Descriptive Statistics
(Devore Chapter One)

1016-345-01: Probability and Statistics for Engineers*

Spring 2013

Contents

Tuesday 9 April 2013

0 Terminology

We now turn to the consideration of data and the techniques of descriptive statistics which cane be used to summarize a set of data points.

Some terminology:

- A **Population** is a set of objects of interest (people, cards in a deck, cars manufactured in a given year, etc).

- A **Census** is a full accounting of all of the properties of the objects in a population.
- A **Sample** is one or more objects drawn from a population. One of the goals of statistics is to use a smaller sample in place of a full census to learn about the population.
- Sometimes the "full population" may not actually exist. For example, if I roll a die one or more times, I can make probabilistic statements about the outcome of that die roll, but there isn't really an underlying population of all die rolls. Instead, the probabilities of different outcomes can be thought of as resulting from an imaginary or **hypothetical population**, also known as a **conceptual population**.

When we come to numerical descriptions of data, we will make a connection between properties of the full population and properties of a random variable constructed by selecting one member of the population at random. Those properties can be generalized to a smaller sample taken from that population. But for now let's discuss some ways of getting a handle on the set of values in a sata sample.

1 Pictorial and Tabular Descriptions of Data

Imagine we have the set of 20 results, e.g., scores (out of 60) for students on a test:

$$56, 45, 37, 41, 41, 36, 27, 31, 41, 40, 48, 43, 43, 33, 44, 41, 35, 28, 37, 29$$

Note that this is a particular kind of data, where the variable being measured (in this case the score) falls into one of a countable set of values (in this case integers). The different kinds of variables we can measure include:

- **Discrete** (not discreet) variables, whose possible values can be listed in a (possibly infinite) sequence. Example: number of calls coming into a call center in an hour.
- **Continuous** variables, whose possible values consist of an interval (finite or infinite) of real numbers. Example: a person's height in centimeters.
- **Categorical** variables can take on one of a set of non-numerical categories. Example: the sex (male or female) of a child.

The example we're currently looking at concerns a discrete variable.

How do we get a handle on these numbers? One thing we can do is to list them in order from lowest to highest:

$$27, 28, 29, 31, 33, 35, 36, 37, 37, 40, 41, 41, 41, 41, 43, 43, 44, 45, 48, 56$$

That's starting to show us something. We can see a lot of results in the 40s, for example.

1.1 Stem-and-Leaf Displays

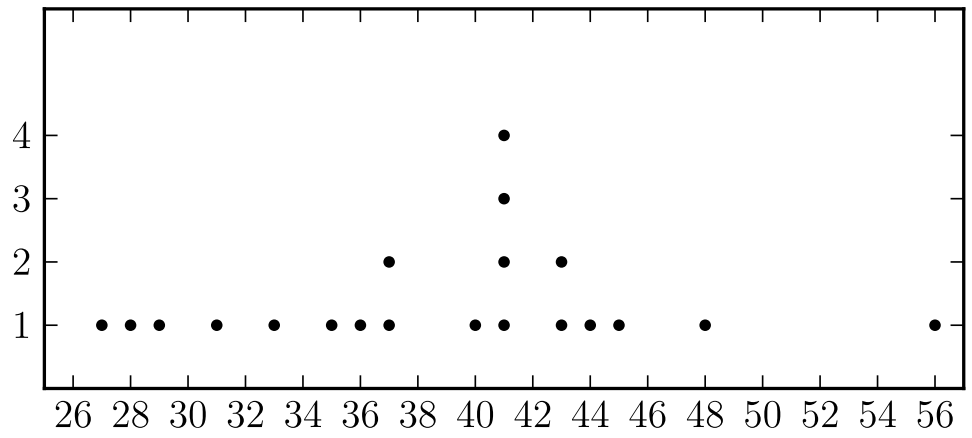
Collecting together the ordered list by decades is the basis for a simple tool in statistics (and a feature in many software packages): the stem-and-leaf display:

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

This contains the same information, in a form that draws the eye to features like the spike in the 40s.

