Descriptive Statistics for STAT112

Charles Fleming

January 10, 2016

Although statistics owes its origins to the state, agricultural experimenters made it into a science in the early decades of the 20^{th} century. Modern statistics began with agriculture but ends with pharmaceuticals. Much statistical activity, today, is being sustained by very handsome salaries which are financed by the pharmaceutical industry.

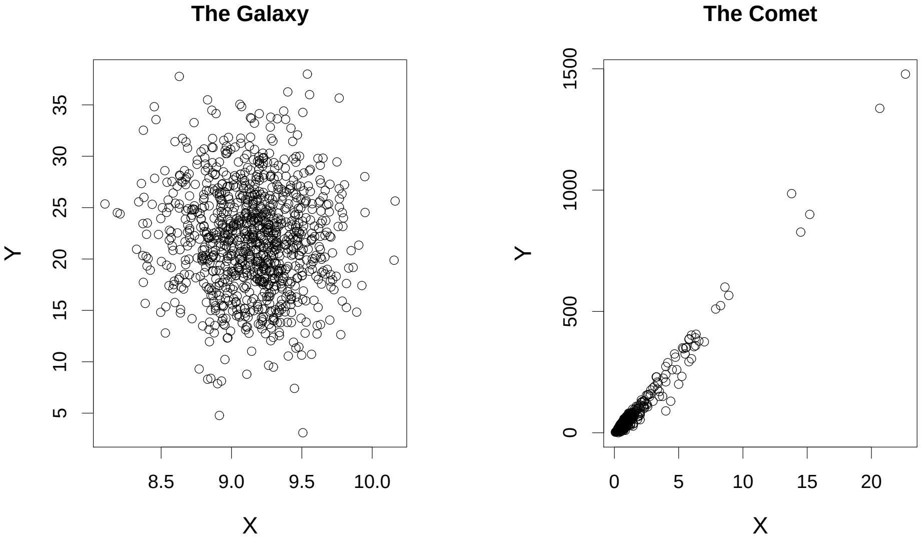
The earliest citation of the word statistics in the Oxford English Dictionary 2^{nd} edition is in 1770 in W. Hooper's translation of Bielfield's *Elementary Universal Education*, "The science, that is called statistics, teaches us what is the political arrangement of all the modern states of the known world." In Webster's dictionary of 1828, the definition of statistics is: "A collection of facts respecting the state of society, the condition of the people in a nation or country, their health, longevity, domestic economy, arts, property and political strength, the state of the country, &c." From this early connotation of statistics with the state, the meaning of statistics has been generalized by scientists to span any analysis of data from any experimental science.

For a student of modern statistics, making sense out of a mass of formulas and computer programs may seem like a daunting job ultimately studied in vain only for the sake of getting a passing grade. While for a statistician, the vast collection of formulas and promising computational software might entice him in vain to produce something out of nothing. For both the student and the professional, two guiding *rules of thumb* should be kept in mind when doing statistical work:

- If something is not obvious in the data, then probably there is nothing there.
- Statistics does not create information; it only clarifies it, like a computer enhancement of a fuzzy picture.

Even though these two principles seem to be self-evident, the first one is used to guard against someone trying to lie with statistics, and the second one expresses the concept of controlling variability in the data. They are very important concepts.

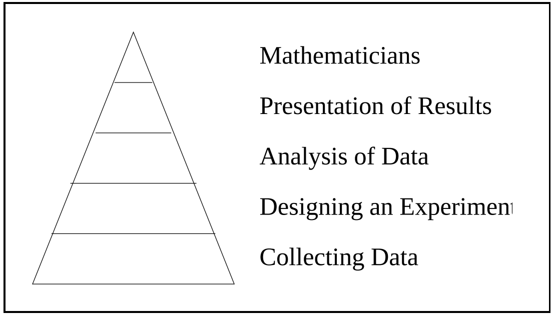
In spite of the sophisticated mathematics which underlies statistical methods, these two principles simply rely on the connotations of imagery. To draw a picture of the data is the first order of business for a statistician because a picture will tell him what he is up against. Look at the two plots of data shown below. It is obvious that there is a linear structure in the data on the right while the picture on the left being featureless will attract little attention for providing any useful information.



The utility of statistics lies in the use of theoretical benchmarks, in order to judge: "*How big is big? How little is little?*" as in determining whether the difference between the characteristics of sample A and those of sample B are significant or negligible. In other words, is the inference which we want to make about a set of experimental observations true? The answers to these questions lie at the heart of statistics. Their validity is determined only by the invocation of the important idea of confidence intervals. From the underlying concept of confidence intervals comes the method of substantiating a statistical inference and the objective of taking us to the culmination of this course.

Without question, the most difficult and most expensive undertaking of a statistical project is the collection of the data. The relative sizes of the divisions of labor for a statistical enterprise may be represented by a pyramid of resources. At the base of the pyramid is its foundation. It is massive. Everything depends on it. The collection of data is the foundation of the experimental sciences. The quality of the data must be the highest attainable; the procedure of obtaining the data must be carefully planned. Although it requires fewer personnel and less financial resources, the planning of an experiment requires shrewd utilization of limited financial resources to produce informative data. This requires a careful experimenter to undertake many preliminary tests to guarantee a successful experiment. The easy part of a statistical undertaking is the analysis of the data. It can be done by pressing a button provided that the computer programs are trouble free. The whole project is useless unless the results are published. This is the time when proficiency

Required Resources



in writing correct English which we learned in school and polished later by practice and by reading good literature bears dividends a hundred fold on our investment. At the top of the pyramid sits a little group of mathematicians and computer scientists who derive the mathematical formulas and write the computer programs.

Although the design of an experiment requires very careful planning and judicious allocation of resources, the process of acquiring the data is very difficult. The overhead which is associated with the collection of data forms the greatest part of the expense. The work force consisting of supervisors, enumerators, and clerical workers require salaries, benefits, training, supplies, and facilities. It suffers from attrition; personnel who make mistakes; computers which malfunction. The largest statistical organization in the United States is the Census Bureau. It employees thousands of people, and it is a multi-billion dollar operation. Whether the statistical undertaking requires a huge organization or it is a semester project, it must start with the basics.

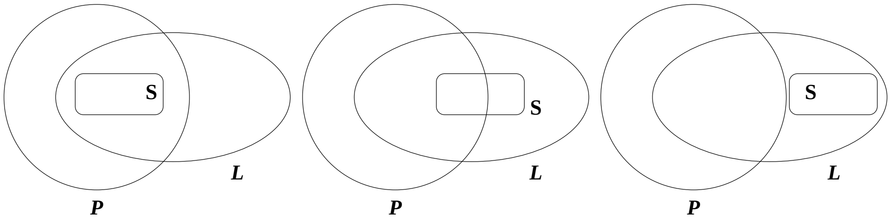
Definition 1.

