Stem and Leaf

```
MTB > set c1
DATA> 42 43 55 26 62 37 33 41 19 54 20 85
DATA> 46 10 17 60 53 42 37 42 55 28 48
DATA> end
MTB > set c2
DATA> 24 43 58 80 43 49 61 44 67 49 53 56
DATA> 59 52 62 54 57 33 46 43 57
DATA> end
MTB > name c1 'control'
MTB > name c2 'trtment'
MTB > stemplot c1

    3    1 079
    6    2 068
    9    3 377
   (7)   4 1222368
    7    5 3455
    3    6 02
    1    7
    1    8 5
```

```
MTB > stemplot c2

    1    2 4
    1    2
    2    3 3
    2    3
    6    4 3334
    9    4 699
   (3)   5 234
    9    5 67789
    4    6 12
    2    6 7
    1    7
    1    7
    1    8 0

MTB > stemplot c2;
SUBC> increment by 10.

    1    2 4
    2    3 3
    9    4 3334699
   (8)   5 23467789
    4    6 127
    1    7
    1    8 0
```

Measures of location

n: number of items in the list
Rank: position of a number in a sorted list
MIN: (minimum) The smallest number (number at rank 1)
MAX: (maximum) The largest number (number at rank n)
Median: Divides the data in half; number at rank $(n+1)/2$ (or average of two closest ones)
Quartiles: Divides the data into quarters

- Q1 is median of lower half of list (including median if it's data)
- Q3 is median of upper half of list (including median if it's data)

(fancier formulas on the computer)

Measures of spread/variability

Range: subtract min from max:
$Range = MAX - MIN.$

IQR: (interquartile range) Subtract Q1 from Q3:
$IQR = Q3 - Q1.$

Five-number summary

MIN Q1 Median Q3 MAX

```
MTB > describe c1-c2

                  N     MEAN   MEDIAN   TRMEAN
control          23    41.52    42.00    40.95
trtment          21    51.90    53.00    51.89

                SEMEAN      MIN      MAX       Q1
control           3.58    10.00    85.00    28.00
 trtment          2.61    24.00    80.00    43.50
```

Other measures of Center

Mean: Average of all the observations

$$(42 + 43 + 55 + 26 + \cdots + 37 + 42 + 55 + 28 + 48)/23 = 41.52$$

$$(24 + 43 + 58 + 80 + \cdots + 33 + 46 + 43 + 57)/21 = 51.90$$

Median: _____

Mode: Most frequent observation.

- More useful for categorical (qualitative) data (male/female, Democrat/Republican/ Wrestler, etc.)
- Sometimes useful for continuous (quantitative) data (height, weight, test score), when there is a lot of data and a clearly-defined mode (more later).

Box Plots

```
MTB > boxplot c1

                    -----------------
    -------------I       +       I--------------------
                    -----------------
  ------+---------+---------+---------+---------+---------+------
        15        30        45        60        75

MTB > boxplot c2

                         ---------------
  ---------------------I        +    I---------
                         ---------------
  ------+---------+---------+---------+---------+---------+------
        30        40        50        60        70
```

Outliers

Lower fence: $Q1 - 1.5 \times IQR$
Upper fence: $Q3 + 1.5 \times IQR$

- Data outside the fences are outliers
- Sometimes 3 is used instead of 1.5

Comparing Distributions

```
MTB > stack c1 c2 c3;
SUBC> subscripts in c4.
MTB > boxplot c3;
SUBC> by c4.

C4

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


                               -----------
2              -------------I     +   I------         *
                               -----------
  ------+---------+---------+---------+---------+---------+------
        15        30        45        60        75
```

BACK TO BACK STEM AND LEAF, BY HAND:

```
          1 079
        4 2 068
        3 3 377
  9964333 4 1222368
 98776432 5 3455
     721 6 02
         7
        0 8 5
```

Histograms

Two Kinds

- Frequency Histograms (Siegel and Morgan)
- Percent Histograms (Moore)

Both useful.