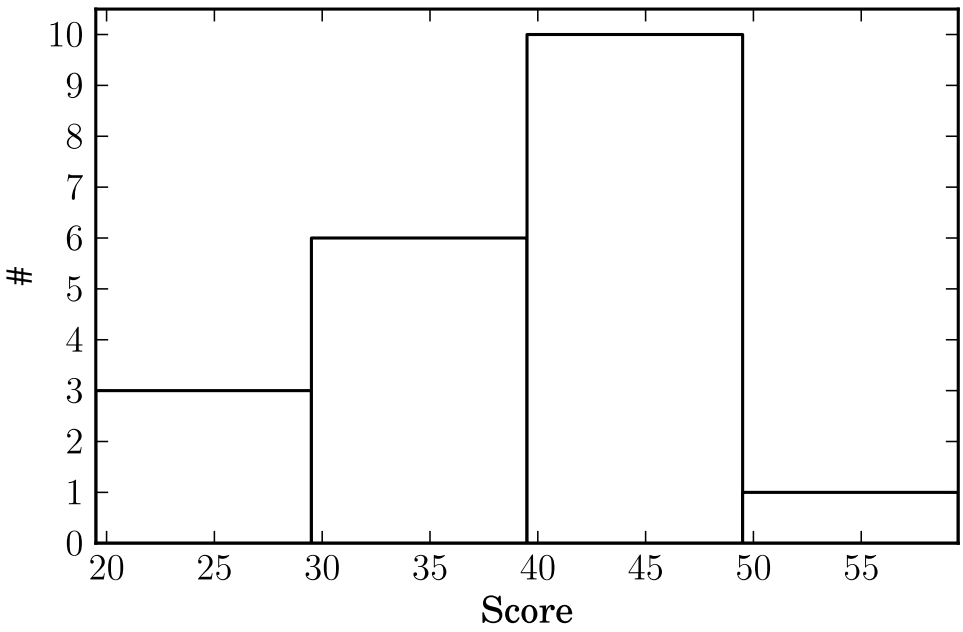
1.2 Dotplots

Another way to represent the data is to lay out an evenly-spaced scale and put a dot for each value. Since there are two 37s, we put two dots on top of each other there:



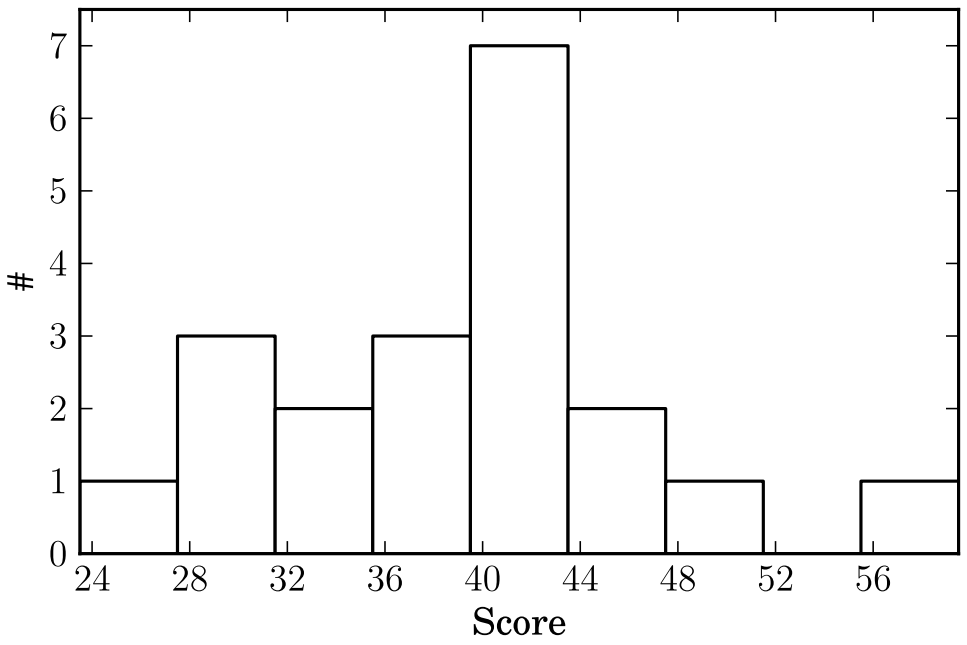
1.3 Histograms

Stem-and-leaf plots and dotplots, which represent each individual data point, work well for relatively small datasets, but if we had 200 or 2000 numbers to categorize rather than 20, it would be difficult to interpret something with so many individual numbers. A **histogram** is a good way to summarize data. For example, we could take our data and count 3 scores in the 20s, 6 in the 30s, 10 in the 40s, and 1 in the 50s. A histogram is just a bar chart of those numbers:



(For the purposes of the display, "the 30s", which includes the integers 30–39, is shown as ranging from 29.5–39.5, since that's the set of real numbers that would round off to those integers.) Note that this looks a lot like a sideways version of the stem-and-leaf plot, where the height of each bar is just the length (number of leaves) associated with each stem.

