10.001: Data Visualization and Elementary Statistical Analysis

R. Sureshkumar

January 15, 1997

Statistics deals with the *collection* and the *analysis* of data in the presence of *variability*. Variability can be caused by the limitations on the resolution of measuring devices or the fluctuations in the conditions under which the measurement is made or the inherent non-constancy of the statistical variable itself or a combination of many of these factors.

1 An Example

Let's consider an example.[1] A manufacturer of metal alloy parts wants to address customer complaints concerning the non-uniformity in the melting points of an alloy his/her company produces and markets. Suppose we are given this problem as consultants. How can we get quantitative measures of the variability (dispersion) as well as the central tendency (location) in the melting point?

Collect a sample representative of the population.

First of all, we need to collect data for the melting point of the alloy parts produced in different batches. The statistical information we seek here is for the entire *population* of alloy parts which is distributed and sold to the customers. However, it is obviously an impractical task to measure the melting point of each one of the alloy parts produced

[1] See **/mit/10.001/Examples/MapleTutorial/figures.ms** for figures which are referred to in this document. Note that figures.ms is a maple file. Copy this into your working directory, load it into a maple session and print it.

in the company. Hence, we need to collect a *sample* which is representative of the entire population. In order for the sample to faithfully represent the statistical properties of the population, the process of sample selection should be done very carefully, by eliminating bias as much as possible. Let's denote such a sample by X, the size of the sample by n and each one of the sample elements by x_1, x_2,...,x_n.

The following data are for a sample of size 50, i.e., it contains the melting points of 50 alloy parts sampled randomly from the production line. The melting point measurements are rounded off to the nearest integer value in order to comply with the accuracy of the measuring procedure.

320 326 325 318 322 320 329 317 316 331 320 320 317 329 316
308 321 319 322 335 318 313 327 314 329 323 327 323 324 314
308 305 328 330 322 310 324 314 312 318 313 320 324 311 317
325 328 319 310 324

We will use this example to illustrate the key ideas and concepts of elementary statistical analysis.

2 Data Visualization/Graphical Analysis

A qualitative idea of the central tendency and the dispersion of the sample data can be obtained by representing it graphically. Typically, graphical visualization precedes quantitative statistical analysis. Figure 1 shows a *scatter plot* of the data, i.e., the data points x_1, x_2, ..., x_n plotted on a 2d graph against their indices. Some of the key statistical features of the data revealed by the scatter plot are: (a). a qualitative idea of the nature of variability, in this case it looks like a random scatter around an average melting point of approximately 320 and (b). the maximum and minimum values of the melting point which are 335 and 305 respectively.

A qualitative idea of how the data are *distributed* between the maximum and minimum values can be obtained by constructing a *frequency* plot, typically presented as a histogram. Here, we plot the number of observations of x_i, say f_i, vs. x_i. A frequency histogram for the melting point data above is given in Figure 2. From the frequency plot, we can see that the melting point of 320 occurs most frequently, i.e., $f(320) = 5$. The observation with the largest frequency is called the *mode* of the sample.

We can also tabulate the number of observation within a given interval and plot

the frequency data thus obtained against the mid-points of the corresponding intervals. For instance, we can create 7 equal intervals: [302.5,307.5), [307.5,312.5), [312.5,317.5), [317.5,322.5), [322.5,327.5), [327.5,332.5), [332.5,337.5). The interval $[a, b)$ contains the data greater than or equal to a and less than b. In Figure 3, you can see a plot of the number of observations of the melting point measurements within each one of these 7 intervals vs. the midpoint of the corresponding interval. As Figure 3 reveals, the distribution is approximately symmetric about a melting point value of 320.

3 Quantitative Description of the Data

There are primarily 2 types of measures we are interested in. The first type is known as the measures of *location* or measures of central tendency. The most commonly used description of this type is the *sample mean* denoted by μ and computed as

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{1}$$

For the data given above, $\mu \approx 320$.