A stem and leaf plot is a tilted histogram

For a frequency histogram

1. Divide the interval from the MIN to the MAX into K equal "class intervals" (say, every 10 units) — these are now the stems
2. Count how many observations fall in each class interval.
3. Make a bar plot showing the counts (instead of listing the data, as in a stem and leaf plot)

For a percent histogram

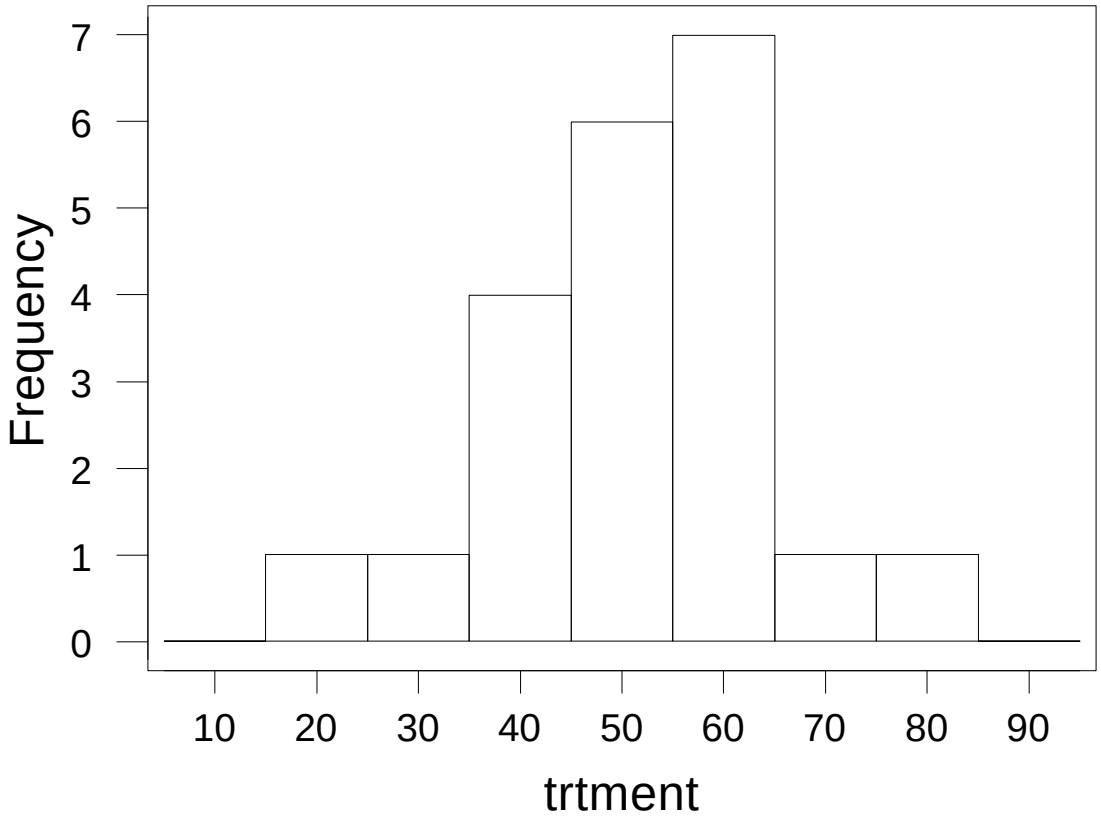
3. Make a bar plot showing the percent density of the counts:

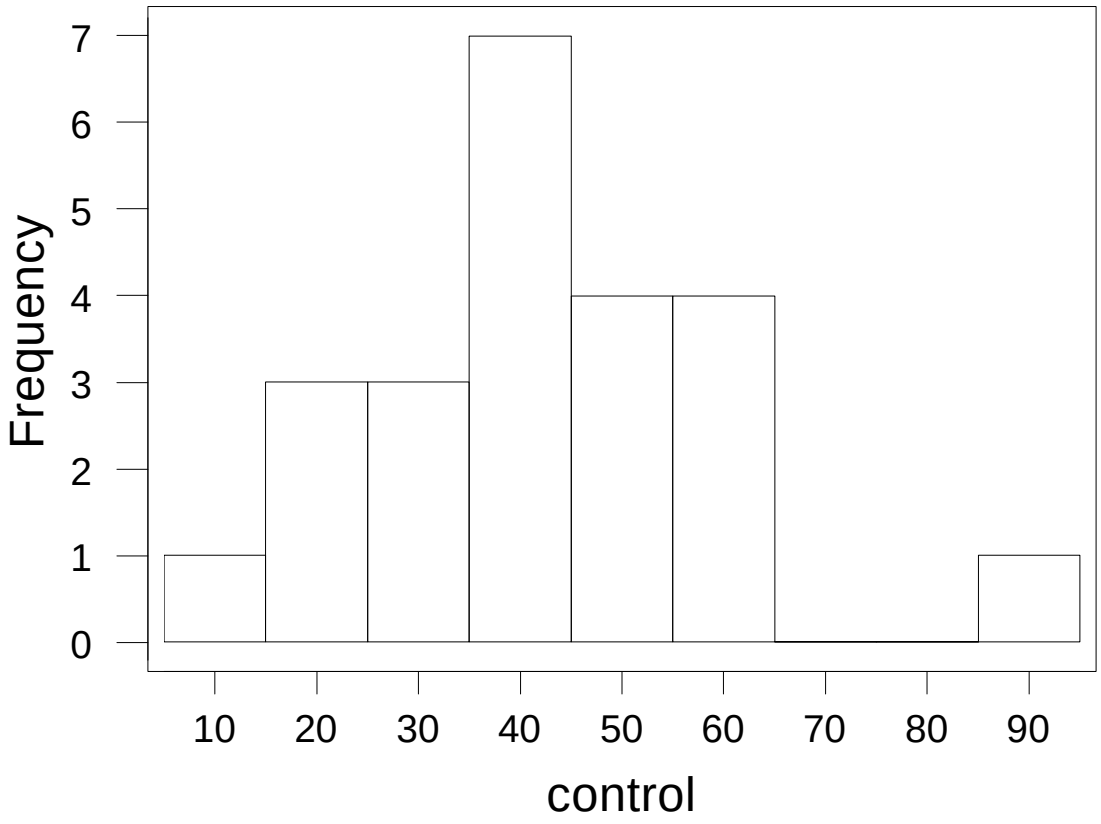
   (a) Divide the class interval counts by the total n to get percents

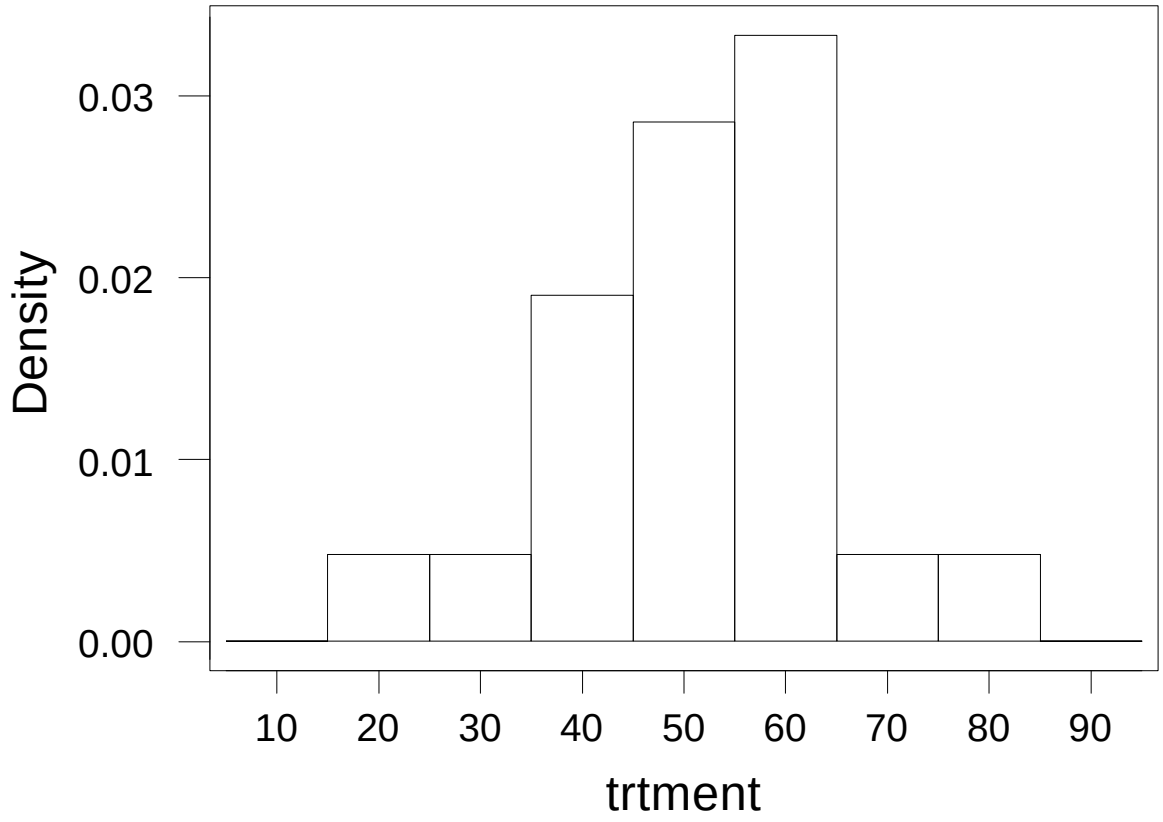
   (b) Divide the percents by the width of the class intervals (divide by 10 if the intervals are every 10 units) to get percent density