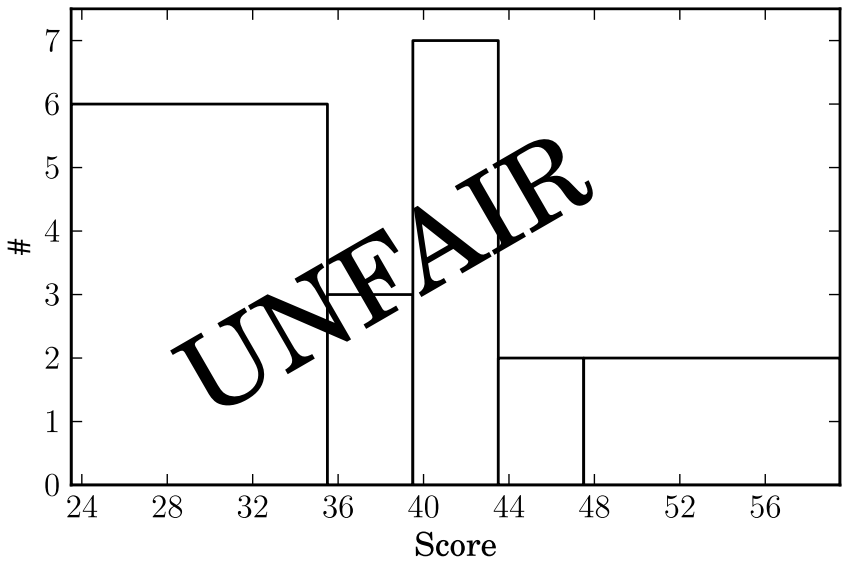
A histogram conceals some of the detail in a stem-and-leaf plot, but it's also more flexible; the categories need not be decades. For example, if I wanted to, I could make each category four points wide, and use as categories 24-27, 28-31, 32-35, etc. Since we have 1 in the first category, 3 in the second, 2 in the third, etc, the histogram looks like this:



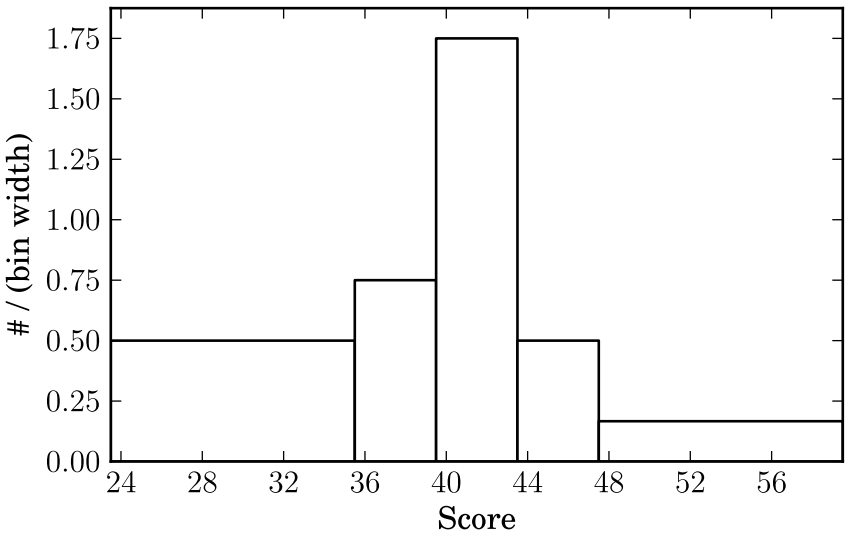
Note that the dotplot above is also basically a histogram, with each bin one point wide.