The mean value of the data set is a single number which we extract from the entire data set using Eq. 1. In that sense, it is a function which maps the sample data to a single number representing a measure of location, i.e., it tells us what the average value of the data set is. This is true of almost all quantitative statistical measures: On one hand, they allow us to represent large sets of data compactly. On the other hand, we have discarded much of the detail in the process of arriving at a compact quantitative representation. This is why it helps to do a qualitative analysis using statistical plots. Moreover, we should try to combine the various quantitative measures to get a comprehensive description of the data.

Now, what are the other measures of location? The *median* value of the data set is defined as the observation so that half of the observations in the sample has values less than the median value. In other words, if we *sort* the data in the ascending order such that $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_{n-1} \leq x_n$, the median value is the middle value of the sorted data if n is an odd number, or it is the arithmetic mean of the two middle values if n is an even number. We will denote the median of the data set X by x_{med}. For instance, if we sort the melting point data above in the ascending order, we can find that $x_{med} = (x_{25} + x_{26})/2 = 320$.

Well, when do we use the median as a representative measure in preference to the

mean? The answer is when a sample of relatively small size contains extreme points, the mean may not be a representative measure, in such cases we tend to use the median of the sample. As an example, consider the data $(2.16, 2.37, 2.84, 3.01, 17.3)$, the mean value is larger than 4 of the 5 observations in the sample due to the relatively large weight of the last datum. Here, the median is a better representation of the central tendency.

In the melting point data given above, the mean, the median as well as the mode of the sample are practically the same (320). This is indicative of the *symmetry* of the distribution, as seen from Figure 3.

The idea of the median can be generalized to the concept of the *percentile* measure. Once the data has been sorted in ascending order, we can seek the observation x_P such that P percent of the (sorted) observations are below x_P. In this case, we call x_P the Pth percentile of the sample. The 25th, the 50th and the 75th percentiles are often referred to as the first, the second and the third *quartiles* respectively. The difference between the third and the first quartiles is called the *interquartile range*. For the melting point data of our example, the first and third quartiles are 316 and 325 respectively so that the interquartile range is 9. The interquartile range is often used to describe the variation of the data. The advantage of using such a measure over the *range* of the sample is that it avoids extreme data points often resulting from observations with relatively low confidence levels.

The *box and whisker* plot, or the box plot for short, graphically represents the median, the first and the third quartiles and the extrema of the sample data. A box plot of the melting point data is given in Figure 4. In the case of multiple samples, we can use a box plot to represent each sample. In that case, the widths of the boxes can be used to represent the relative sizes of the samples.

4 Variance, Standard Deviation and Coefficient of Variation

The most commonly used measure of variation (dispersion) is the sample standard deviation, σ. The square of the sample standard deviation is called the sample variance,

defined as[2]

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{2}$$

However,

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i^2 - 2\mu x_i + \mu^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\mu(\sum_{i=1}^{n} x_i) + n\mu^2 \\ &= (\sum_{i=1}^{n} x_i^2) - 2n\mu^2 + n\mu^2 \\ &= (\sum_{i=1}^{n} x_i^2) - n\mu^2. \end{aligned} \tag{3}$$

So an alternate equation for computing the variance is given by

$$\sigma^2 = \frac{1}{n-1}\left[(\sum_{i=1}^{n} x_i^2) - n(\sum_{i=1}^{n} x_i)^2\right]. \tag{4}$$

The advantage of Eq. 4 over Eq. 2 is that it allows for the computation of $\sum x_i^2$ required for the evaluation of σ and $\sum x_i$ required for the evaluation of μ in one loop, whereas Eq. 2 requires the precomputed value of μ before we can compute σ. For this reason, Eq. 4 is used often in the computations of the mean and variance.