Population. *A population is an arbitrary collection of things which contain the desired information. Its composition is determined by the experimenter, and it must be precisely defined in terms of substance and time.*

List. *A directory of the population is called a list, and it is used to identify the elements of the population.*

Sample. *A subset of the list is called a sample.*

The definitions of population, list, and sample may be depicted by anyone of these three schematic diagrams:



Usually, the sample contains a mixture of some elements of the population and elements outside the population. Ideally, all the elements of a sample should lie within the population. The worst sample will be the one which contains no elements of the population. A sample is always drawn from the list.

Example 1. *Examples of a Population*

> *All English letters in the textbook. The population is static.*
>
> *Everyone alive at this moment in Washington, D.C.. If the temporal aspect were not specified, the population would be constantly changing,*
>
> *Students present in class. This population is co-incidentally a subset of the previous one.*

As these examples illustrate, the concept of a population is defined according to the eyes of the beholder. Its composition is completely arbitrary, but it is defined in a way as to contain within it all the desired information. There is something about the population which is being sought; otherwise, there is no need to conduct an experiment. An economist who might want to estimate the number of men who have been involuntarily unemployed for more than six months will define one kind of population, while a sociologist who might want to learn the distribution of ages of mothers whose children are younger than five years old according to their disposable income will define a different population. A precise definition of a population is very important, so that someone else can clearly understand its scope. Once a population is defined, its elements have to be identified, in order to acquire the information.

Example 2. *Examples of a List*

> *The textbook can serve as a list. That is, a list and a population can be the same.*
>
> *A telephone directory.*
>
> *A class roster.*

In practice, a list is seldom very accurate because the maintenance of a list cannot keep up with a constantly changing population. A good example is the population of people present in Washington, D. C.. Choosing a telephone directory for the list of this population might be convenient, but it is notoriously inaccurate. Some residents have unlisted telephone numbers, others may have moved away but whose telephone number is still listed. Obviously, those who are commuting from work, being born, or dying make the population inherently dynamic; therefore, the population of people present in Washington, D. C. was defined with respect to time. Likewise, a list must exactly correspond with the population. The less complete and less accurate a list of the population is, the worse a sample will represent the population.

Example 3. *Examples of a Sample*

> *A page of a book.*
>
> *Everyone in Washington, D.C. who is in jail at this moment.*
>
> *Every student on the roster whose family name begins with a vowel.*

Usually, it is infeasible to study every element of a list. Instead, characteristics of a sampling of a list are studied and generalized to describe the population provided that the sample is representative of the population. Of course, no sample will be absolutely representative of a population because every list contains imperfections. If a sample is drawn at random from a good list of the population, then, for practical purposes, the sample can be assumed to represent the population.

Question 1. *Are the characteristics of prisoners who reside in the Washington, D. C. jail representative of the city's population at large?*

Clearly, a sample consisting of jailed men of Washington cannot contain sufficient information, if at all, to determine the total financial liabilities of women in the city. It is not enough to define a population precisely and to find an accurate list of it, drawing a sample of sufficient size which adequately represents the population can be difficult to do.

Some samples are unique.

Definition 2. *If $\mathcal{S} = \mathcal{L} = \mathcal{P}$, then $\mathcal{S}$ is called a* ***census****.*

We obtain information about the population by means of an experiment whereby a sample is drawn from the list. The sample provides us with information. In chemistry and physics, the population is the universe; the list is the observable universe. The sizes of both are infinite. But in business, social sciences, and surveys, the population is finite.

Finite brings connotations of counting. Counting things may seem like the easiest of all assignments. Is it not easy to count the number of fingers on a hand? But in general, counting is actually very difficult. How can someone accurately count small birds flying in large flocks through tree tops during the Spring migration? Counting is but one procedure which is utilized for collecting data. Once the set of data has been collected and cleaned of errors, the analysis of the data may commence.

1 Make a Picture

If the experiment was carefully designed, then the set of data which came from it should contain obvious features. To see the obvious, it is necessary to make a picture of the data. There are several standard techniques with which to exhibit the obvious, if the obvious is there.

2 Leaf and Stem Plot

A leaf and stem plot is perhaps the simplest technique to make a simple picture of the data quickly and inexpensively. Given this set of scores of an examination:

$$\mathcal{S} = \left\{80,\ 40,\ 78,\ 93,\ 85,\ 82,\ 87,\ 80,\ 71,\ 99,\ 85,\ 86,\ 80,\ 91,\ 90,\ 83,\ 75\right\}$$

split each number into two. The set of first digits are listed vertically in one column and the set of second digits will be listed in ascending order in that row which corresponds to the first digits. For example,

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

Gaps in the data are included, in order to give a better sense of the distribution of the data. This leaf and stem plot suggests at a glance that the frequencies of 70's, 80's, and 90's in the set of data are not uniform. Rather, there is a preponderance of 80's in the midst of few low and high scores. In response to the obvious structure revealed by the leaf and stem plot, we are lead to believe that there is something about the data worth reporting.

3 Descriptive Statistics

The concept of descriptive statistics is taken from physics. The physicists, in order to explain the motion of a large system of particles, describe the system in terms of its center of mass and moment of inertia. Before the invention of computers, short cuts were needed to communicate the essence of large sets of data to others because making a picture of the data by lithographers for publication was a very expensive and slow process. Ideas of center of mass and moment of inertia were employed to reduce a large set of data to two numbers. This practice is still used today even when cheap computational resources make it very easy to describe data by means of a picture. Although information is lost by describing data in terms of descriptive statistics, they are welcomed bits of information with which to embellish a picture.

