STAT 100 **Notes** **January 21, 2010**

Means & SDs & Normal Distributions

SDs:

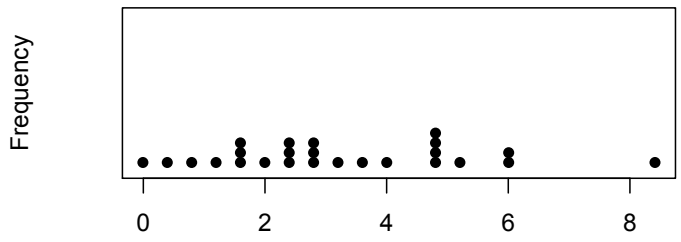
The SD of a set of numbers is the "typical" size of a deviation from the mean. But that is just the general idea. We need more for a definition. If x1, x2, x3, …,xn is a set of data, then its SD is computed as follows:

1. Compute the mean, call it $\bar{x}$ (say "x-bar")
2. Subtract the mean from each number in the data set to form the deviations
3. Square all the deviations and add them up
4. Divide the sum from 3. by n-1
5. Take the square root of the result in 4.

So if my numbers are 1,2,5,
1. mean is 4.0
2. deviations are -3, -2, +1
3. squared deviations are 9, 4, 1 and their sum is 14
4. 14/2 = 7
5. The square root of 7 is 2.65

So the SD of {1,2,5} is 2.65

Here is another data set, shown as a "dot-plot".



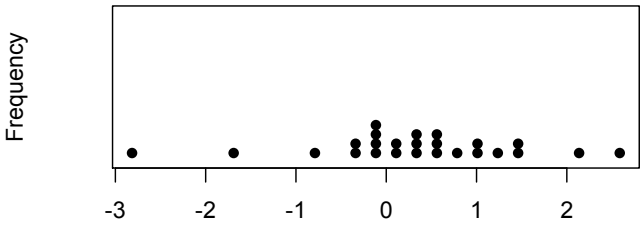
Q: Guess the SD. What is a typical deviation in this case?

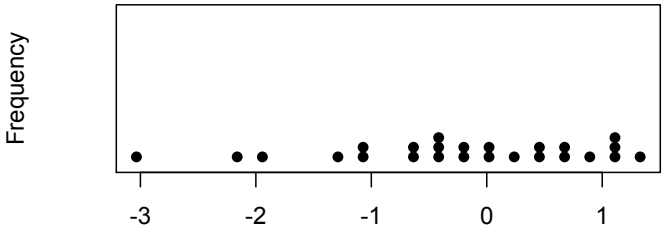
Not much point in calculating it by hand – calculators and computers do it easily. But you need to know something about it. For example, if I add the number 20 to the above data set, would it change the SD very much? The answer is yes, because the calculation involves a squaring of the deviation for 20, which is much bigger than the square of all the other deviations. So the SD is not a very good measure of "typical deviation" when there are "outliers" in the data. It is best as a descriptive summary for a data distribution symmetric around its mean.

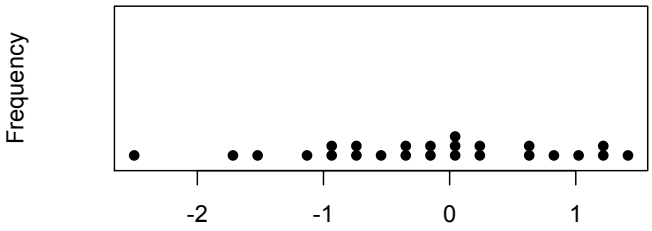
Normal Distributions

A model that is very useful for statistical summaries is the Normal Distribution. ("All models are wrong but some are useful".)

Here are a few samples of data (sample size 25) from a Standard Normal Distribution – that is, a Normal Distribution that has mean=0 and standard deviation =1.

