S. Gopalakrishnan

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Sample Median

- Order the values of a data set of size n from smallest to largest.
- If n is odd, the sample median is the value in position $(n+1)/2$;
- if n is even, it is the average of the values in positions $n/2$ and $n/2+1$.
- Sample median; loosely speaking, it is the middle value when the data set is arranged in increasing order.

This file is meant for personal use by test_m7p4dnspx6@gradeai.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Example Problem

Find the sample median of the Data described

Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

This file is meant for personal use by [overlapping text] only.
Sharing or publishing the contents in part or full is liable for legal action.

Example Problem

In a study reported in Hoel, D. G., A representation of mortality data by competing risks, Biometrics, 28, pp. 475488, 1972, a group of 5-week-old mice were each given a radiation dose of 300 rad. The mice were then divided into two groups; the first group was kept in a germ-free environment, and the second in conventional laboratory conditions. The numbers of days until death were then observed. The data for those whose death was due to thymic lymphoma are given in the following stem and leaf plots (whose stems are in units of hundreds of days); the first plot is for mice living in the germ-free conditions, and the second for mice living under ordinary laboratory conditions.

This file is meant for personal use by test_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Example Problem

Germ-Free Mice

1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

Conventional Mice

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Sample Mode

Another statistic that has been used to indicate the central tendency of a data set is the sample mode, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called modal values.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Example Problem

The following frequency table gives the values obtained in 40 rolls of a die.

Age	Frequency
1	9
2	8
3	5
4	5
5	6
6	7

Find (a) the sample mean, (b) the sample median, and (c) the sample mode.

This file is meant for personal use by t_m_p4admsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

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

This file is meant for personal use by test_m7p4adrsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Example Problem

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

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

This file is meant for personal use by t_m7p4dnsx6@gmail.com only.
This file is meant for personal use by t_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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:

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

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

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.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

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.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Sample Standard Deviation

The positive square root of the sample variance is called the sample standard deviation.
The quantity s, defined by

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

This file is meant for personal use by test_m7p4dnsx6@gradient.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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.

This file is meant for personal use by test_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

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.

This file is meant for personal use by test_m7p4dnsx6@gradalcom only.
Sharing or publishing the contents in part or full is liable for legal action.

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.

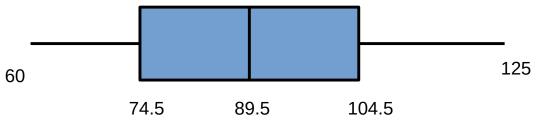


Figure: Box plot

This file is meant for personal use by tm_ep4admsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

This file is meant for personal use by [illegible]_7p4dnsp6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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 $\bar{x} + ks$.

This file is meant for personal use by [overlapping watermark text]@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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 $k \geq 1$,

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

This file is meant for personal use by t_m_p4dnsp6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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 $(x_1 - \bar{x})^2 \geq k^2 s^2$ when $i \notin S_k$. Dividing both sides of the preceding inequality by nk^2s^2 yields that

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

This file is meant for personal use by tm_7p4adnsxa6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

This file is meant for personal use by test_mtp4udnsxa@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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 $s = 124,542.9$

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

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.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

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.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

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

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

This file is meant for personal use by test_m7p4vdrsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

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

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

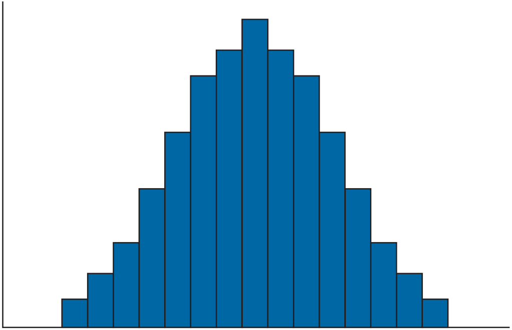
Multiplying the numerator and the denominator of the right side of the preceding by $\frac{k^2}{s^2}$ gives

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

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.