However, if you closely examine Eq. 2 and Eq. 4, one important difference can be pointed out: Eq. 2 guarantees a non-negative variance because variance is given there as the sum of squares. This is not necessarily true of Eq. 4 where we subtract $n(\sum_{i=1}^{n} x_i)^2$ from $\sum_{i=1}^{n} x_i^2$. From a computational perspective, we know that this can cause difficulties for large samples prone to potential roundoff errors. So we are interested in developing an algorithm which computes (a). both the mean and the variance in the same loop and (b). variance as a sum of squares. How can this be accomplished? Well, we can resort

[2]Note that the sample variance is defined with $n-1$ in the denominator, if we use n instead of $n-1$ in Eq. 2, the quantity computed is referred to as the *population variance*. For large values of n, the sample variance is practically equal to the population variance. Here, the same symbols, μ and σ, are used to denote the mean and standard deviation respectively of the sample and the population. This is not to imply that they are the same, the sample mean and variance provide at the best an estimate of the population mean and variance. We will make the distinction between the sample and the population statistics from the context.

to developing recursive relations. Applying Eq. 1 for the the first $p-1$ and p data and subtracting one from the other, we get

$$p\mu_p = (p-1)\mu_{p-1} + x_p, \tag{5}$$

where μ_p denotes the mean value of the first p data of the sample. We can now compute the sample mean recursively by letting $\mu_1 = x_1$ and subsequently applying Eq. 5 for $p = 2, 3, \cdots, n$. We can also construct a simple recursion relation for computing σ^2 by applying Eq. 4 for the first $p-1$ and p data in the sample. This gives the two equations

$$\begin{aligned}(p-2)\sigma_{p-1}^2 &= x_1^2 + x_2^2 + \cdots + x_{p-1}^2 - (p-1)\mu_{p-1}^2 \\ (p-1)\sigma_p^2 &= x_1^2 + x_2^2 + \cdots + x_p^2 - p\mu_p^2.\end{aligned} \tag{6}$$

subtracting the first of Eq. 6 from the second one gives

$$(p-1)\sigma_p^2 = (p-2)\sigma_{p-1}^2 + (p-1)\mu_{p-1}^2 + x_p^2 - p\mu_p^2, \tag{7}$$

which can be rewritten (to get rid of subtractions) by substituting for μ_{p-1}^2 from Eq. 5 as follows:

$$(p-1)\sigma_p^2 = (p-2)\sigma_{p-1}^2 + p(x_p - \mu_p)^2/(p-1), \;\; p = 2, 3, \cdots, n. \tag{8}$$

Now, once we initialize $\mu_1 = x_1$ and $\sigma_1 = 0$, we can compute the sample mean and variance using Eq. 5 and 8 for $p = 2, 3, \cdots, n$ within the same loop. Note that the variance thus computed is guaranteed to be non-negative.

The coefficient of variation of the sample data, denoted by CV is defined as

$$\text{CV} = \frac{\sigma}{\mu}. \tag{9}$$

Note that CV is independent of the units of measurement.

5 Frequency Distribution Revisited

In section 2, we computed the number of observations of the melting point within a given interval for 7 intervals and plotted the interval frequency vs. the interval midpoint (refer to Figure 3). Such information obtained by grouping data into various intervals can be used to infer and model how the data is distributed in the entire population. In the context of our example, this means that by studying the statistical properties of the melting point

data of the sample grouped into different intervals, we can gain inferences on how the melting point is distributed among the entire population of alloy parts produced.

The first step in this kind of analysis is to define the intervals and find the frequency of observation within each interval. Let's say that we chose m intervals of equal length of Δx. Let's denote the mid-points of these intervals by P_j, for $j = 1, 2, \cdots, m$. An observation x_k belongs to the interval j if

$$P_j - \Delta x/2 \leq x_k < P_j + \Delta x/2. \tag{10}$$

Using the above criterion, we can find the interval frequency f_j for $j = 1, 2, \cdots, m$. The frequency distribution plot in Figure 3 shows f_j vs. P_j with $m = 7$. Note that the sample size n has to be equal to sum of the interval frequencies, i.e., $n = \sum_{i=1}^{m} f_j$.