Table 1

	Data Set I				Data Set II			
	Value	Mean	Deviation	Square of Deviation	Value	Mean	Deviation	Square of Deviation
1	1	5	-4	16	5.2	5	.2	.04
2	2	5	-3	9	5.3	5	.3	.09
3	3	5	-2	4	4.8	5	-.2	.04
4	4	5	-1	1	4.9	5	-.1	.01
5	5	5	0	0	5.0	5	.0	.00
6	6	5	1	1	5.2	5	.2	.04
7	7	5	2	4	4.8	5	-.2	.04
8	8	5	3	9	4.7	5	-.3	.09
9	9	5	4	16	5.1	5	.1	.01
sum	45		0	60	45		0	.36
Standard Deviation				$\sqrt{\frac{60}{9}} = \sqrt{6.667} = 2.582$	Standard Deviation			$\sqrt{\frac{.36}{9}} = \sqrt{.04} = .2$

The meaning of center of mass is easy to understand. It is that point or number where all the data can be concentrated. In statistics, the center of mass is called the mean. If someone were to place the mean on top of his finger, the set of data will balance perfectly. The moment of inertia or the variance as it is called in statistics is a measure of the dispersion of the data about the mean. To illustrate these two important descriptive statistics, consider the tabulation shown in Table 1 of two sets of different data which coincidentally have the same center of mass, 5, but the sets of data are dispersed differently about the mean. Although the two sets of data have the same center of mass, further review of the table agrees with our intuition that the second set of data must be better because its values do not deviate from the mean as much as the values of the first set. That our intuition tells us something should not be dismissed. In fact, we will rely on our intuition to guide us throughout the course of studying statistics. In order to express the sense of our intuition analytically, we will use a mathematical description of the dispersion of the data about the mean.

The difference between a value and the mean is the deviation of the value from the mean. The larger the deviations, the more dispersed the values are from the mean. Taking the sum of the deviations, therefore, should be a good measure of the dispersion of the data from the mean; however, as can be seen in the table, the sums of the deviations are equal to zero. In fact, the sum of the deviations will always be equal to zero, because the negative deviations exactly cancel the positive deviations. To get rid of the negative signs, the deviations are squared and the sum of squared deviations is divided by the number of observations. This quotient is called the moment of inertia by physicists. Statisticians call it the variance. There are two kinds of variances depending on whether the set of data completely corresponds to every element of a population or the set of data came from a sampling of the population. The tabulated data shown in Table 1 may represent all the values for a population or it may represent the values of a sample. It is impossible to tell without more information.

Definition 3. *Mean*

1. *The average of all values which are associated with a population, $\mathcal{P}$, is called the* ***population mean*** *and is denoted by μ.*

2. *The average of all values coming from a sample, $\mathcal{S}$, is called the* ***sample mean*** *and is denoted by $\bar{x}$.*

Definition 4. *Variance*

1. *The average of all squared deviations of the values associated with a population from the population mean, μ, is called the* ***population variance*** *and is denoted by σ^2.*

2. *The "average" of all squared deviations of values of a sample from the sample mean, $\bar{x}$, is called the* ***sample variance*** *and is denoted by s^2.*

Unlike English which is a natural language and has inherent limitations which are likely to be met when trying to explain abstract ideas, the language of mathematics uses symbols which permit us to express ideas very succinctly and thereby make them much more easily comprehensible. Stating the definitions of mean and variance again but mathematically will illustrate the immense utility of the mathematical language. Take note of the following conventions which are employed in the formulas:

1. $i \in \mathcal{P}$ means *i in* the population.

2. $i \in \mathcal{S}$ means *i in* the sample.

3. N is the size of $\mathcal{P}$

4. n is the size of $\mathcal{S}$

	Mean	Variance	Standard Deviation
Population	$\mu = \frac{\sum_{i \in \mathcal{P}} x_i}{N}$	$\sigma^2 = \frac{\sum_{i \in \mathcal{P}} (x_i - \mu)^2}{N}$	$\sqrt{\sigma^2} = \sigma$
Sample	$\bar{x} = \frac{\sum_{i \in \mathcal{S}} x_i}{n}$	$s^2 = \frac{\sum_{i \in \mathcal{S}} (x_i - \bar{x})^2}{n-1}$	$\sqrt{s^2} = s$

5. **Be careful** about the denominator of the sample variance.

To demonstrate the utility of mathematical symbols and syntax in developing and explaining mathematical concepts, notice in the following example how much more succinct and clear the mathematical statement is compared to its English equivalent.

Example 4. *Write in mathematical symbols: The quotient of the difference between the sum of the square of the observed value x_i over all elements in $\mathcal{S}$ and the product of n times the square of the sample mean by the size of $\mathcal{S}$. Answer:* $\frac{\sum_{i \in \mathcal{S}} x_i^2 - n\bar{x}^2}{n}$

A glance at a mathematical treatise which was written before the 18^{th} century probably almost completely in prose usually in Latin rather than in a modern language should convince anyone of the revolutionary improvement in communicating complex abstract ideas that the lexicon of symbols have brought to mathematics. The prospect of reading, much less trying to comprehend these treatises, is discouraging enough that we depend on the historian of mathematics to translate them into modern mathematics.

4 Greek Alphabet

Because of the profound Greek legacy which is so conspicuous in mathematics, and for that matter, in all fields of abstract reasoning, Greek letters and derivatives of Greek words are commonly found in English and in mathematics. Ancient Egyptian and Babylonian scholars knew some empirical mathematics, that is, they knew rules of thumb, such as the observation that the 3-4-5 triangle is a right triangle, but they did not know why the rules were valid. The Greeks proved that the rules of thumb were indeed true by using deductive reasoning from self-evident principles. The Greeks' profound legacy in mathematics and science lies in their insistence of using reason or what we call deductive logic to explain the natural causation of phenomena. As important as that inheritance may seem to be for the advancement of science and of the maintenance of modern society, scholars of antiquity will argue that the greatest and purely original contribution of the Greeks to western civilization is not the institution of democracy or philosophy or mathematics or medicine or science, but the tragedy, because through tragedy they could advance the principle that the administration of justice applies to everyone both mortal and divine. In practical matters for us, Greek letters appear so often in our study of statistics that the Greek alphabet is shown below and ought to be learned.

No

Known

Portrait

Archimedes
(Ἀρχιμήδης)
287-212 B.C.

Archimedes Being Accosted by
a Roman Soldier

Table 2: Greek Alphabet

ἄλφα	A	α	alpha	ἰῶτα	I	ι	iota	ῥῶ	P	ρ	rho
βῆτα	B	β	beta	κάππα	K	ϰ	kappa	σίγμα	Σ	σ ς	sigma
γάμμα	Γ	γ	gamma	λάμβδα	Λ	λ	lambda	ταῦ	T	τ	tau
δέλτα	Δ	δ	delta	μῦ	M	μ	mu	ὖ ψῑλόν	Υ	υ	upsilon
ἒ ψῑλόν	E	ε	epsilon	νῦ	N	ν	nu	φῖ	Φ	φ	phi
ζῆτα	Z	ζ	zeta	ξῖ	Ξ	ξ	ksi	χῖ	X	χ	chi
ἦτα	H	η	eta	ὂ μῑκρόν	O	ο	omicron	Ψῖ	Ψ	ψ	psi
ϑῆτα	Θ	ϑ	theta	πῖ	Π	π	pi	ὦ μέγα	Ω	ω	omega

How Classical Greek was pronounced is unknown. If the best modern scholar of Greek were to be transported back into time and placed in the Agora at lunchtime in downtown Athens in 450 B.C., he would not be able to understand or be understood by anyone. The same difficulty which our scholar will experience is the same one that would probably be experienced today, if someone from rural Louisiana and rural China were to converse in English. Both may speak English, but their strong accents will probably prevent them from understanding each other.