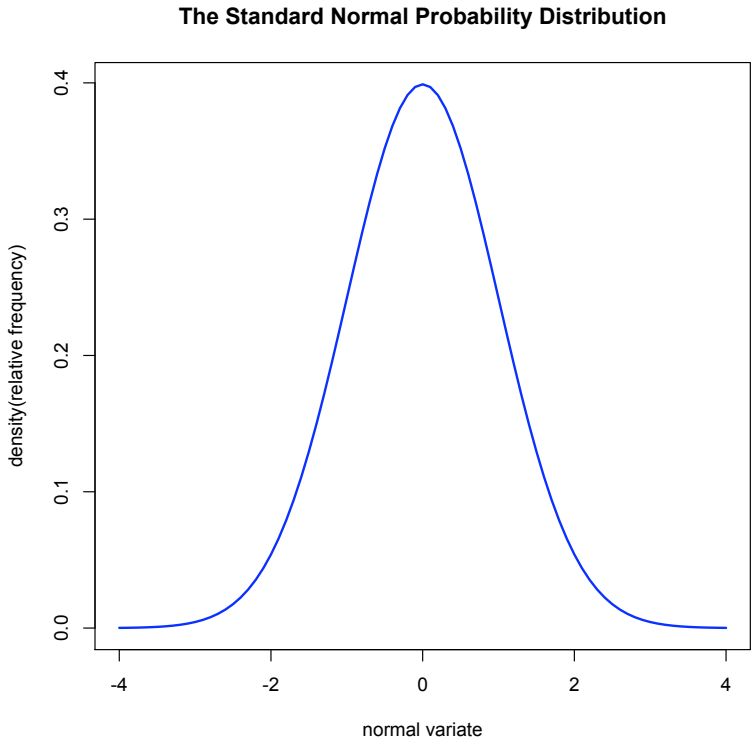






Note: These samples of 25 values do not have a bell-shaped distribution, because they are sample distributions, not probability distributions. But they have all been generated by randomly sampling a real Normal probability distribution with mean 0 and SD=1. The samples will not even have mean 0 nor SD 1, although they will be close. But the variability of means and SDs in such samples will be a primary interest for applications, as we shall see.

The page below shows the standard normal distribution: you can think of it as the shape that a histogram would have if you had an infinite-sized sample of data. It has no jagged edges since with lots of data the width of rectangles is as small as you want. Note that the vertical scaling is determined by the "relative" frequency nature of probability. We want the probability that a randomly selected value from the distribution is somewhere between negative infinity and positive infinity to be 1.



The way this curve relates to probabilities is that the area under the curve represents probability. For example, the area to the right of 0, under the curve, is ½. Remember area is just width x height, so if you filled up the area under the curve with small squares, and added up their area, the total would be 1, and the total to the right of 0 would obviously (by symmetry) be ½. (Of course, I am assuming you use the vertical scale for the vertical side of the squares and the horizontal scale for the horizontal size of the squares, in calculating "area". The actual square inches of the graph is irrelevant for this.)

How much area lies between -1 and +1 in the standard normal distribution? The answer is approximately 68%. Between -2 and +2, it is about 95%, and between -3 and +3 it is about 99.7%. Data analysts have computer codes to compute a probability for any interval in this distribution. For example, it turns out that the proportion of values in the interval (-1.3, +0.6) in the probability distribution is 62.9%. Of course, in any sample of say 25 random values from this distribution, the proportion in (-1.3, 0.6) will not necessarily be 62.9% and may not even be close to that. So what use is a model that has exact properties in theory (probability) if it does not work in practice (samples)?

The answer is that, although the samples will not have a distribution that looks like the probability distribution that produced the sample, the model does better at helping to describe some properties of the sample, like sample mean and sample SD. The next step in this discussion is to examine what happens in samples from the standard normal probability distribution to the sample mean and sample SD?

The sample mean and sample SD

From now on we use the shorthand "N(0,1)" for "standard normal probability distribution". Also, when we are talking about generating samples from a probability distribution, we often call it sampling a "population".

Here are some examples of sample means and sample SDs from a N(0,1) population. The sample size in each case is 25:

Mean	SD
-0.23	1.12
0.24	1.23
-0.04	0.95
0.04	1.11
-0.30	1.09

Now suppose that on the basis of one of these samples, we wanted to estimate the probability mean. (Although we know it is 0, lets suppose we did not know that). Naturally, our estimate would be the sample mean. How wrong is it? On the evidence so far, it looks like we would be within about .3 of the true value, which is 0. This is where some theory can help. The theory says that 95% of the time the sample mean will be within about 0.4 SDs of the true mean. Where did that 0.4 come from?