A tunable feature in a histogram is the bin width, and choosing this width is important to producing an intelligible histogram: too few bins, and you just get a couple of big blocks that don't carry much information; too many bins and you only have a few items in each bin, and the whole thing looks ratty and conceals the broad trends. Devore gives a rule of thumb that the number of bins of a histogram ought to be about the square root of the number of data points.

Note that you don't have to choose bins the same width, but if your bin widths are not the same, you have to be careful not to create a deceptive histogram. Suppose we wanted to choose the lowest bin to cover 24-35, combining the three lowest bins, and the highest bin to cover 44-59, combining the three highest. Well, now there are six of the data points total in the lowest interval and two in the highest, but if we plot that it's not really fair:



This is because we naturally look at the area of the bars as a gauge of how much data they represent, and so for example, the last bin, being two units high and 12 points wide, looks like it represents more data than the one next to it, two units high and only four points wide. So the natural thing to do with unevenly binned histograms is to make the height of each bar equal to the number of points divided by the width of the bins.



Now the area of each bar is equal to the number of points represented.

2 Numerical Representations of Data

Pictures are nice, but often we need to boil down properties of the data to a few numbers.

2.1 Sample Mean and Median

The first thing we might want to know about a numerical dataset, is what is the typical or average value.

What we usually mean by "average" is more precisely called the **mean**: add up all the numbers and divide by how many there are. If we call the observed values x_1, x_2, etc up to x_n (so that there are n of them), the mean, which we write as $\overline{x}$, is defined as

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

In the example from Tuesday, this is

$$\overline{x} = \frac{776}{20} = 38.8 \tag{2.2}$$

This is more concretely called the **sample mean** because $x_1, x_2, \ldots x_n$ (which I might write as $\{x_i | i = 1 \ldots n\}$) is a sample out of a hypothetical population.

There are some drawbacks to using the mean as an illustration of the typical value. For example, in this case there were also three students who had dropped the course by the time of the exam (so in effect they got 0), and also one who got a 60, which I left out because I wanted to have a multiple of four. If we include those, so that there are 24 scores, the average becomes

$$\overline{x} = \frac{776 + 60}{24} = 34.8 \tag{2.3}$$

Whether those scores are included or not makes a pretty big difference to the sample mean, but even with those included, only a third of the sample values (8 of 24) lie below the mean, and two-thirds lie above it. An alternative approach is to split the data set into its upper and lower half, and thus find the value which half lie above and half below:

0 0 0 27 28 29 31 33 35 36 37 37 | 40 41 41 41 41 43 43 44 45 48 56 60

Any value between 37 and 40 would work, but as a rule we take a number halfway in between, 38.5. This is called the **median** and we write it $\widetilde{x}$. In general, $\widetilde{x}$ is the middle value (when the values are listed in order) if n is odd, and the average of the two middle values if n is even (as it is here). Note that the median is not sensitive to those exremely high or low values the way the mean is; it doesn't matter that they're 0; they could be any number below 38.5. (This makes it particularly sensible in this case, because students who dropped the class would probably have scored somewhere in the lower half of the class on the test.)

If we had used the original data set, the median would have been 40.5; still higher, but not as big a change.

27 28 29 31 33 35 36 37 37 40 | 41 41 41 41 43 43 44 45 48 56

2.2 Population Mean and Median

Just as you can define the mean and median for a sample, you could do the same thing for the whole population. This is easy to write down if the population has a finite number of members, call it N. The **population mean** is written μ and defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.4}$$

The only difference to the definition of the sample mean is that the sum is over the whole population and not just the sample.

Similarly, the **population median** $\widetilde{\mu}$ is the value such that half of the population lies above and half below it.

Note that there is some inconsistency in the notation, because the population mean μ is not written with a bar. The notation is summarized as

