Basic Statistics Concepts for Empirical Science Workshop

January 29 2006
Jim Mahoney

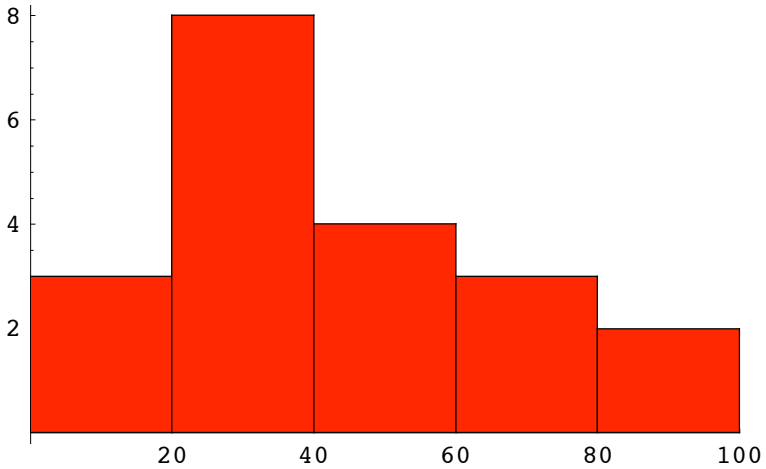
Some notes on basic ideas in statistics, particularly random errors and how results are reported in scientific journals.

the Mean

First, just to have some numbers to work with, suppose we pick a number at random from 0 to 100, assuming each has the same probability. Doing this N=20 times might give the following numbers and histogram.

```
minNumber = 0;
maxNumber = 100;
howMany = 20;
x = Table[Random[Integer, {minNumber, maxNumber}], {howMany}]
<< Graphics`Graphics`;
Histogram[x];
```

```
{21, 30, 22, 14, 23, 59, 23, 20, 62,
 92, 21, 82, 1, 77, 28, 54, 19, 53, 65, 54}
```



The graph shows how many times we got a number in a certain range.

I'll call the individual numbers x_i. In other words, $x_1 = 97$, $x_2 = 75$, and so on.

The mean, or arithmetic average, of these number is their sum divided by how many you have. Intuitively the mean describes the center of the numbers. (However, be aware that there are other measures of the center, especially the "median".)

Using the "Σ" symbol to mean "sum", or "add 'em all up", and "N" for "howMany", the formula for the mean is

$$\text{mean}(x) = m = \langle x \rangle = \bar{x} = \frac{\Sigma x_i}{N} \tag{1}$$

In this case,

```
N[Mean[x]]
```

```
41.
```

You can do this calculation with any tool you like: a handheld calculator, an online data analysis tool, Excel, *Mathematica*, or whatever.

the Standard Deviation

The standard deviation, usually indicated by the symbol σ (sigma) or just the letter "s", describes how far the numbers are from the mean.

Without going into the reasons why here, the formulas look like this :

$$\begin{aligned} \text{standard\_deviation}\,(x) = \sigma &= \sqrt{\text{mean}(\,(x_i - m)^2\,)} = \sqrt{\text{mean}(x_i{}^2) - m^2} \\ &= \sqrt{\frac{\Sigma\,(x_i - m)^2}{N}} = \sqrt{\frac{\Sigma\,(x_i)^2}{N} - \left(\frac{\Sigma\, x_i}{N}\right)^2} \end{aligned} \tag{2}$$

For the numbers up above,

```
<< Statistics`DescriptiveStatistics`;
N[StandardDeviationMLE[x]]
```

```
25.2527
```

One particularly confusing thing about the formula for standard deviation is that there are actually two versions, one with an N in the denominator, and one with $N - 1$. Most of the time (that is, as long as N isn't too small) the difference won't matter. Different calculators use different conventions as to which of these they compute, so you'll need to read the documentation closely or do some tests to see which one your particular program is giving you.

$$\text{standard\_deviation\_best\_guess\_of\_parent\_population}\,(x) = s = \sigma\sqrt{\frac{N}{N-1}} \tag{3}$$

