36-201 Spring 1999
Solutions to Homework 1

1. `Moore, 4.20 (p.  231).` Find the five-number summary and make a simple boxplot of the same data. List separately outliers identified with the 1.5*IQR rule, if any.

```
MINITAB STEM PLOT (number of calories in hotdogs):

     1    10 7
     1    11
     1    12
     5    13 5689
     8    14 067
    (1)   15 3
     8    16
     8    17 2359
     4    18 2
     3    19 015
```

The shape is possibly bimodal, or just irregular. The question asked "which observation was the 'EAT SLIM VEAL DOGS'?" This must be the 107 (the min of the data).

```
MINITAB BOX PLOT (not required):

                       --------------------------
   --------------------I        +               I----------
                       --------------------------

  ----+---------+---------+---------+---------+---------+--
     112       128       144       160       176       192

FIVE NUMBER SUMMARY, FENCES

Min            Q1          Med         Q3            Max
107            139         153         179           195

IQR = Q3 - Q1 = 40

Lower fence = Q1 - 1.5*(IQR) = 139 - 1.5 * 40 = 79
Upper fence = Q3 + 1.5*(IQR) = 179 + 1.5 * 40 = 239

No data falls outside the fences --> No outliers [just as in the boxplot]!

(Note: Minitab's 5-number summary:

 MEDIAN    MIN       MAX       Q1        Q3
 153.00    107.00    195.00    138.50    180.50
```

```
gives slightly different results, but still no outliers.

)
```

2. `Moore, 4.33 (p. 243).` The "average" household wealth of readers of Forbes magazine ni 1995 was either \$800,000 or \$2.2 million, depending on which "average" was used.

   Money typically has a skewed-right distribution, so $MEAN > MEDIAN$. I suspect that \$800,000 is the median and \$2.2 million is the mean.

   If you thought money had a different-shaped distribution, or you thought there could be more than one shape, we will accept your answers as long as they are tied to the shapes you thought about.

3. `Moore, 4.26 (p. 235), 4.34 (p. 243), and 4.49 (p. 256)`

```
      RUTH   MARIS

           0 8
           1 346
        52 2 368
        54 3 39
   9766611 4
       944 5
         0 6 1
```

The distribution of home runs per season for Ruth is shifted up (higher values) from the distribution of Maris. Except for the record breaking 61 homers, only the bottom four years of Ruth's data overlap Maris' at all.

```
BEFORE REMOVING THE 60 AND 61        AFTER REMOVING THE 60 AND 61

Mean(Ruth)     =  43.93              Mean(Ruth)     = 42.79
Mean(Maris)    =  26.1               Mean(Maris)    = 22.2

Median(Ruth)   =  46                 Median(Ruth)   = 46
Median(Maris)  =  24.5               Median(Maris)  = 23
```

The difference in Maris's means before and after removing the 61 is $26.1 - 22.2 = 3.9$; the difference in his medians is $24.5 - 23 = 1.5$. The reduction in the mean is greater since means are *less robust against* (or *more sensitive to*) changes in extreme data.

The difference in Ruth's means is $43.93 - 42.79 = 1.14$; the difference in his medians is $46 - 46 = 0$.

- The difference in Ruth's means is smaller than Maris's because *means are more sensitive to more extreme data* (or *less sensitive to less extreme data*) and the 60 is not as extreme for Ruth, as the 61 is for Maris.
- It's less clear to me why the difference in the Ruth's medians is less, but it helps that Ruth has more observations than Maris in about the same range of scores, so more similar observations pile up around the middle of the sorted observations.

Here are Minitab's 5-number summaries, and side-by-side boxplots:

```
MTB > desc c1-c2

                  N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
ruth             15    43.93    46.00    44.38    11.25     2.90
maris             9    26.44    26.00    26.44    16.52     5.51

                MIN      MAX       Q1       Q3
ruth          22.00    60.00    35.00    54.00
maris          8.00    61.00    14.00    33.00

MTB > stack c1-c2 c3;
SUBC> subs c4.
MTB > boxp c3;
SUBC> by c4.

C4

                                    ---------------
1                    ----------------I       +     I--------
                                    ---------------

            --------------------
2     ------I           +      I-----------------------------
            --------------------
    ----+---------+---------+---------+---------+---------+--C3
       10        20        30        40        50        60
```

Note that 75% of Ruth's data (top boxplot) is greater than 75% of Maris' data, and almost all of Ruth's data is above the lower half of Maris' data. Thus Ruth typically had better years than Maris. The narrower IQR (width of box) for Ruth than for Maris also suggests that Ruth was more consistent year-to-year than Maris.

4. `Moore, 4.53 (p.  257) Not everything about this question is a calculation!`

   Weight gains of chicks, on old vs. new corn:

```
MTB > set c1
DATA> 380 321 366 356
DATA> 283 349 402 462
DATA> 356 410 329 399
DATA> 350 384 316 272
DATA> 345 455 360 431
DATA> end
MTB > name c1 'old corn'
MTB > set c2
DATA> 361 447 401 375
DATA> 434 403 393 426
DATA> 406 318 467 407
```

```
DATA> 427 420 477 392
DATA> 430 339 410 326
DATA> end
MTB > name c2 'new corn'
MTB > desc c1-c2

                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
old corn       20    366.3    358.0    366.2     50.8     11.4
new corn       20   402.95   406.50   403.56    42.73     9.55

              MIN      MAX       Q1       Q3
old corn    272.0    462.0    333.0    401.2
new corn   318.00   477.00   379.25   429.25

MTB > stack c1-c2 c3;
SUBC> subs c4.
MTB > boxp c3;
SUBC> by c4.

C4

                            -----------------
1           ----------------I     +         I----------------
                            -----------------

                                        ------------
2                       ----------------I     +    I------------
                                        ------------
          ----+---------+---------+---------+---------+---------+--C3
            280       320       360       400       440       480
```

(a) Weight gain on the new corn (lower boxplot) does appear to be better overall than on the old corn (upper boxplot). The difference is not as striking as for the Ruth vs. Maris data, but at least 75% of the new corn weight gains were better than 50% of the old corn weight gains. A typical chick would gain around 360 grams on the old corn (mean = 366, median = 358) and around 405 grams (mean = 403, median = 406.5) on the new corn.