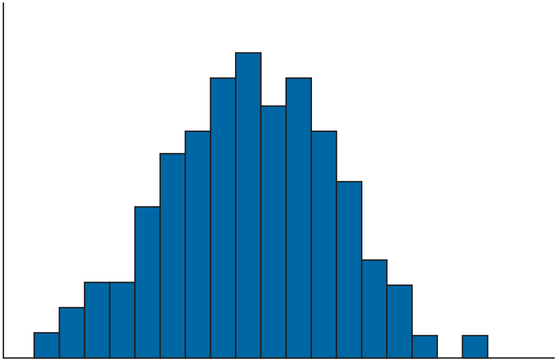
This file is meant for personal use by test_m7p4drnx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Histogram of normal data set



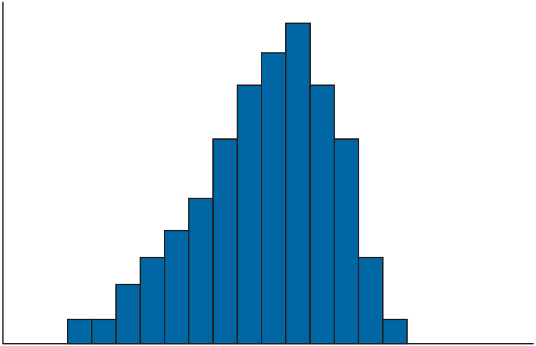
This file is meant for personal use by test_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Histogram of approximately normal data set



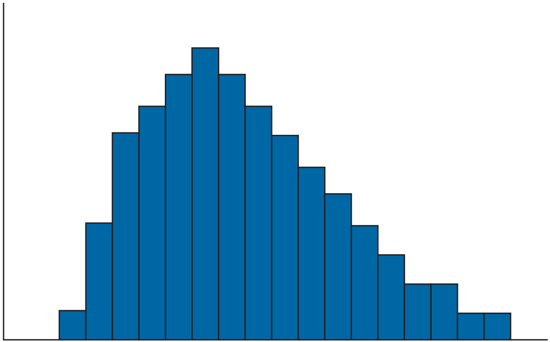
This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Histogram of a data set skewed to the left



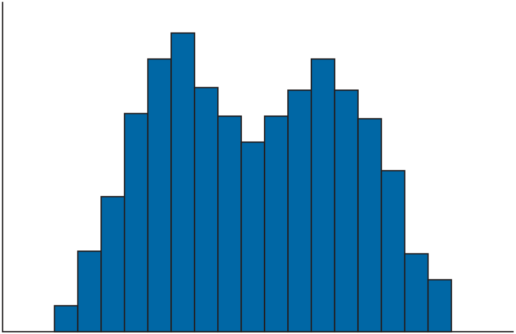
This file is meant for personal use by test_m7p4dmx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Histogram of a data set skewed to the right



This file is meant for personal use by test_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Histogram of Bimodal dataset



This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

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

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

This file is meant for personal use by [overlapping watermark text] only.
Sharing or publishing the contents in part or full is liable for legal action.

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.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

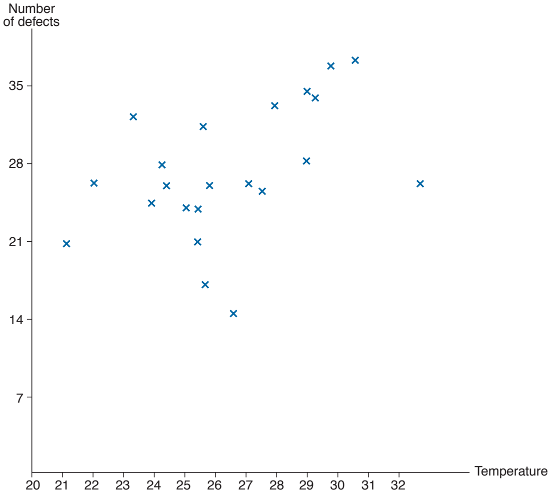
Paired data sets: Example Problem

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

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Example Problem



This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Paired data sets and correlations

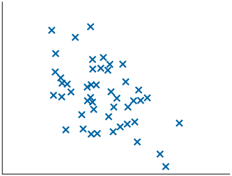
Let s_x and s_y denote, respectively, the sample standard deviations of the x values and the y values. The sample correlation coefficient, call it r , of the data pairs $(x_i, y_i), i = 1, \ldots, n$ is defined by