The difference is this : σ is the true standard deviation of those particular numbers, while s is the best estimate of the standard deviation of the larger population that those numbers came from.

the Normal Distribution

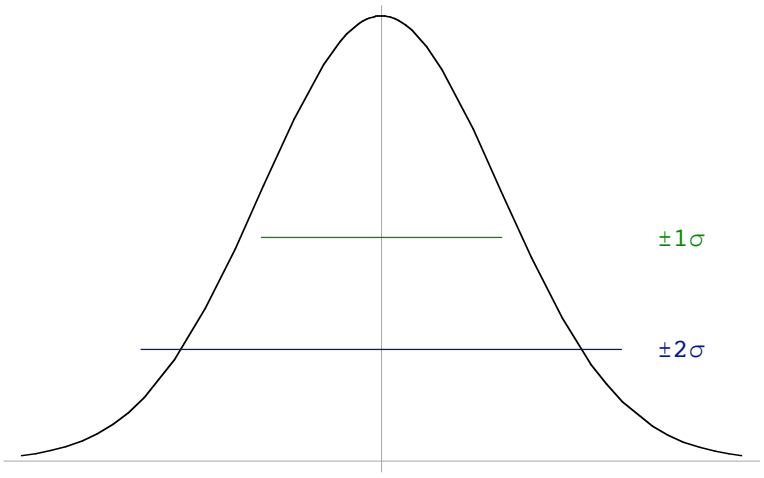
The real power of the standard deviation is in talking about numbers that follow a Normal Distribution, also called a Bell Curve.

In that case, about 2/3's of all the numbers are within 1 σ of the mean, while about 95% of all the numbers are within 2 σ. This is important enough to have a formula line.

$$\begin{array}{ll} m \pm 1\,\sigma & \text{contains 68 \% of normally distributed data} \\ m \pm 2\,\sigma & \text{contains 95.5 \% of normally distributed data} \end{array} \tag{4}$$

Here's another histogram. Even though it doesn't have red bars, the ideas is still the same : the x axis shows a data value, while the y axis represents how many of the data points have that value, just like the plot at the top of this page.

```
<< Statistics`ContinuousDistributions` ;
ndist = NormalDistribution[0, 1];
normPDF = PDF[ndist, xx];
DisplayTogether[
  Plot[normPDF, {xx, -3, 3},
   {Ticks → None, AxesStyle → {RGBColor[.6, .6, .6]}}],
  Graphics[{ RGBColor[0, 0, .5], Text["±2σ", {2.5, .1}],
    Line[{{-2, .1}, {+2, .1}}]}],
  Graphics[{ RGBColor[0, .5, 0], Text["±1σ", {2.5, .2}],
    Line[{{-1, .2}, {+1, .2}}]}]
 ];
Null
```



the Standard Deviation of the Mean

This is all very nice, I hear you saying - but what if my data don't follow a Normal Distribution? After all, the numbers that we started with (0 to 100 with equal probability) aren't normal.

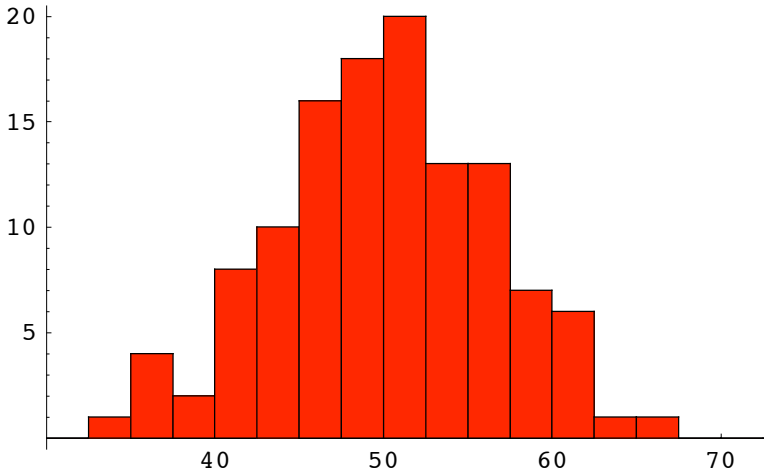
Well, no, they're not. BUT what if that "experirment" (picking 20 numbers and taking the mean) is done over and over again? What does the distribution of the means look like?