	Sample	Population
mean	$\overline{x}$	μ
median	$\widetilde{x}$	$\widetilde{\mu}$

Practice Problems

1.11, 1.15, 1.17, 1.29

Thursday 11 April 2013

2.3 Quartiles, Boxplots, Fourth Spread, and Outliers

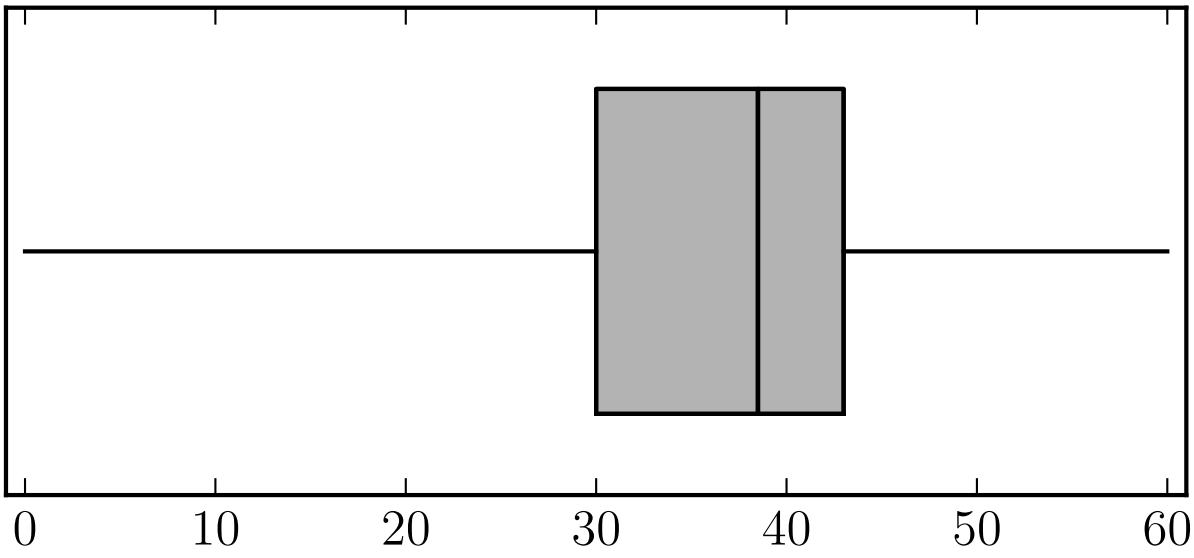
Having divided the data into halves, we could also divide it into fourths, and use that to get a sense of how spread out they are:

0 0 0 27 28 29 | 31 33 35 36 37 37 | 40 41 41 41 41 43 | 43 44 45 48 56 60

The bottom fourth runs from 0 to 29; the second fourth runs from 31 to 37; the third fourth from 40 to 43, and the top fourth from 43 to 60. Of course, that's now eight numbers, but doesn't convey that much information, so we can average the points on the boundaries to get what's called the "five-number summary":

- The smallest value is 0
- The lower fourth boundary is 30
- The median is 38.5
- The upper fourth boundary is 43
- The largest value is 60

That can be summarized in a plot:



We can see at a glance that the data spread out more below the median than above it. One measure of the amount of spread is the difference between the lower fourth and upper fourth, i.e., how much the middle half of the data are spread out. We call this the **fourth spread** and write it f_s In our case that is

$$f_s = 43 - 30 = 13 \tag{2.5}$$

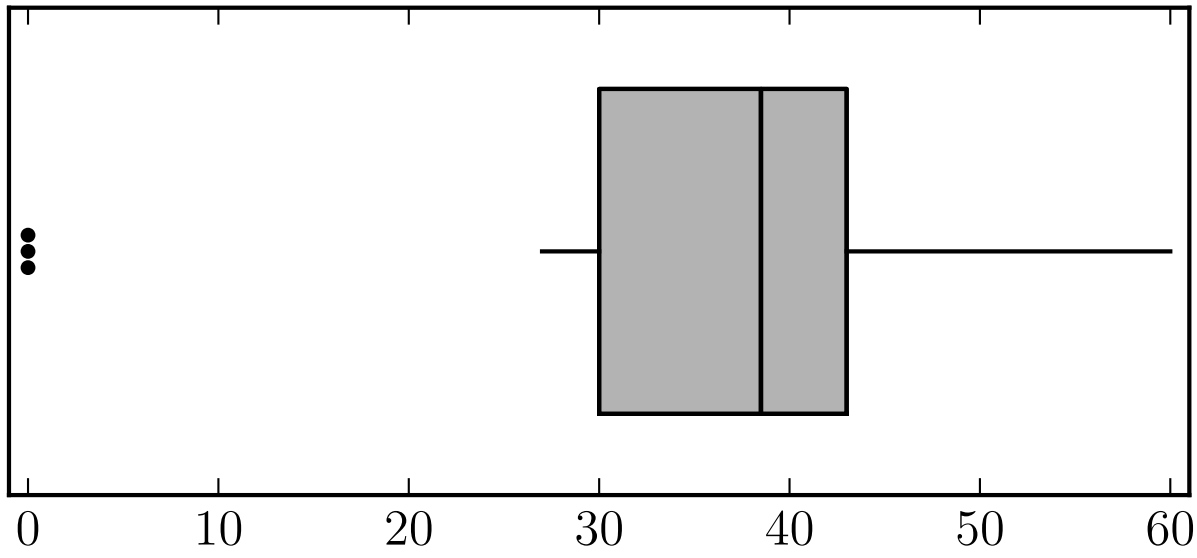
The presence of the three scores at zero is sort of concealing the spread of most of the data. We call those scores "outliers" since they lie far away from most of the data. One precise definition of an outlier, which we'll use in this course, is in terms of the fourth spread:

- An **outlier** is any data point more than $1.5f_s$ away from the closest fourth boundary (upper or lower)
- An **extreme outlier** is any data point more than $3f_s$ away from the closest fourth boundary (upper or lower)
- A **mild outlier** is an outlier which is not extreme.

In our case, $f_s = 13$, so $1.5f_s = 19.5$ and $3f_s = 39$. Since the upper and lower fourths are 30 and 43, in our example

- x is an extreme outlier if $x < -9$ or $x > 82$ (both of which are actually impossible given the range of possible scores on the exam)
- x is a mild outlier if $-9 \leq x < 10.5$ or $62.5 < x \leq 82$

We can show outliers in a boxplot by putting dots for the individual values and then only having the "whiskers" go out to the lowest and highest non-outlier values, which in this case are 27 and 60.



(If there were extreme outliers, we'd represent them with open circles.)

2.4 Variance and Standard Deviation

The fourth spread is sort of an anologue to the median. To make a measure of variability analogous to the mean, we should look for some sort of average difference from the typical value.

It turns out to be easier to think about the population first, so let's think about how all of the values differ from the population mean μ. One idea would be to take the average of $x_i - \mu$, but that quickly runs into a problem. It is

$$\frac{(x_1 - \mu) + (x_2 - \mu) + \ldots + (x_N - \mu)}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu) \tag{2.6}$$

Now, we can reorganize the terms in the numerator, to collect the N different x_is and the N copies of μ, and that gives us

$$\frac{(x_1 + x_2 + \ldots + x_n)\overbrace{-\mu - \mu \ldots - \mu}^{N \text{ copies}}}{N} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) - \frac{N\mu}{N} \tag{2.7}$$

But the expression in the big parentheses is just μ, so we end up with

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu) = \mu - \mu = 0 \ . \tag{2.8}$$

The problem is that the terms with $x_i < \mu$ were negative and taken together they cancelled out those with $x_i > \mu$. We want something that measures how far away each member of the

population is from μ, in a form where values that are smaller and larger will both contribute positively. So instead we take the average of $(x_i - \mu)^2$; we're guaranteed that each term is either zero or positive, and so the sum will be zero or positive. (And it'll only be zero if all of the terms are zero, i.e., each value in the population is the same.) This is the **population variance**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \ . \tag{2.9}$$

Since the population variance is guaranteed not to be negative, we can take its square root and get what's called the **population standard deviation**

$$\sigma = +\sqrt{\sigma^2} \ . \tag{2.10}$$

Now let's think about a sample of n objects drawn from the population. We know the sample mean $\overline{x}$ is an estimate of the underlying population mean μ, and the sample median $\widetilde{x}$ is an estimate of the population median $\widetilde{\mu}$. We'd like to construct something from the sample to approximate the population variance σ^2. What you'd like to do is average $(x_i - \mu)^2$ over the sample, but we don't know the population mean μ because it's a property of the underlying population, not just of the sample. so the best thing we can do is use the sample mean $\overline{x}$ as a stand-in. But the average of $(x_i - \overline{x})^2$ turns out to be an understimate of the population variance σ^2, because you're using the same data to estimate the mean and the variance. Actually

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ is an estimate of } \left(\frac{n-1}{n}\right)\sigma^2 \ . \tag{2.11}$$

The demonstration of this is a little involved, but we can see it for the simple case where $n = 1$. If there's only one item in the sample, than the sample mean must be $\overline{x} = x_1$. But then the average of $(x_1 - \overline{x})^2$ is zero no matter what σ is, which is what (2.11) says for $n = 1$.