First we need a formula:

SD of sample mean = SD of population / $\sqrt{}$sample size

Now consider the first sample above with mean -0.23 and SD 1.12. If we use 1.12 as our estimate of the population SD, then we compute

SD of sample mean = 1.12 / $\sqrt{}$25 = 1.12/5= .224

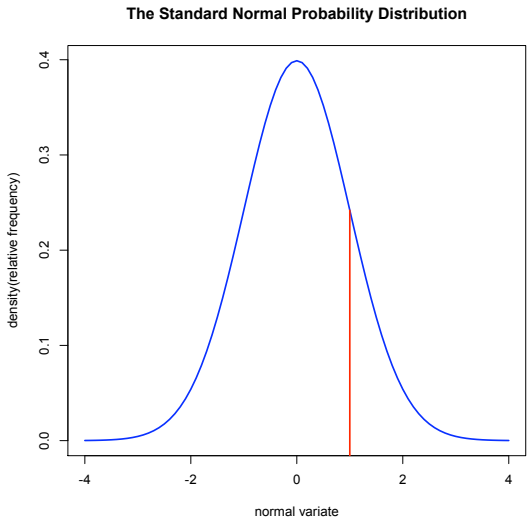
If the sample mean has a normal distribution , then 95% of the time it should be within 2 SDs of its mean. So 95% of the time, we expect -0.23 is within 2 x .224 = .448 of the population mean. Or we could say, population mean = -0.23 +- .448.
Note that this one time is one of the 95%, since the true mean 0 really is in this interval.

In fact , for the other four samples, estimates of the mean would be
.24 +- .49
-0.04 +- .38
0.04 +- .44
-0.30 +- .44
and in each case the true value 0 is in the interval.

Next we consider normal populations that are not standard – N(mean, SD) where the mean is not necessarily 0 and the SD is not necessarily 1.

N(mean, SD): Any Normal distribution can be related to the standard normal by expressing its values in terms of "SDs from the mean".

If someone have an IQ of 115, and I know IQs are designed to be Normally distributed and have mean 100 and SD 15, then this person's IQ is 1 SD above the mean. Like this.



Note that, although I am talking about a property of the N(100, 15) distribution, the graph is showing the N(0,1) distribution, and the red line is in the right place to judge how high the persons IQ is relative to the overall distribution of IQs.

Q: From what I have told you so far, it is possible to figure out that the IQ of 115 is higher than 84% of the population. Do you see why?

So any probabilities associated with Normal distributions can be determined from the probabilities for the N(0,1) distribution.

The variability of sample Means

Now the real reason for introducing SDs and Normal distributions is to describe the variability of a sample mean (from one sample of size n to another sample of size n).

Look again at our five samples of size 25 that had the following summaries:

Sample	Mean	SD
1	-0.23	1.12
2	0.24	1.23
3	-0.04	0.95
4	0.04	1.11
5	-0.30	1.09

Suppose again that we do not know the mean and SD of the population that generated these samples. And, suppose after the first sample is observed, we want to guess the variability of the Mean in subsequent samples. Sounds impossible but it can be done pretty well.

From the formula we showed before

SD of sample mean = SD of population / $\sqrt{}$sample size

Now the "variability of the Mean in subsequent Samples" is measured by "SD of sample mean" so the formula should help. But the problem is we are supposing we do not know the SD of the population. However, the sample SD should be close to it so we use it instead. Then

SD of sample mean = (approx) SD of sample / $\sqrt{}$sample size
$1.12/\sqrt{25} = 1.12/5 = .224$

OK. How good an approx is it? Look at the other means {0.24,-0.04,0.04,-0.30}. The SD of these 4 means is .224. Wow! That is exactly right. Well, this method is not always so accurate but it is good enough to be very useful.
Do you see why there should be some surprise in this accuracy? I have used ONE sample to provide information about FOUR other samples. The information I extracted was the variability of the means of the FOUR samples, and I based it on a completely different sample. Of course, it was important that all five samples were random samples from the same population. That population happened to be normal but that was not really a requirement of the method.