Over the centuries, the Greek language has changed. There is Homeric Greek (700 B.C.), Classical Greek (450 B.C.), Koine Greek (1 A.D.), Byzantine or Medieval Greek, and Modern Greek. The grammar and spellings have become simpler and the pronunciation less aspirated and palatalized; that is, it sounds progressively softer from era to era. Some teachers of Classical Greek use Modern Greek pronunciation instead of the hypothesized classical pronunciation since, they argue, no one knows how Classical Greek was spoken anyway; therefore, who should care except those in small academic circles. We will use the Classical Greek pronunciation as developed by the 16th century Dutch scholar Desiderius Erasmus and modified slightly only recently.

The way that the Greeks wrote numbers must surely have inhibited them from developing numerical methods for doing mathematical computations. A tabulation of the Ancient Greek numerals from 0 to 99 is shown in Table 3

Table 3: Ancient Greek Numerals

0	0	1	2	3	4	5	6	7	8	9
0		α′	β′	γ′	δ′	ε′	ϛ′	ζ′	η′	ϑ′
1	ι′	ια′	ιβ′	ιγ′	ιδ′	ιε′	ιϛ′	ιζ′	ιη′	ιϑ′
2	ϰ′	ϰα′	ϰβ′	ϰγ′	ϰδ′	ϰε′	ϰϛ′	ϰζ′	ϰη′	ϰϑ′
3	λ′	λα′	λβ′	λγ′	λδ′	λε′	λϛ′	λζ′	λη′	λϑ′
4	μ′	μα′	μβ′	μγ′	μδ′	με′	μϛ′	μζ′	μη′	μϑ′
5	ν′	να′	νβ′	νγ′	νδ′	νε′	νϛ′	νζ′	νη′	νϑ′
6	ξ′	ξα′	ξβ′	ξγ′	ξδ′	ξε′	ξϛ′	ξζ′	ξη′	ξϑ′
7	ο′	οα′	οβ′	ογ′	οδ′	οε′	οϛ′	οζ′	οη′	οϑ′
8	π′	πα′	πβ′	πγ′	πδ′	πε′	πϛ′	πζ′	πη′	πϑ′
9	ϙ′	ϙα′	ϙβ′	ϙγ′	ϙδ′	ϙε′	ϙϛ′	ϙζ′	ϙη′	ϙϑ′

Although there is a pattern in the Greek numerals, computational mathematics did not develop until after the invention of Arabic numerals which we use today. Numerical analysis which encompasses computational mathematics began as a science upon the momentous developments of calculus made by Newton and Leibniz in the 17th century. By means of this new science of numerical analysis, efficient and accurate methods were derived for constructing mathematical tables which came to be closely guarded commercial secrets particularly in Great Britain. Supposedly, the accuracy and completeness of navigational tables led to the maritime superiority of Great Britain, so that British ships could ply the oceans more quickly than ships from other countries as is cited, for example, about a British ship which could cross the Pacific Ocean two weeks faster than a competing nation's ship.

In the tabulation of Greek numerals there is a conspicuous blank spot for 0. The Greeks did not have a numeral for zero nor did any other civilization. Such a simple idea of a numeral zero evaded the greatest mathematicians of antiquity, and the world had to wait for the Hindi of India to invent this indispensable numeral. It can be seen in the following table of the multiples of 100 that the decimal system with placeholders also evaded the Greeks.

100	200	300	400	500	600	700	800	900	1000
ρ′	σ′	τ′	υ′	φ′	χ′	ψ′	ω′	ϡ′	͵α

Without the advantages which Arabic numerals and the decimal system provide and without the mathematical paradigms given by the Greeks, statistics could not have developed into a practical science. A tremendous renaissance occurred in statistics with the invention and subsequent availability of inexpensive computational resources of the digital computer and hand calculator. Computational mathematics including statistics is currently enjoying a golden age. A popular topic of contemporary statistical research is, in fact, statistical computing.

5 Other Descriptive Statistics

There are other descriptive statistics besides the mean and variance which are useful in communicating the essence of a set of data. For example,

Definition 5.

1. *The **range** is the difference between the maximum and minimum values.*

2. *The **median** divides an ordered set of data into two parts, each part having the same number of elements*

3. *The **quartiles** divide an ordered set of data into four equal parts. There are several methods for calculating the quartiles. We will use the method proposed by John Tuckey found in Understanding Robust and Exploratory Data Analysis written by David Hoaglin, Frederick Mosteller, and John Tukey.*

   (a) *If n is odd then the 1^{st} quartile divides the first half into two where the median is included. Algebraically, the 1^{st} quartile $= \frac{n+3}{4}$*

   (b) *If n is even then the 1^{st} quartile divides the first half into two where the median is excluded. Algebraically, the 1^{st} quartile $= \frac{n+2}{4}$*

   (c) *Likewise for the 3^{rd} quartile; if n is odd then the 3^{rd} quartile $= \frac{3n+1}{4}$, and when n is even, the 3^{rd} quartile $= \frac{3n+2}{4}$*

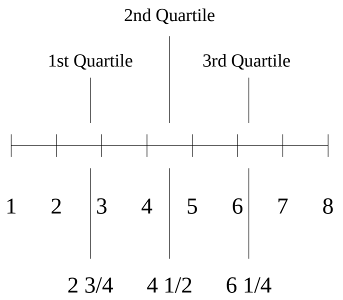


Figure 1

4. *A* ***percentile*** *is that number, x_p, such that at least p% of the data is less than x_p and at most (1-p)% of the data is greater than x_p.*