Once we have collected the interval frequencies, we can compute the mean and the variance from the grouped data. We will use the subscript g to denote the statistical measures obtained from grouped data. In particular,

$$\begin{aligned} \mu_g &= \frac{\sum_{j=1}^{m} f_j P_j}{\sum_{j=1}^{m} f_j} = \frac{\sum_{j=1}^{m} f_j P_j}{n} \\ (n-1)\sigma_g &= \sum_{j=1}^{m} f_j (P_j - \mu_g)^2. \end{aligned} \tag{11}$$

For instance, μ_g for the melting point data grouped as in Figure 3 is 320.3.

The *cumulative frequency*, F_j is defined as the sum of all the frequencies of the intervals $\leq j$. You can compute F_j recursively as $F_j = F_{j-1} + f_j$, $j = 2, 3, \cdots, m$ with $F_1 = f_1$. Note that $F_m = n$.

5.1 Modeling the Melting Point Data: Gaussian/Normal Distribution

Even a cursory examination of Figure 3 reveals that the sample frequency distribution is approximately symmetric about a mean value of 320. Moreover, the maximum frequency occurs very close to the mean value. Furthermore, the frequency diminishes rapidly as we move either to the left or to the right of the mean value. These observations suggest the use of a bell shaped distribution to model the data. Or in mathematical terms, we seek to model the population frequency distribution using a Gaussian (normal) distribution. The Gaussian distribution is one of the most widely used probability distributions

with applications not only in statistical analysis of data but in theory of probability and stochastic processes. The mathematical expression for the normal distribution is given by

$$N(x:\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{12}$$

where $N(x:\mu,\sigma)$ denotes a normal or Gaussian distribution of variable x with mean μ and standard deviation σ. It can be shown that

$$\int_{-\infty}^{\infty} N(x:\mu,\sigma)dx = 1, \tag{13}$$

$$\int_{-\infty}^{\infty} xN(x:\mu,\sigma)dx = \mu, \text{ and} \tag{14}$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 N(x:\mu,\sigma)dx = \sigma^2. \tag{15}$$

Eq. 13 is equivalent to saying that that the distribution function is normalized such that the area above the x axis and under the $N(x:\mu,\sigma)$ curve is unity. Eq. 14 says that the expectation of the statistical variable is equal to the mean. Eq. 15 says that the second moment of the distribution about the mean value is equal to its variance. The points of inflection of $N(x:\mu,\sigma)$ are given by $x = \mu \pm \sigma$.

Now how can we relate the *discrete* frequency distribution of the melting point data to the *continuous* normal distribution? We can rephrase this question as: what is the appropriate frequency function which will approach $N(x:\mu,\sigma)$ in the limit of the interval length $\Delta x \to 0$ and the number of observations $n \to \infty$? Such a function can be constructed by suitably scaling f_j so that after scaling, the area under f_j vs. P_j curve is unity. We can then use the scaled f_j for comparisons against $N(x:\mu,\sigma)$.

Now, if we plot a histogram of f_j vs. P_j, the contribution to the area from interval j is $f_j\Delta x$. So the total area from m intervals is $\Delta x\sum_{j=1}^{m} f_j = n\Delta x$. Hence, in order to be consistent with the normalization given by Eq. 13, we should compare $f_j/(n\Delta x)$ with $N(x:\mu,\sigma)$. Such a comparison for the melting point data is shown in Figure 5. Note that in Figure 5, the continuous curve corresponds to $n\Delta x N(x:\mu,\sigma) = 250N(x:320.1,6.7)$ is plotted for $305 \le x \le 335$ and the points correspond to f_j for $j = 1,2,\cdots,7$. As we can see from Figure 5, the agreement is quite satisfactory, implying that the distribution of the melting points in the entire population of alloy parts can be modeled as a Gaussian distribution.

