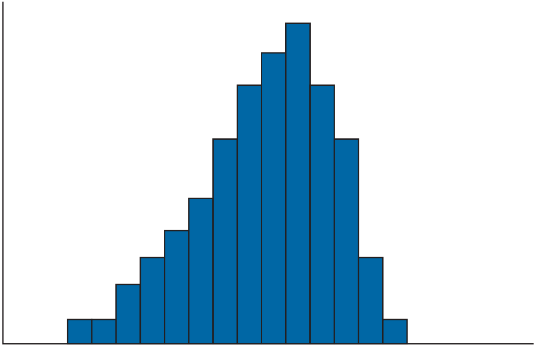
Histogram of normal data set



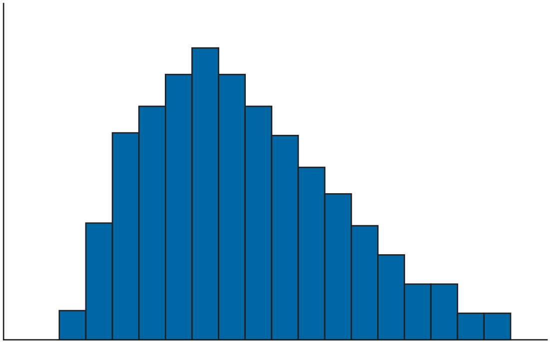
Histogram of approximately normal data set



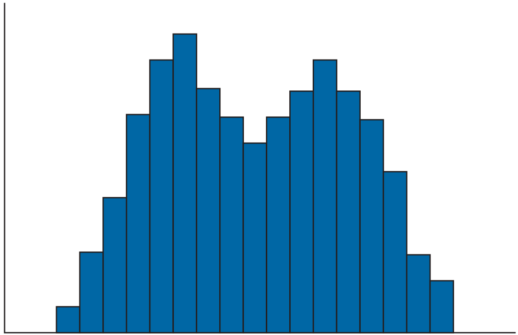
Histogram of a data set skewed to the left



Histogram of a data set skewed to the right



Histogram of Bimodal dataset



Empirical rule of normal datasets

If a data set is approximately normal with sample mean $\bar{x}$ and sample standard deviation s, then the following statements are true.

1. Approximately 68 percent of the observations lie within

$$\bar{x} \pm s$$

2. Approximately 95 percent of the observations lie within

$$\bar{x} \pm 2s$$

3. Approximately 99.7 percent of the observations lie within

$$\bar{x} \pm 3s$$

Example Problem

The following stem and leaf plot gives the scores on a statistics exam taken by industrial engineering students.

9	0,1,4
8	3,5,5,7,8
7	2,4,4,5,7,7,8
6	0,2,3,4,6,6
5	2,5,5,6,8
4	3,6

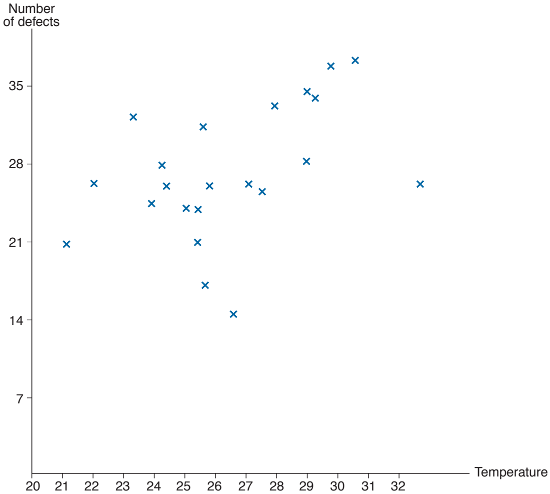
Use it to assess the empirical rule.

Paired data sets: Example Problem

TABLE 2.8 *Temperature and Defect Data*

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

Example Problem



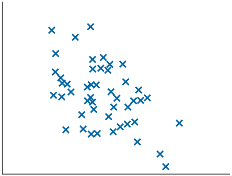
Paired data sets and correlations

Let s_x and s_y denote, respectively, the sample standard deviations of the x values and the y values. The sample correlation coefficient, call it r , of the data pairs $(x_i, y_i), i = 1, \ldots, n$ is defined by

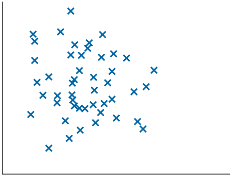
$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \end{aligned}$$

When $r > 0$ we say that the sample data pairs are positively correlated, and when $r < 0$ we say that they are negatively correlated.

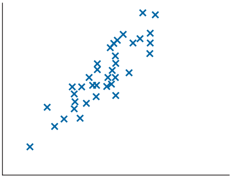
Different correlations



$r = -.50$



$r = 0$



$r = -.90$