Care must be taken when reading a publication involving statistics or when using a computer program because the definition of the quartile or the percentile might differ from the one given here. For example, in the statistical software package known as R the quartiles are defined in terms of interpolated distance, that is, the 1^{st} quartile is equal to $\frac{1}{4}^{th}$ of the range from the minimum value, the 3^{rd} quartile is $\frac{3}{4}^{ths}$ of the range from the minimum value; the median is $\frac{1}{2}$ the distance. For instance, given $\{1 \ \ 2 \ \ 3 \ \ 4 \ \ 5 \ \ 6 \ \ 7 \ \ 8\}$, then as shown Figure 1, $q_1 = 2\frac{3}{4}$, $q_2 = 4\frac{1}{2}$, and $q_3 = 6\frac{1}{4}$ instead of $2\frac{1}{2}$, $4\frac{1}{2}$,and $6\frac{1}{2}$, respectively, which we would have reported. We will use our definition of quartiles because the mid-points are easier to find than interpolating the range appropriately for the quartiles. While descriptive statistics are used for describing a set of data for the benefit of a reader, they are not very useful in substantiating a statistical conclusion. Instead, a plot of the data, the theoretical derivation of the statistical model, and a thorough analysis of the data constitute a substantial defense or attack on a statistical conclusion. In practical matters, then, the differences over the definitions of the sample quartiles and percentiles are not that important to worry about.

Example 5. *For the set of data $\mathcal{S} = \{4.7 \ \ 4.8 \ \ 4.8 \ \ 4.9 \ \ 5.0 \ \ 5.1 \ \ 5.2 \ \ 5.2 \ \ 5.3\}$.*

1. *Arrange from smallest to largest.*

2. *The median is 5.*

3. *n=9 is odd; therefore, 1^{st} quartile=4.8 and 3^{rd} quartile=5.2*

Example 6. *Suppose the set of data consists of $\mathcal{S} = \{1 \ \ 2 \ \ 3 \ \ 4 \ \ 5 \ \ 6 \ \ 7 \ \ 8\}$.*

1. *The median is invented by the convention of taking the mid-point of 4 and 5; therefore, the median is 4.5.*

2. *Likewise, we will take the mid-point for the quartiles. Because n=8 is even, the median is excluded, so that the* 1^{st} *quartile=2.5 and the* 3^{rd} *quartile=6.5.*

Table 4: The Number of Heads from Flipping Three Coins 70 Times

0	1	1	1	1	1	2	2	3	1	1	0	0	2
1	1	1	2	1	2	3	2	1	1	1	1	0	0
3	1	1	1	2	2	3	1	0	1	0	0	1	1
2	2	2	1	0	1	2	1	2	0	0	2	1	1
2	1	1	2	1	2	1	1	2	1	2	2	2	2

From the ever popular experiment of flipping coins, many useful examples may be drawn for illustrating simple concepts in statistics and probability. The results of obtaining the number of heads of one such experiment of flipping three coins 70 times are tabulated in Table 4, and the calculations for the sample mean and sample variance are shown in Example 7.

Example 7.

1. $\bar{x} = \frac{11\times0+33\times1+22\times2+4\times3}{70} = \frac{89}{70} = 1.271428$

2. $s^2 = \frac{11(0-\frac{89}{70})^2+33(1-\frac{89}{70})^2+22(2-\frac{89}{70})^2+4(3-\frac{89}{70})^2}{70-1} = \frac{1023}{1610} = .635403$

3. *s=.797121*

Although the table of data in this example is small enough to be easily displayed, publishing a much larger table might not be practical, but its essence can still be presented through the use of descriptive statistics. Of course, presenting both a tabulation together with a set of descriptive statistics provides more information than either alone. The best presentation of the data consists of a tabulation, descriptive statistics, and a picture.

6 Histograms

Leaf and stem plots are sometimes used especially when a quick and rough graphical display of the data is needed. Pie charts like the one shown in Figure 2 are commonly employed whenever there is a need to illustrate the relative proportion of each of several components comprising a set of data, but histograms command the greatest popularity. Since the purpose of making a picture of the data lies in the goal of making outstanding features of the data easily perceptible, the artistic talents, that discerning eye, those creative instincts

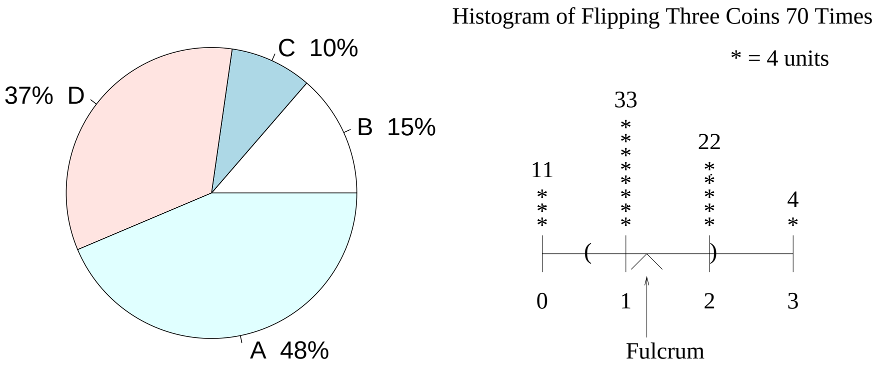


Figure 2

of the author determine the form of a good histogram. There are many kinds of histograms and some of them are given names like butterfly charts, frequency charts, and bar charts. There is no theoretically correct form of a histogram, but there are some general rules:

- Use enough bins to make the histogram look nice.
- Embellish the histogram with a title and legend.
- Use artistic skill when making a histogram.

Law 1 (Nolan's Placebo). *An ounce of image is worth a pound of performance.*

Always bear in mind that nothing apparent may exist in the information until a picture of the data is drawn.

The histogram shown in Figure 2 portrays the frequency of getting heads from the experiment of flipping three coins. There are 11 0's, 33 1's, 22 2's and 4 3's in the data. To make the histogram more compact, an asterisk represents four units and it is noted in a legend. The number of occurrences appears at the top of each column to assist the reader in associating the histogram with the tabulation of outcomes. A fulcrum was drawn on the histogram only for the purpose of illustrating the concept of the center of mass or what statisticians call the sample mean. At that point, the histogram will balance. All the weight of the data can be concentrated at that point. The parentheses mark the distance of one standard deviation on either side of the mean, and they convey a sense of the dispersion of the data about the mean. If the asterisks were to be removed from the histogram, of the pieces that are left, the fulcrum and parentheses represent the mental images that one should see when studying the two descriptive statistics, the mean and the variance. Upon reading the mean and the variance, they should present in the mind's eye an image of the center of mass and dispersion of the data about the mean.