5.2 Predicting Probability from the Model Distribution

Once we have modeled the data using a distribution, we have at our disposal a *predictive* tool to evaluate the probability that a certain observation can be made between any 2 melting points x_a and x_b such that $x_a < x_b$. This probability, denoted by $\Pi(x_a, x_b)$, is the area under the distribution curve between $x = x_a$ and $x = x_b$; i.e.,

$$\Pi(x_a, x_b) = \int_{x_a}^{x_b} N(x : \mu, \sigma) dx. \tag{16}$$

Note that Eq. 13 is equivalent to the statement that the probability of making an observation between $-\infty$ and ∞ is unity.

How do we compute the integral in Eq. 16? First of all, we would like to define a new variable

$$z \equiv (x - \mu)/\sigma \tag{17}$$

which is independent of the units of measurement. Moreover, $z = 0$ for $x = \mu$ and note that z measures the how far we are away from the mean in units of the standard deviation. We now define the *standard normal distribution*, $N_s(y : 0, 1)$ as

$$N_s(y : 0, \sigma) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2). \tag{18}$$

Note that the standard normal distribution is a normal distribution with 0 mean and unit variance. We now define the standard cumulative distribution function $\Phi(z)$ based on $N_s(y : 0, 1)$ as

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-y^2/2) dy. \tag{19}$$

However, from Eq. 16, we have

$$\begin{aligned} \Pi(x_a, x_b) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{x_a}^{x_b} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{z_a}^{z_b} \exp(-\frac{z^2}{2}) dz \ \text{ (Note: } z = \frac{x-\mu}{\sigma}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_b} \exp(-\frac{z^2}{2}) dz - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_a} \exp(-\frac{z^2}{2}) dz \\ &= \Phi(z_b) - \Phi(z_a). \end{aligned} \tag{20}$$

The values of $\Phi(z)$ can be found in standard mathematical tables. In maple, invoke the stats package using with(stats); and do ?statevalf; to get the syntax of the

statevalf function which can be used to evaluate distribution values. In particular, $stateval f[cdf, normald](value)$ will give the numerical value of $\Phi(z)$ at $z = value$. It is important to note that $\Phi(-\infty) = 0$, $\Phi(0) = 1/2$ and $\Phi(\infty) = 1$. Moreover,

$$\Phi(-z) = 1 - \Phi(z). \tag{21}$$

See Figure 6 in the maple worksheet figures.ms for a plot of $\Phi(z)$.

The following series formula is also used to compute z, convergence is slower as the value of z becomes large.

$$\Phi(z) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \sum_{k=0}^{\infty} a_k z^{2k+1}, \tag{22}$$

where

$$\begin{aligned} a_k &= \frac{1-2k}{2k(1+2k)} a_{k-1}, \ \ n \geq 1. \\ a_0 &= \frac{1}{\sqrt{2}}. \end{aligned} \tag{23}$$

The following is a table which provides, z, $\Phi(z)$ accurate to 6 decimal digits and the value of k at which the series expansion of Eq. 22 was truncated to obtain that accuracy. Only $z \geq 0$ need be computed (see Eq. 21).

0.0	0.500000	0
0.2	0.579260	4
0.4	0.655422	5
0.6	0.725747	6
0.8	0.788145	7
1.0	0.841345	8
1.2	0.884930	9
1.4	0.919243	10
1.6	0.945201	12
1.8	0.964070	13
2.0	0.977250	14
2.2	0.986097	16
2.4	0.991802	17
2.6	0.995339	19
2.8	0.997445	20
3.0	0.998650	22
3.2	0.999313	24
3.4	0.999663	26

3.6	0.999841	28
3.8	0.999928	30
4.0	0.999968	32
4.2	0.999987	35
4.4	0.999995	37
4.6	0.999998	40
4.8	0.999999	42
5.0	1.000000	45

For very large values of z ($z >> 1$), we may use the asymptotic relation

$$\Phi(z) = 1 - \exp(-z^2/2)/(z\sqrt{2\pi}), \tag{24}$$

instead of the series expansion given in Eq. 22. For $z = 4$, the asymptotic result gives $\Phi(z) = 0.999966$, this is correct to 5 decimal places.