You might guess that the mean would be near 50 each time, half way between 0 and 100. And you'd be right. OK, how near? And what is the shape of the distribution?

Here are the numbers you might get if you did this "experiment" 120 time, along with the corresponding histogram.

```
nTrials = 120;
experiment :=
  Mean[Table[Random[Integer, {minNumber, maxNumber}], {howMany}]];
data = N[Table[experiment, {nTrials}]]
Histogram[data, HistogramRange → {30, 70}];
```

```
{48.65, 42.05, 43.35, 52.3, 35.5, 51.1, 48.95, 49., 42.6, 49.15,
 41.35, 43.95, 46.4, 43.4, 64.8, 60.55, 52.25, 48.35, 53.15, 52.05,
 42.7, 50.3, 48.25, 56.45, 52.6, 49.9, 58.7, 45.35, 53.45, 40.3,
 48.75, 40.2, 55.6, 55.55, 51.35, 51.1, 43.2, 57.05, 56.05, 54.2,
 55., 57.4, 47.25, 46.2, 54.75, 50.05, 50.75, 46.45, 60.1, 55.7,
 57.5, 50.55, 45.25, 59.9, 45.7, 50., 54.25, 44.1, 50.95, 46.4,
 59.7, 48.4, 57.15, 45.15, 57.45, 53.1, 61.2, 49.1, 56.8, 56.65,
 44.9, 32.8, 48.7, 61.45, 60.3, 53.15, 47.95, 52.85, 58.45, 35.15,
 47., 50.65, 40.55, 49.65, 53.35, 44., 47.75, 47.2, 35.8, 36.35,
 46.8, 50.1, 52.35, 47.05, 38.9, 53.75, 41.45, 52.2, 46.1, 52.,
 58.85, 53.8, 51.05, 61.25, 42.65, 66.9, 56.55, 46.1, 40.55, 49.55,
 57.55, 41.25, 53.85, 50.75, 46.35, 49.05, 39.85, 47.6, 52.2, 48.7}
```



For the mean and standard deviation of these "experiments" we get

```
Mean[data]
```

```
49.9667
```

```
StandardDeviationMLE[data]
```

```
6.56957
```

These "experiments" give data which is starting to look like a Normal Distribution. This is in fact a general rule : the distribution of the mean of almost anything is normally distributed if you do it enough times.

This data has has a mean near 50 and a standard deviation a lot smaller than the one from the original random numbers. How much smaller?

Well, there's a formula for that which works out like this :

$$\sigma_{\text{mean of n\_trials}} \simeq \frac{\sigma_{\text{once}}}{\sqrt{\text{n\_trials}}} \tag{5}$$

For the experiment just performed, $\sigma_{\text{once}} = 25$ or so (see the top of this page), and n_trials = 120,
so $\sigma_{\text{once}} \,/\, \sqrt{\text{n\_trials}} = 25/\sqrt{120} = 2.8$ which is in the same ballpark as 6.5 - even if they aren't quite the same.

Anytime you're doing an experiment over and over to beat down the random errors and get better accuracy, this is what you're up against: to get 10 times better accuracy, you need to do it 100 times more often. (Gulp.)

Experimental Measurements

Now for the punch line : how would you report the results of an experiment like this in a scientific journal? The language would look something like this.

> In an experiment to measure the mean of 20 random numbers from 0 to 100,
> we find that the result is 49.97 $\pm$ 13.2 at the 2σ ($p < 0.05$) confidence level.

Saying "2σ" is the same as saying"$p < 0.05$"; in both cases, you're claiming that if someone else chooses 20 random numbers from 0 to 100 and takes their mean, then 95% of the time they'll get the same result you did, namely a number within 13.2 of 49.97. The scientific literature uses these sorts of phrases interchangeably.

So ... are we having fun yet?