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

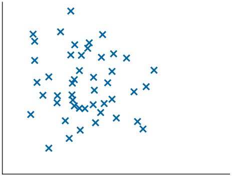
When $r > 0$ we say that the sample data pairs are positively correlated, and when $r < 0$ we say that they are negatively correlated.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

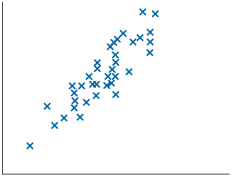
Different correlations



$r = -.50$



$r = 0$



$r = -.90$

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Properties of r

1. $-1 \leq r \leq 1$
2. If for constants a and b,with $b > 0$,
$$y_i = a + bx_i, \qquad i = 1, \ldots, n$$
then $r = 1$.
3. If for constants a and b,with $b < 0$,
$$y_i = a + bx_i, \qquad i = 1, \ldots, n$$
then $r = -1$.
4. If r is sample correlation coefficient for the data pairs $x_i, y_1, i = 1, \ldots, n$ then it is also the sample correlation coefficient of the data pairs.
$$a + bx_i, \quad c + dy_i, \quad i = 1, \ldots, n$$
provided that b and d are both positive or both negative.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Example Problem

The following data gives the resting pulse rates (in beats per minute) and the years of schooling of 10 individuals.

Person	1	2	3	4	5	6	7	8	9	10
Years of school	12	16	13	18	19	12	18	19	12	14
Pulse rate	73	67	74	63	7384	60	62	76	71	

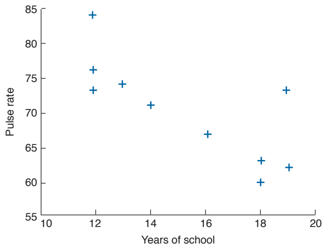
This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Example Problem

The following data gives the resting pulse rates (in beats per minute) and the years of schooling of 10 individuals.

Person	1	2	3	4	5	6	7	8	9	10
Years of school	12	16	13	18	19	12	18	19	12	14
Pulse rate	73	67	74	63	7384	60	62	76	71	

A scatter diagram of these data is presented in the figure. The sample correlation coefficient for these data is r = 0.7638.



This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

We will now prove the first three properties of the sample correlation coefficient r . That is, we will prove that $|r| \leq 1$ with equality when the data lie on a straight line. To begin, note that,

$$\sum \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0$$

$$\sum \frac{(x_i - \bar{x})^2}{s_x^2} + \sum \frac{(y_i - \bar{y})^2}{s_y^2} - 2 \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \geq 0$$

$$n - 1 + n - 1 - 2(n-1)r \geq 0$$

showing that

$$r \leq 0$$

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Note also that $r = 1$ if and only if there is equality in first equation on the previous slide. That is, $r = 1$ if and only if for all i

$$\frac{\bar{y} - y_i}{s_y} = \frac{\bar{x} - x_i}{s_x}$$

$$y_i = \bar{y} - \frac{s_y}{s_x}\bar{x} + \frac{s_y}{s_x}x_i$$

That is, $r = 1$ if and only if the data values (x_i, y_i) lie on a straight line having a positive slope.
To show that $r \geq -1$, with equality if and only if the data values (x_i, y_i) lie on a straight line having a negative slope, start with

$$\sum \left(\frac{x_i - \bar{x}}{s_x} + \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0$$

and use an argument analogous to the one just given.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Probability

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Sample space and Events

This set of all possible outcomes of an experiment is known as the **sample space** of the experiment and is denoted by S. Some examples are the following.

1. If the outcome of an experiment consists in the determination of the sex of a newborn child, then

   $$S = g, b$$

   where the outcome g means that the child is a girl and b that it is a boy.

2. If the experiment consists of the running of a race among the seven horses having post positions 1, 2, 3, 4, 5, 6, 7, then

   $$S = \{\text{all orderings of } (1, 2, 3, 4, 5, 6, 7)\}$$

   The outcome (2, 3, 1, 6, 5, 4, 7) means, for instance, that the number 2 horse is first, then the number 3 horse, then the number 1 horse, and so on.