(Note that the mean and median are about the same for both groups, since both data sets are approximately symmetric; see part (c) below for more details).

(b) We need to know more about this experiment before we can conclude that the change in corn caused the greater weight gains in the new corn chicks. Some things to look for are

- Maybe the new corn chicks were of a faster-growing breed than the old-corn chicks;
- Maybe the people taking care of the new-corn chicks knew they were on the new corn and subconsciously did other things, like keep them warmer or give them a bit more water, that would help them grow faster, even under the old corn.
- Etc. (there are many possibilities)

When we design an experiment we try to design it in such a way that influences such as those I've listed above can be ruled out. If we are successful at this, then any differences between the new corn and old corn chicks *can* be said to be caused by the change in corn.

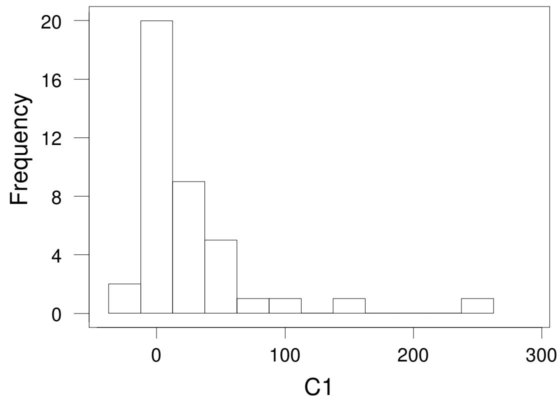
(c) The old corn (upper boxplot) data looks approximately symmetric, with just a little skew suggested in the middle 50% of the data. The new corn data (lower boxplot) also looks reasonably symmetric, though the whiskers (tails) suggest some mild skewing to the left. The IQR (width of the box, variability) is a little less for the new corn chicks than the old corn chicks. You can also see this smaller variability in the lower SD for new corn than old corn, in the MINITAB descrptive statistics above.

5. `Percent differences between measured and reported calories, per item:`

```
 2.0  -28.0   -6.0    8.0    6.0   -1.0   10.0   13.0   15.0  -4.0
-4.0  -18.0   10.0    5.0    3.0   -7.0    3.0   -0.5  -10.0   6.0
41.0   46.0    2.0   25.0   39.0   16.5   17.0   28.0   -3.0  14.0
34.0   42.0   15.0   60.0  250.0  145.0    6.0   80.0   95.0   3.0
```

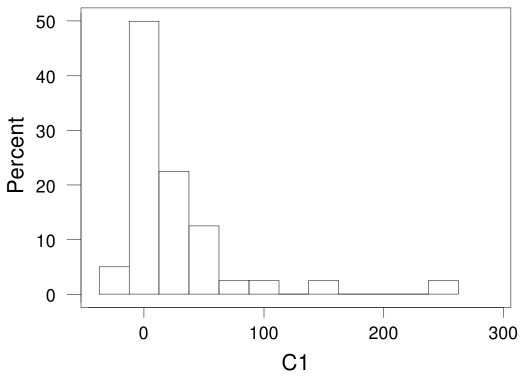
`Use this data in Minitab to print out various graphical summaries, and say which is easiest to interpret.`

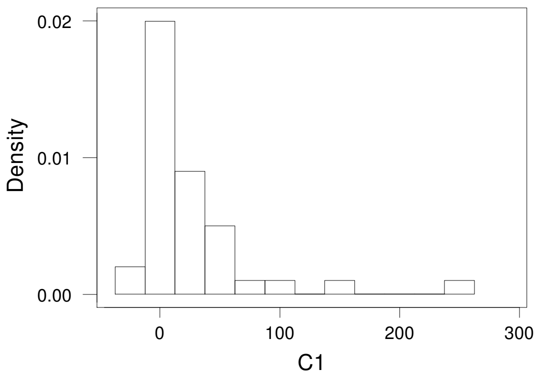
(i) Frequency Histogram.



(ii) Percent Histogram (change ”options” when you make the histogram).

There are two “percent” histograms listed in Minitab under the “options” button. I prefer the “density” one for reasons I outlined in class, but either will do for this assignment.





(iii) Boxplot.

```
MTB > boxp c1


              -------
        ------I +   I------   *  *         O                    O
              -------
        ------+---------+---------+---------+---------+---------+C1
              0        50       100       150       200       250
```

(iv) Stem and Leaf Plot.

```
MTB > stem c1

Stem-and-leaf of C1        N  = 40
Leaf Unit = 10


    1   -0 2
   10   -0 110000000
  (18)   0 000000000011111111
   12    0 2233
    8    0 444
    5    0 6
    4    0 89
    2    1
    2    1
    2    1 4
    1    1
    1    1
    1    2
    1    2
    1    2 5
```

(iv) Dotplot.

```
MTB > dotp c1


              .
            . : :
            ::::.:    .
       . . .::::: ...:.  .   .  .         .                    .
        -----+---------+---------+---------+---------+---------+-C1
             0        50       100       150       200       250
```

All the plots tell the same story—the data is extremely skewed right with some very high outliers. The stem and leaf plot gives the shape of the data and I can recover the original observations from it, so that seems very useful. The boxplot automatically identifies outliers, which is also nice. [Any reasoned answer is fine here.]