How do we make use of the information in the table above? For instance, let's ask: what is the probability $\Pi(x_a, x_b)$ that an observation is one standard deviation within the mean? Here, $x_a = \mu - \sigma$ and $x_b = \mu + \sigma$, so that $z_a = -1$ and $z_b = 1$. So the probability (see Eq. 20) $\Pi(\mu - \sigma, \mu + \sigma) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \approx 0.6827$. This implies that we expect 68.27% of the entire population of the alloy parts to have melting points between 313 and 327 if we use $\mu = 320$ and $\sigma = 7$.

Confidence levels are simply probabilities converted into percentages. For instance, the confidence level of the interval $[\mu - \sigma, \mu + \sigma]$ is 68.27% from the probability calculated above. It can be shown that the probability that an observation x falls within the range $[\mu - 1.96\sigma, \mu + 1.96\sigma]$ is 0.95 for the normal distribution. So the confidence level for this interval is 95%.

The discrete analogue of $\Phi(z)$ is the cumulative frequency data scaled with n plotted against $z_j = (P_j - \mu)/\sigma$. We expect F_j/n to approach $\Phi(z)$ in the limit $n \to \infty$ and $\Delta x \to 0$.

6 On Sorting Data

How do we sort the sample data in ascending order? A straightforward scheme is to find the smallest of the n sample points and assign that as x_1, then to find the smallest of the remaining $n - 1$ and assign that as x_2 and repeat the procedure until we have x_2. This

process requires a total of $(n-1)+(n-2)+\cdots+2+1=(n^2-n)/2$ comparisons. So for large values of n, the algorithm scales as n^2. A common n^2 algorithm for sorting, which can be found in almost any elementary book on computer science or statistical analysis is the *bubble sort* algorithm. The algorithm is given as follows.
1. Initialize n, x_1, x_2,...,x_n.
2. Define $K=n-1$, $L=0$.
3. while $(L=0)$
$L=1$
for $j=1,2,...,K-1$
if $(x_j > x_{j+1})$ then swap x_j and x_{j+1}, L=0
end for loop
K=K-1
end while loop
4. Print the sorted values of x_1, x_2,...,x_n.

Either of the algorithms given above should not be used for large values of n since more efficient methods which scale like $n\ln_2(n)$ exist in literature. These algorithms are based on clever partioning of the data into suitable subsets. A number of such methods may be found in chapter 8 of the NRC book. One of the most popular methods used in practice is the *quicksort* method. Note that the CPU time of the quicksort method depends on the initial ordering of data. While the best case scenario is order $n\ln_2(n)$, the worst case could be as bad as order n^2. The *heapsort* method is order $n\ln_2(n)$, irrespective of the ordering of the input data. The decision to use a quicksort over heapsort is typically made when sorting has to be done for a large number of data sets (i.e., many different samples), in which case, on the average CPU time for quicksort is typically smaller as compared to that for heapsort. In NRC, the quicksort function is called *void sort(unsigned long n, float arr[])* (page 333 of the NRC book) and the heapsort function is called *void hpsort(unsigned long n, float arr[])* (page 337 of the NRC book).

7 Using Maple for Statistical Analysis

The stats package in maple provides a number of subpackages and functions for data visualization, sorting, tabulating interval frequencies, computations of the measures of location and dispersion, computations of distributions and linear regression. Many of these functions are illustrated in the tutorial **/mit/10.001/Examples/MapleTutorial/stats.ms**.