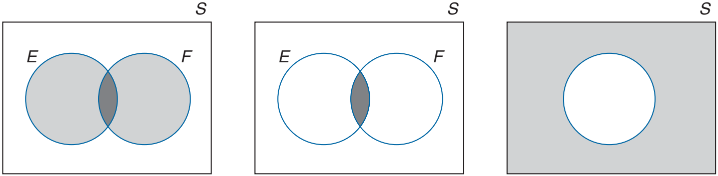
3. Suppose we are interested in determining the amount of dosage that must be given to a patient until that patient reacts positively. One possible sample space for this experiment is to let S consist of all the positive numbers. That is, let

   $$S = (0, \infty)$$

   where the outcome would be x if the patient reacts to a dosage of value x but not to any smaller dosage.

This file is meant for personal use by [illegible]@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

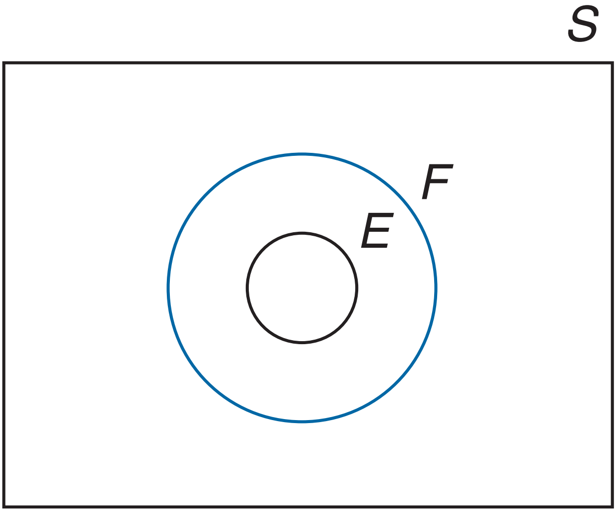
Venn Diagrams



(a) Shaded region: $E \cup F$ (b) Shaded region: EF (c) Shaded region: E^c

This file is meant for personal use by tm_7p4bnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Venn Diagrams



$E \subset F$

This file is meant for personal use by [illegible]@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Algebra of Events

Commutative Law	$E \cup F = F \cup E$	$EF = FE$
Associative Law	$(E \cup F) \cup G = E \cup (F \cup G)$	$(EF)G = E(FG)$
Distributive Law	$(E \cup F)G = EG \cup FG$	$EF \cup G = (E \cup G)(F \cup G)$

This file is meant for personal use by test_m7p4dmsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Algebra of Events

Commutative Law	$E \cup F = F \cup E$	$EF = FE$
Associative Law	$(E \cup F) \cup G = E \cup (F \cup G)$	$(EF)G = E(FG)$
Distributive Law	$(E \cup F)G = EG \cup FG$	$EF \cup G = (E \cup G)(F \cup G)$

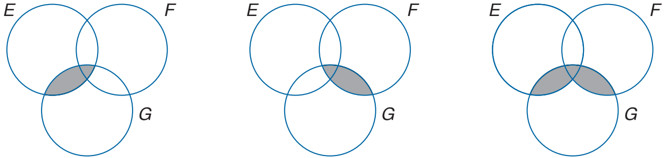
De Morgan's Laws

$$(E \cup F)^c = E^c F^c$$

$$(EF)^c = E^c \cup F^c$$

This file is meant for personal use by t_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proof of Distributive Law



(a) Shaded region: EG

(b) Shaded region: FG

(c) Shaded region: $(E \cup F)G$
$(E \cup F)G = EG \cup FG$

This file is meant for personal use by test_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Axioms of Probability

Axiom 1

$$0 \leq P(E) \leq 1$$

This file is meant for personal use by test_mp4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Axioms of Probability

Axiom 1

$$0 \leq P(E) \leq 1$$

Axiom 2

$$P(S) = 1$$

This file is meant for personal use by test_m7p4bdmx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Axioms of Probability

Axiom 1

$$0 \leq P(E) \leq 1$$

Axiom 2

$$P(S) = 1$$

Axiom 3 For any sequence of mutually exclusive events $E_1, E_2, \ldots$ (that is,events for which $E_i E_j = \emptyset$ when $i \neq j$),

