First name: Roland
Last name: Angerer
Date: 16-12-2003
Homework number: 1
Homework title: Exercise 1.27

Problem Description

Give a detailed explanation of the numerical inferiority of the one-pass formula for computing the standard deviation compared with the two-pass formula given in Example 1.16.

Example 1.16 Standard Deviation. The *mean* of a finite sequence of real values $x_i, i = 1, \ldots, n$, is defined by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the *standard deviation* is defined by

$$\sigma = \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}$$

Use of these formulas requires two passes through the data: one to compute the mean and another to compute the standard deviation. For better efficiency, it is tempting to use the mathematically equivalent formula

$$\sigma = \left[\frac{1}{n-1} (\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) \right]^{1/2}$$

to compute the standard deviation, since both the sum and the sum of squares can be computed in a single pass through the data.

Unfortunately, the single cancellation at the end of the one-pass formula is often much more damaging numerically than all of the cancellations in the two-pass forumla combined. The problem is that the two quantities being subtracted in the one-pass formula are apt to be relatively large and nearly equal, and hence the relative error in the difference may be large (indeed, the result can even be negative, causing the square root to fail).

Problem Solution

In the two-pass formula only the difference between a number and the mean is squared (always positive), while in the one-pass formula each value and the mean are squared, resulting in larger numbers (and errors).

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(\bar{x} + \Delta x_i - \bar{x})^2$$
$$= \sum_{i=1}^{n} \Delta x_i^2$$
$$\sum_{i=1}^{n}(x_i^2 - n\bar{x}^2) = \sum_{i=1}^{n}(\bar{x} + \Delta x_i)^2 - b\bar{x})^2$$
$$= \sum_{i=1}^{n}(\bar{x}^2 + 2\bar{x}\Delta x_i + \Delta x_i^2) - n\bar{x}^2$$
$$= \sum_{i=1}^{n}(2\bar{x}\Delta x_i + \Delta x_i^2)$$

Results

I varified the theoretical conclusions with a Maple script ...

```
> with(stats):

// Generate random numbers and display Maple standard deviation
> generator := rand(10**6-100..10**6+100):
> nb := [seq(generator(), i=1..100)]:
> describe[standarddeviation](nb);
```

$\frac{1}{25}\sqrt{2081614}$

```
// Calculate the mean of the random numbers
> mean := 0:
> for i from 1 to 100 do
>    mean := mean + nb[i];
> end do:
> mean := evalf(mean / 100);
```

mean := .100000376010^7

```
// Calculate the standard deviation according to the two-pass formula
> std1 := 0:
> for i from 1 to 100 do
>    std1 := std1 + (nb[i] - mean)**2;
> end do:
> std1 := sqrt(std1/99);
```

std1 := 58.00193657

```
// Calculate the standard deviation according to the one-pass formula
> std2 := 0:
> for i from 1 to 100 do
>    std2 := std2 + nb[i]**2;
> end do:
> std2 := sqrt((std2 - 100*mean**2)/99);
```

std2 := 54.77225575

```
/* Calculate the difference (if any) between two-pass formula and
```

```
one-pass formula for the standard deviation */
> std1 - std2;
3.22968082
```

Discussion and Comments

Proof for the mathematical correctness:

$$
\begin{aligned}
\sigma &= \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2} = \\
&= \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)\right]^{1/2} = \\
&= \left[\frac{1}{n-1}\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\bar{x}+\sum_{i=1}^{n}\bar{x}^2)\right]^{1/2} = \\
&= \left[\frac{1}{n-1}\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\right]^{1/2} = \\
&= \left[\frac{1}{n-1}\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right]^{1/2}
\end{aligned}
$$