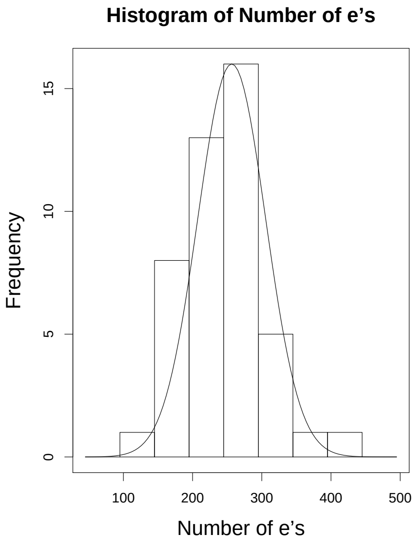


Figure 3

The histogram in Figure 2 is sufficient to tell us that there is some structure to the data. At a glance, it is apparent that the distribution is not symmetrical; rather, it is skewed to the left in favor of the 0's and the 1's. There is evidently some property of the data which warrants investigation which otherwise would be missed, if we had only been presented with a set of descriptive statistics. The utility of pictures in portraying data cannot be emphasized too much, and its importance is recognized by the substantial attention it receives in contemporary statistical research.

After reviewing many histograms, certain traits constantly reappear like a bell-shaped appearance similar to the one shown in Figure 3. It is not a mere coincidence that the two descriptive statistics, the sample mean and the sample variance of a set of data also characterize the location and spread of a bell shaped histogram. Later, we will discover that the bell shaped curves which are curves of the Normal probability distribution are defined by the mean and the variance. It can be proven that as the number of experimental observations increase to infinity the histogram of the sample z-scores will converge to that of the Normal distribution with mean 0 and variance 1. The shape of these bell shaped curves reflect the basic properties of the data. Some of them are narrow and centered while others are flat or some are off-centered suggesting to us concepts of accuracy and precision.

Imagine a target at which a statistician is aiming his hopes during the planning of a carefully designed experiment from which the answer of a problem will be obtained and pinned to the target. If the experiment is well conceived and smartly performed, the set of data produced by it will hit the bull's-eye. Miscalculations, mistakes, interference, bad planning, and the insidious complications perpetrated by uncontrollable circumstances afflict the best of experimental designs, so that in general the elements of the data are typically scattered over the target. That set of data which is tightly grouped in the bull's-eye is said to have high accuracy and high precision or equivalently have low bias and high precision. Three other cases can happen as depicted in Figure 4. The set of data can have high precision but is off the mark in which case it is said to have high bias. It is possible that the data is widely scattered but its center of mass is in the bull's-eye. That data will have low bias and low precision. Finally, the data can be bad on both counts; it will be biased with low precision. The best set of data is the one with no bias and high precision. The others are unwanted and unacceptable. Naturally, the more expensive the experiment, the more stress the statistician experiences because he might not get a second chance to do the experiment again. Therefore, statistical techniques have been created to be used in the analysis of the data for the purpose of salvaging as much information from imperfect experimental data as possible with the hope that enough good information can be gleaned from the data to defend an answer with a good degree of confidence.

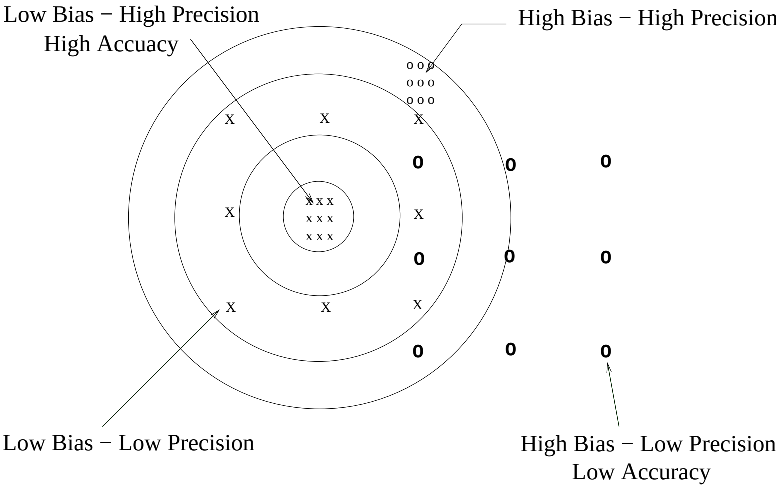

Low Bias − High Precision
High Accuacy
High Bias − High Precision
Low Bias − Low Precision
High Bias − Low Precision
Low Accuracy

Figure 4

7 Box Plot

John Tukey
1915-2000

The box plot was invented by John Tukey in the 1970's. This clever invention for displaying important features of data in a simple picture finds application in industrial quality control or similar production environments, because it is easy to make and easy to interpret. Statistical quality control has proven to be indispensable in improving the efficiency not only of manufacturing processes but every conceivable kind of operation. Box plots provide a quick way to evaluate data for discovering outliers which could mean that a malfunction is occurring and for revealing trends which could mean the process might not be conforming to specifications. The ends of the box are defined by the 1^{st} and 3^{rd} quartiles. The height of the box is immaterial, but its length implies that the box contains 50 percent of the data. The median is marked by a + symbol. Extending from either side of the box are whiskers. They are attached to the box at points called hinges. Their lengths are determined by the interquartile range (IQR).

Definition 6.

1. *The Inter Quartile Range, **IQR**, is the difference between the 3^{rd} quartile and the 1^{st} quartile.*

2. *Right **Inner Fence**=1.5*IQR+3^{rd} quartile*
   *Left **Inner Fence**=-1.5*IQR+1^{st} quartile*

3. *Right **Outer Fence** =3*IQR+3^{rd} quartile*
   *Left **Outer Fence** =-3*IQR+1^{st} quartile*

Once the IQR has been calculated, the whiskers are drawn from the hinges to a length of 1.5xIQR units as a solid line. Likewise the outer fences are drawn from the hinges