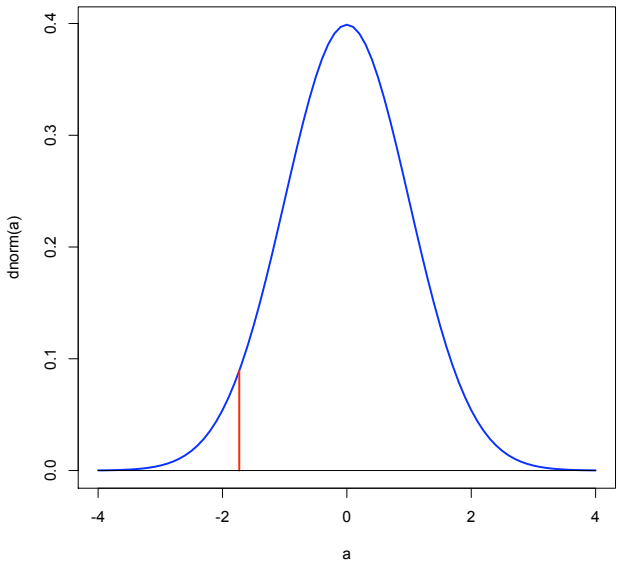
Applications of the Method

An example:

You own an auto-repair shop and you record your revenue each day. A large factor in your revenue is the hourly charge to customers for labour. You have been charging the same rate for several months but you would like to increase the rate so you can pay more to your deserving staff. But you are worried that a raise in the labour rate will cut into your revenues if your customers spread the word about your higher rate. Prior to the rate rise your revenues averaged $2000 with a SD of $200 and was fairly stable over several months. After the hike your revenues averaged $1900 with a SD of $190 based on only twelve days of operation. Can you decide already that the rate hike is hurting revenues, or might the drop be simply random variation?

The key question is how well do the twelve post-hike revenues estimate the post-hike average that your shop will experience if the hike persists for a longer time? In other words, is the $1900 average the result of sample variation from a probability distribution with mean $2000 and SD $200?

We can compute an estimate of the variability of the mean of twelve random values from the N(2000, 200) distribution using our formula. SD of the average = approx $200/\sqrt{12} = 57.7$. That means the observed sample mean of \$1900 is just under 2 SDs below the mean of \$2000. Actually (2000-1900)/57.7 = 1.73 - in other words, 1.73 SDs below the mean. Let's see where this is on our Normal distribution:



So, \$1900 is a pretty low value if the true mean of the distribution underlying it had mean \$2000. But is it low enough to provide strong evidence that our assumption (mean=\$2000) is untenable? While this is a personal decision, statisticians usually require a value in the 5 % tail before rejecting the assumption of the calculation. By looking carefully at the graph, or looking up the value via calculator or computer, the tail area to the left of the red line above is 0.042 – i.e. 4.2% of the probability is to the extreme of the observed value. So if we use the traditional 5% critical value, we would take this data as evidence that the rate hike has actually reduced revenues.

A recap of theory

The above discussion is an example of hypothesis testing. Remember the logic "If something happens that is unusual under ordinary circumstances, then this is evidence that the circumstances are not ordinary." The additional complication involving means, the SD, and the Normal distribution is very often involved in this "hypothesis testing" process.

Why the Normal distribution? Isn't it just a model that could be very wrong? A piece of theory that we will visit frequently in future explains that Normal distributions pop up naturally when we are examining means, no matter what the sampled distribution is. More on that later under the banner of "The Central Limit Theorem".

A rough table of the Normal Distribution:

x=SDs above the mean	Probability less than x	For Remembering (approx)
-4	0.00003	0
-3	0.0013	0
-2	0.023	2.3%
-1	0.16	16%
0	0.50	50%
+1	0.84	84%
+2	0.977	97.7%
+3	0.9987	100%
+4	0.99997	100%

Or, even more simply and approximately:

Within 1 SD of the mean 68%
Within 2 SDs of the mean 95%
Within 3 SDs of the mean 100%

KLW 2010/01/21