In general, (2.11) tells us that we should define the **sample variance** by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.12}$$

and the **sample standard deviation** by

$$s = +\sqrt{s^2} \tag{2.13}$$

Note that for practical computations of both σ^2 and s^2 one typically uses a shortcut that comes from expanding the square:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}\left[(x_i)^2 - 2x_i\overline{x} + (\overline{x})^2\right] = \sum_{i=1}^{n}(x_i)^2 - \sum_{i=1}^{n}(2x_i\overline{x}) + \sum_{i=1}^{n}(\overline{x})^2 \\ &= \sum_{i=1}^{n}(x_i)^2 - 2\overline{x}\sum_{i=1}^{n}x_i + n(\overline{x})^2 = \sum_{i=1}^{n}(x_i)^2 - 2\overline{x}(n\overline{x}) + n(\overline{x})^2 = \sum_{i=1}^{n}(x_i)^2 - n(\overline{x})^2 \end{aligned} \tag{2.14}$$

This means that

$$s^2 = \frac{1}{n-1} S_{xx} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i)^2 - \frac{n}{n-1} (\overline{x})^2 \ . \tag{2.15}$$

By a similar argument

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i)^2 - (\mu)^2 \ . \tag{2.16}$$

2.5 Calculating the Mean and Variance given $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} (x_i)^2$

Given a sample of n numbers, we can easily calculate their sum

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{2.17}$$

and the sum of their squares

$$\sum_{i=1}^{n} (x_i)^2 = (x_1)^2 + (x_2)^2 + \cdots + (x_n)^2 \tag{2.18}$$

For example, given the $n = 20$ numbers we considered before, $\sum_{i=1}^{n} x_i = 776$ and $\sum_{i=1}^{n} (x_i)^2 = 31086$.

From $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} (x_i)^2$, we can calculate, from (2.1), the sample mean

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.19}$$

and, from the shortcut formula (2.15), the sample variance

$$s^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} (x_i)^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right] \ . \tag{2.20}$$

(It's usually better to do it this way to avoid introducing errors from rounding off $\overline{x}$.)

So for our example,

$$\overline{x} = \frac{1}{20}(776) = 38.8 \tag{2.21}$$

and

$$s^2 = \frac{1}{19} \left[31086 - \frac{1}{20} (776)^2 \right] = \frac{977.2}{19} \approx 51.4316 \approx 51.4 \ . \tag{2.22}$$

The sample standard deviation is $s \approx \sqrt{51.4316} \approx 7.17$.

It's a good idea to practice this using some of the quizzes from previous quarters!

Practice Problems

1.33, 1.35, 1.39, 1.49, 1.51, 1.69, 1.73

Tuesday 16 April 2013

3 Probability Plots (Devore Section 4.6)

Back in Chapter One, we defined a bunch of properties of a sample: mean, variance, standard deviation, median, and various percentiles. We also defined the corresponding properties for the underlying population. Since then, we've been considering random variables, which can be thought of as based on conceptual populations, and defined analogous properties of mean, variance, standard deviation, median and percentiles associated with the underlying probability distribution.