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i), \qquad n = 1, 2, \ldots, \infty$$

We call $P(E)$ the probability of the event E.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Proposition 1

Since $E \cup E^c = S$

$$1 = P(S) = P(E \cup E^c) = P(E) + P(E^c)$$

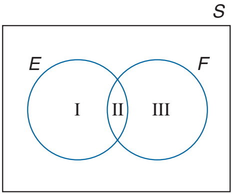
Equivalently

$$P(E^c) = 1 - P(E)$$

This file is meant for personal use by test_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proposition 2

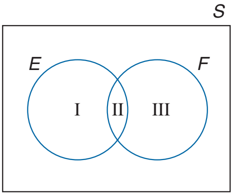
$$P(E \cup F) = P(E) + P(F) - P(EF)$$



This file is meant for personal use by test_m7p4dnsx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Proposition 2

$$P(E \cup F) = P(E) + P(F) - P(EF)$$



Proof

$$P(E \cup F) = P(I) + P(II) + P(III)$$

$$P(E) = P(I) + P(II)$$

$$P(F) = P(II) + P(III)$$

which shows that

$$P(E \cup F) = P(E) + P(F) - P(EF)$$

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Sample spaces with equally likely outcomes

For a large number of experiments, it is natural to assume that each point in the sample space is equally likely to occur. That is, for many experiments whose sample space S is a finite set, say $S = \{1, 2, \ldots, N\}$, it is often natural to assume that

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{N\}) = p$$

Now it follows from Axioms 2 and 3 that

$$1 = P(S) = P(\{1\}) + P(\{2\}) = + \cdots + P(\{N\}) = Np$$

which shows that

$$P(\{i\}) = p = \frac{1}{N}$$

from Axiom 3 it follows that for any event E,

$$P(E) = \frac{\text{Number of points in} E}{N}$$

In words, if we assume that each outcome of an experiment is equally likely to occur, then the probability of any event E equals the proportion of points in the sample space that are contained in E.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Generalized Basic Principle of Counting

If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes, and if for each of these n_1 possible outcomes there are n_2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments there are n_3 possible outcomes of the third experiment, and if, . . . , then there are a total of $n_1 \cdot n_2 \cdots n_r$ possible outcomes of the r experiments.

This file is meant for personal use by test_mp4admx6@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Generalized Basic Principle of Counting

If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes, and if for each of these n_1 possible outcomes there are n_2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments there are n_3 possible outcomes of the third experiment, and if, . . . , then there are a total of $n_1 \cdot n_2 \cdots n_r$ possible outcomes of the r experiments.

Notation

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Also since $0! = 1$

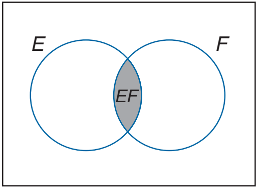
$$\binom{n}{0} = \binom{n}{n} = 1$$

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

Conditional Probability

Conditional probability of E given that F has occurred, and is denoted by

$$P(E|F)$$



A general formula for $P(E|F)$ that is valid for all events E and F. If the event F occurs, then in order for E to occur it is necessary that the actual occurrence be a point in both E and F; that is, it must be in EF. Now, since we know that F has occurred, it follows that F becomes our new (reduced) sample space and hence the probability that the event EF occurs will equal the probability of EF relative to the probability of F. That is,

$$P(E|F) = \frac{P(EF)}{P(F)}$$

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.

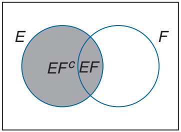
Bayes Formula

Let E and F be events. We may express E as

$$E = EF \cup EF^c$$

As EF and EF^c are clearly mutually exclusive, we have by Axiom 3 that

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)[1 - P(F)] \end{aligned}$$



Equation states that the probability of the event E is a weighted average of the conditional probability of E given that F has occurred and the conditional probability of E given that F has not occurred: Each conditional probability is given as much weight as the event it is conditioned on has of occurring.

This file is meant for personal use by [illegible] only.
Sharing or publishing the contents in part or full is liable for legal action.