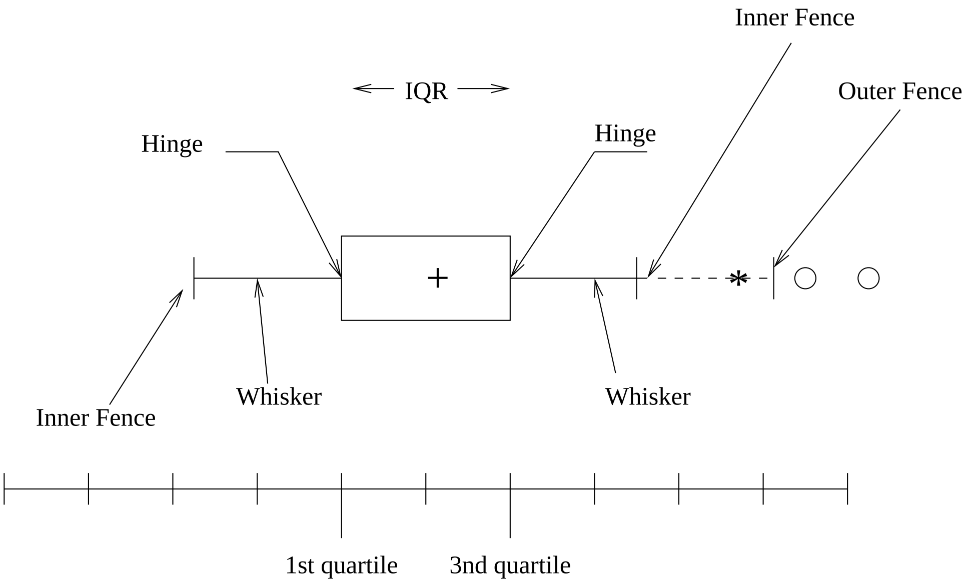


Figure 5

3xIQR units. That portion of the whisker that extends beyond the inner fences is drawn as a dashed line. Any points which lie between the inner and outer fences are denoted by asterisks. The points lying beyond the outer fences are denoted by small circles. The asterisks reveal points which are suspect for being bad; points denoted by circles are very suspect and require investigation into their origins.

When constructing a box plot, there is the stipulation that the fences never go beyond the furtherest data point. The right outer fence shown in Figure 5 extends to its fullest length, but the left whisker is cut short because it is stopped by the smallest data point.

A box plot shows basics statistics at a glance, for instance:

- 50% of the data lies in the box.
- Range corresponds to the extremities of the box plot.
- Points lying beyond the inner fence are problematic.
- Points lying beyond the outer fence are very problematic.
- Box plots are useful when a quick interpretation of the data is desired routinely or when comparing multiple data sets.

To recapitulate the steps for processing a set of data, consider the following example in which we identify the population, list, and sample, compare the various pictures of the data with one another, and associate the descriptive statistics with features of the pictures and with a tabulation of the data.

Example 8. *In the parking lot across the street, there are 20 automobiles. The parking lot attendant recorded the license plate number of each car and submitted the list an hour ago.*

Table 5: Blue Book Value of Twenty Automobiles

Car	Value	Car	Value	Car	Value	Car	Value
1	1	6	8	11	7	16	16
2	6	7	9	12	1	17	10
3	10	8	2	13	5	18	8
4	5	9	5	14	15	19	6
5	4	10	30	15	50	20	3

Should one expect that the list is still correct now?

On the same list, the attendant recorded beside each license plate number the blue book value of the car in thousands of dollars.

Therefore $\mathcal{P} = \{$*cars in parking lot*$\}$ *and* $\mathcal{L}$*=record from the attendant.*

The first order of business is to make a picture of the data like the accompanying leaf and stem plot of the data in tens of thousands of dollars.

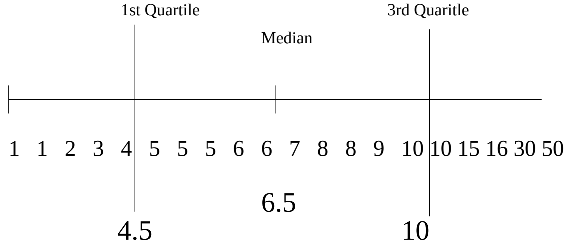
0	1 1 2 3 4 5 5 5 6 6 7 8 8 9
1	0 0 5 6
2	
3	0
4	
5	0

We see immediately that most of the cars have a value under \$10,000.

Let us calculate the usual population descriptive statistics.

- $\mu = \frac{1+6+10+5+\cdots+3}{20} = \frac{201}{20} = 10.05$
- $\sigma^2 = \frac{(1-10.05)^2+(6-10.05)^2+(10-10.05)^2+\cdots+(3-10.05)^2}{20} = \frac{2496.95}{20} = 124.8475$
- $\sigma = 11.17$

The first step in finding the median and the quartiles is to arrange the data in ascending order as in the following:



- 1^{st} quartile=4.5, 2^{nd} quartile=6.5, and 3^{rd} quartile=10

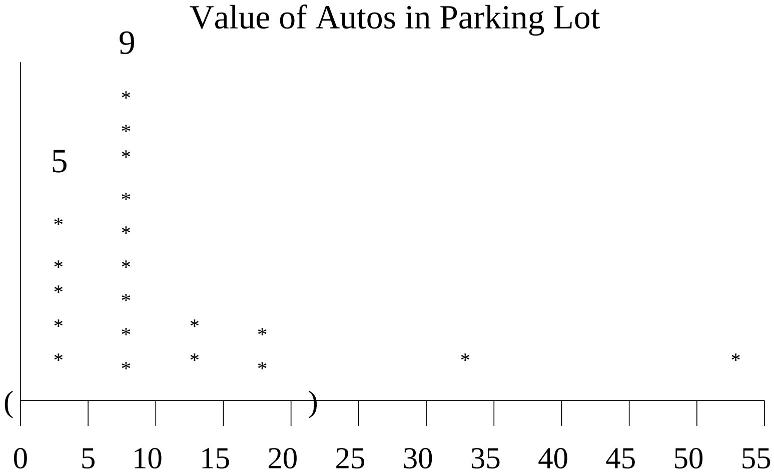


Figure 6

A histogram of the data like the one shown in Figure 6 is more informative than the leaf and stem plot and we see again that most of the values of the automobiles lie under \$10,000, but the bell shaped outline of the histogram is not symmetric but is skewed, and there appears to be a couple of unusually expensive cars in the lot.

Due to our suspicions about the presence of outliers, our attention turns to the box plot of the data as shown Figure 7. Again, the bulk of the data lies under \$10,000, and now the two most expensive cars command our attention because they lie beyond the right outer fence. Our curiosity leads us to wonder if the owners of them are students or perhaps wealthy alumni.

Rather than record all the cars, suppose the attendant instead recorded the value of the first five cars that entered the parking lot. That is, the attendant took a sample, $\mathcal{S} = \{1\ 6\ 10\ 5\ 4\ \}$.