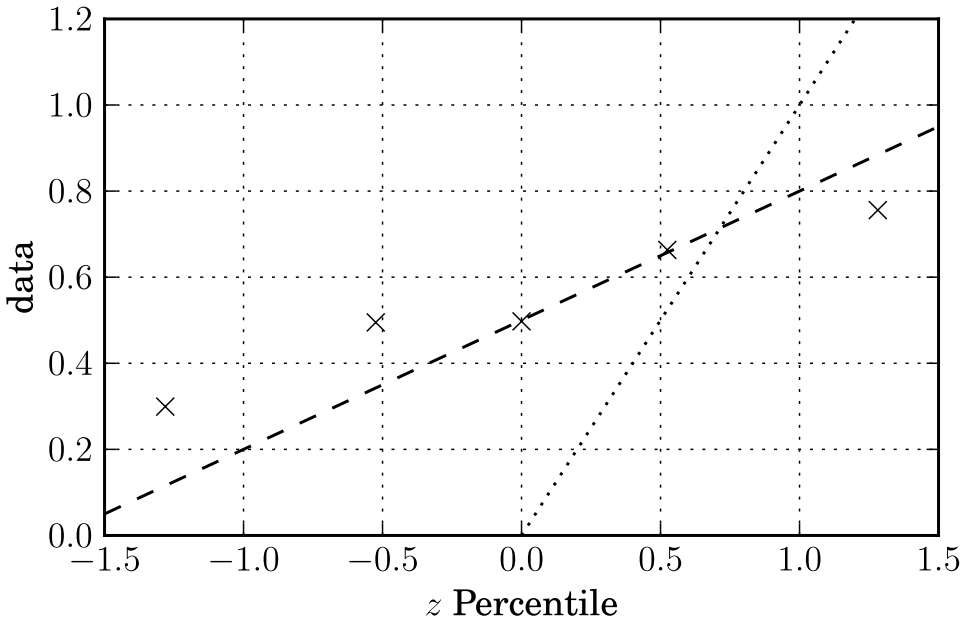
One thing we can do is consider a data sample, and check how likely it is to have originated from a given probability distribution. (We'll only do this qualitatively at this point.) One simple thing to do is compare the median: is the sample median close to the median of the proposed probability distribution? We can also start to ask this about various percentiles of the data: do, for example, the top 10% of the data lie above the 90th percentile of the probability distribution? But which percentiles do we check? We can't really check more percentiles than we have data points, and we'll let the points themselves tell us where to check. Recall that if we have an odd number of points, the middle one (once they're sorted into order) is the median, but if we have an even number, we have to interpolate. The idea, if we have e.g., 5 points, is that two and a half lie below and two and a half lie above the middle value. We can do the same trick with the other points: the lowest value has half a point below and 4.5 above it, so it's the 10th percentile of the sample. With 5 points we get the following correspondence:

Sample #	1	2	3	4	5
Pct below	10	30	50	70	90

(In general, the percentile corresponding to point i after sorting a sample of n points is $100(i - .5)/n$.) We plot each value against the corresponding percentile of the distribution we want, so if we have five samples the lowest value in the sample is plotted against the 10th percentile of the distribution, the next lowest against the 30th percentile, etc. Let's look at this for an example data set, and see if it fits a standard normal distribution, from which we can use z_α with $\alpha = 1 - (i - .5)/n$.

Sample #	1	2	3	4	5
Percentage	10	30	50	70	90
z percentile	-1.282	-0.524	0.000	0.524	1.282
Sample obs.	0.299	0.495	0.497	0.663	0.756

The data points don't agree at all well with the corresponding percentiles of the standard normal distribution, which we can see by drawing a dotted line with $y = x$ on the corresponding plot:



Now, this is not so surprising, since I generated the data using a normal distribution with $\mu = .3$ and $\sigma = .5$. If we work out the percentiles of the distribution with those parameters, the values are much closer:

Sample #	1	2	3	4	5
Percentage	10	30	50	70	90
z percentile	-1.282	-0.524	0.000	0.524	1.282
Sample obs.	0.299	0.495	0.497	0.663	0.756
$N(0.5, 0.3^2)$ *percentile*	0.116	0.343	0.500	0.657	0.884

Now, we may often want to check whether data are consistent with a normal distribution without specifying μ and σ for that distribution. The cool thing is that we don't have to, because any normal random variable $X \sim N(\mu, \sigma^2)$ is related to a corresponding standard normal random variable by

$$Z = \frac{X - \mu}{\sigma} \tag{3.1}$$

or

$$X = \mu + \sigma \cdot Z \tag{3.2}$$

This means that the $100(1 - \alpha)$ percentile x_α is

$$x_\alpha = \mu + \sigma \cdot z_\alpha \tag{3.3}$$

and a sample is consistent with a normal distribution if it lies close to a straight line on a normal probability plot. The dashed line on the plot above is $.5 + .3z$, which passes through the appropriate percentiles for a normal distribution with $\mu = .5$ and $\sigma = .3$.

Practice Problems

4.87, 4.89, 4.91, 4.93, 4.97, 4.111