FP Solutions

A1 Question 1

```
function b = randfp(t, L, U)

    % random sign
    blah = rand;
    if blah>0.5
        b = '+0.1';
    else
        b = '-0.1';
    end

    % t mantissa digits
    for k = 2:t
        % randomly choose t mantissa digits
        blah = rand;
        if blah>0.5
            b = [b '1'];
        else
            b = [b '0'];
        end
    end

    % randomly choose the exponent in the
    % exponent range from L to U
    p = randi(U-L+1)+L-1;
    b = [b 'b' num2str(p)];
```

A1 Question 2

```
function x = fp2dec(b)

    % Find where the "exponent" character is
    e_place = strfind(b,'b');

    % extract the mantissa
    b_mant = b(4:e_place-1);

    r = 0.;
    % accumulate each digit into the value
    for k = 1:length(b_mant)
       if b_mant(k)=='1'
          r = r + 2^(-k);
       end
    end

    %  extract and use the exponent
    p = str2num(b(e_place+1:end));

    x = r * 2^p;

    % sign
    if b(1)=='-'
       x = -x;
    end
```

A1 Question 3

(3) a) Largest value in $\mathbb{F}(7, 4, -8, 8)$.

$+ 0.6666 \times 7^{8}$ ✓

b) $fl(26535.1_7) \otimes fl(10000_7)$

$= 26540_7 \otimes 10000_7$

$= fl(265400000_7)$ ✓

$= fl(0.2654_7 \times 7^9) =$ "overflow" or "Inf"

c) Machine epsilon (these values can be base-10)

$$E = \frac{1}{2}\beta^{1-t} = \frac{1}{2}7^{1-4} = \frac{1}{2}7^{-3} = \frac{1}{2} \times 0.001_7$$

In base-10, $E = 0.00146$ ✓

d) All the values smaller than 1 have the form

$$\pm 0.d_1d_2d_3d_4 \times 7^{p} \quad \text{where } p \leq 0.$$

The mantissa can be any normalized mantissa. What matters is the exponent, p.

$p \in \mathbb{Z}$, $-8 \leq p \leq 0$ yields values < 1

$p \in \mathbb{Z}$, $-8 \leq p \leq 0$ yields values < 1

$0 < p \leq 8$ yields values ≥ 1

Thus, of the 17 possible exponents, 9 of them yield values < 1. ✓

$\therefore \frac{9}{17}$ or 53% are smaller than 1.

A1 Question 4: v2

Prove that $RelErr(ab+c) \leq \frac{|ab|}{|ab+c|} E(1+E) + E$

$$Err = |(a \otimes b) \oplus c - (ab+c)|$$

$$= |ab(1+\delta_1) \oplus c - (ab+c)|$$

$$= |(ab(1+\delta_1)+c)(1+\delta_2) - (ab+c)|$$

$$= |ab(1+\delta_1) + \cancel{c} + ab(1+\delta_1)\delta_2 + c\delta_2 - ab - \cancel{c}|$$

$$= |\cancel{ab} + \delta_1 ab + ab(1+\delta_1)\delta_2 + c\delta_2 - \cancel{ab}|$$

$$= |ab\,\delta_1 + ab\,\delta_2 + ab\,\delta_1\delta_2 + c\delta_2|$$

$$= |ab\,\delta_1(1+\delta_2) + \delta_2(ab+c)|$$

$$\leq |ab\,\delta_1(1+\delta_2)| + |\delta_2(ab+c)| \quad \Delta \text{ inequality}$$

$$\leq |ab|\,E(1+E) + |ab+c|\,E \quad \text{since } |\delta_i| \leq E$$

Thus,

$$RelErr \leq \frac{|ab|}{|ab+c|} E(1+E) + E$$

QED

A1 Forensic Siphoning

a) Method C is the most accurate. It sorts the transactions and adds one credit and one debit, a pair at a time.

b) This method is more accurate because it avoids computing with large values. Since most of the credits and debits cancel out, the net income is much smaller than the possible intermediate values. Sorting the transactions and including them in credit/debit pairs avoids large values.

c) Yes, it seems a crime is being committed. The function returns the smallest net income of the three methods. Two of the methods have the potential for a large cancellation error, so the value returned may be substantially less than the real answer.