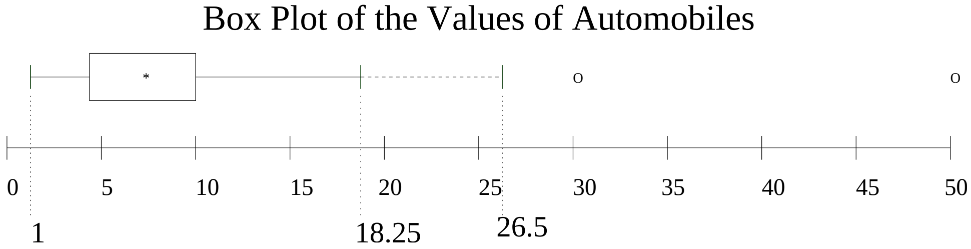


Figure 7

- $\bar{x} = \frac{1+6+10+5+4}{5} = \frac{26}{5} = 5.2$
- $s^2 = \frac{(1-5.2)^2+(6-5.2)^2+(10-5.2)^2+(5-5.2)^2+(4-5.2)^2}{5-1} = \frac{42.8}{4} = 10.7$
- s=3.27
- sample median = 5
- 1^{st} quartile=4, 2^{nd} quartile=5, and 3^{rd} quartile=6.

Juxtaposing the population descriptive statistics with the sample descriptive statistics in one table brings to light an interesting difference between the statistics of the population and of the sample. On the one hand, the sample quartiles do not deviate much from the population quartiles. But, on the other hand, the sample mean and sample variance are very much different than that of their population counterparts. We discover by this example that the sample mean and sample variance behave much differently than the quartiles because quartiles are resilient to the effects of outliers while the mean and variance are very sensitive to the presence of outliers. For that reason, quartiles especially the median are usually given side-by-side with the mean and variance in a statistical report.

	Mean	Variance	Standard Deviation	1^{st} Quartile	Median	3^{rd} Quartile
Population	$\mu = 10.05$	$\sigma^2 = 124.8475$	$\sigma = 11.17$	4.5	6.5	10
Sample	$\bar{x} = 5.2$	$s^2 = 10.7$	$s = 3.27$	4	5	6

8 Correlation Matrix

The correlation matrix comes from the variance-covariance matrix. Recall that the sample variance is: $s^2 = \frac{\sum\limits_{i \in \mathcal{S}} (x_i - \bar{x})^2}{n-1}$ the numerator of which can be written as: $\sum\limits_{i \in \mathcal{S}} (x_i - \bar{x})(x_i - \bar{x})$. It is a sum of squares. This idea of sum of squares can be generalized, for example, to $SS_{xy} = \sum\limits_{i \in \mathcal{S}} (x_i - \bar{x})(y_i - \bar{y})$. We see that with the generalized notation $SS_{xx} = \sum\limits_{i \in \mathcal{S}} (x_i - \bar{x})(x_i - \bar{x})$. If there is a third variate like z, then $SS_{yz} = \sum\limits_{i \in \mathcal{S}} (y_i - \bar{y})(z_i - \bar{z})$, and so on. These sums of squares are consolidated into a compact form by using the notation of matrices as in:

$$\Sigma = \frac{1}{n-1} \begin{bmatrix} SS_{xx} & SS_{xy} & SS_{xz} \\ SS_{yx} & SS_{yy} & SS_{yz} \\ SS_{zx} & SS_{zy} & SS_{zz} \end{bmatrix}$$

A numerical example of a matrix containing variances and covariances is the following:

$$\Sigma = \begin{bmatrix} 2.6567993 & 0.85522664 & -0.13320999 \\ 0.8552266 & 0.95097132 & -0.03052729 \\ -0.1332100 & -0.03052729 & 1.37557817 \end{bmatrix}$$

This matrix is called the variance-covariance matrix, and it is customarily denoted by Σ. A variation of it is the correlation matrix. Both contain the same information, but the correlation matrix makes it easier to relate variates with one another.

An element of the correlation matrix has the form: $\rho_{xy} = \frac{1}{n-1} \frac{SS_{xy}}{\sqrt{SS_{xx}}\sqrt{SS_{yy}}}$. The corresponding correlation matrix of Σ which was shown above is:

$$\begin{bmatrix} 1.00000000 & 0.53804443 & -0.06968107 \\ 0.53804443 & 1.00000000 & -0.02669082 \\ -0.06968107 & -0.02669082 & 1.00000000 \end{bmatrix} = \begin{bmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{yx} & 1 & \rho_{yz} \\ \rho_{zx} & \rho_{zy} & 1 \end{bmatrix}$$

By inspecting the correlation matrix, we see that the first and second variates have a correlation of 0.53804443 while the second and third variates have a correlation of -0.02669082. The diagonal elements are always 1 in the correlation matrix and that can be easily proven. Both the variance-covariance matrix and the correlation matrix are symmetric matrices which mean that the upper triangular array is the mirror image of the lower triangular array.

We do not make conclusions based on descriptive statistics rather they are tools for describing data and discovering features of the data. Even when there is a high correlation between two variable, correlation does not imply causation. Another example of a correlation matrix is the following one which came from responses to a survey.

The correlation matrix is one of the important things to examine when looking at data. It is the starting point for the study of principal components and factor analysis. The benefit of the correlation matrix is that it is simple, and it is shows at a glance how

Table 6: **Correlation Matrix**

	income	edu	job	life	values	valmoney	valfrien	valcaree	valrelig	depend
income	1.000	0.538	-0.069	-0.243	0.095	0.090	-0.343	0.011	0.044	-0.258
edu	0.538	1.000	-0.026	-0.100	-0.018	-0.019	-0.062	0.197	-0.060	-0.219
job	-0.069	-0.026	1.000	0.262	-0.083	0.007	-0.210	0.005	0.331	-0.054
life	-0.243	-0.100	0.262	1.000	0.090	-0.013	-0.041	-0.176	0.116	0.283
values	0.095	-0.018	-0.083	0.090	1.000	0.073	-0.098	-0.339	-0.444	-0.064
valmoney	0.090	-0.019	0.007	-0.013	0.073	1.000	-0.238	-0.344	-0.524	-0.135
valfrien	-0.343	-0.062	-0.210	-0.041	-0.098	-0.238	1.000	-0.262	-0.279	-0.052
valcaree	0.011	0.197	0.005	-0.176	-0.339	-0.344	-0.262	1.000	0.076	0.001
valrelig	0.044	-0.060	0.331	0.116	-0.444	-0.524	-0.279	0.076	1.000	0.124
depend	-0.258	-0.219	-0.054	0.283	-0.064	-0.135	-0.052	0.001	0.124	1.000

variables correlate with each other. For example, income and edu correlate rather highly with a value of .538 while edu and valmoney seem to be somewhat correlated with a value of -0.019, or valcaree and depend are even less correlated with a value of .001.