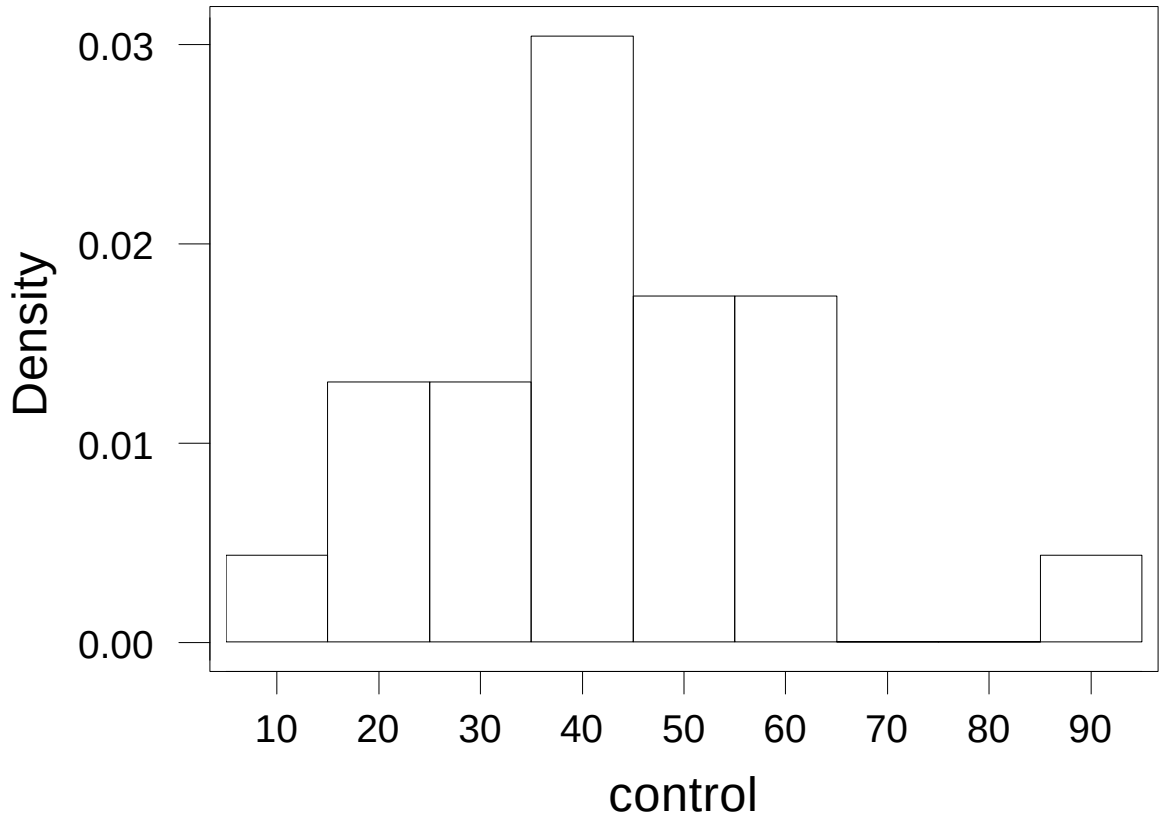
   (c) Make a bar plot of the percent densities

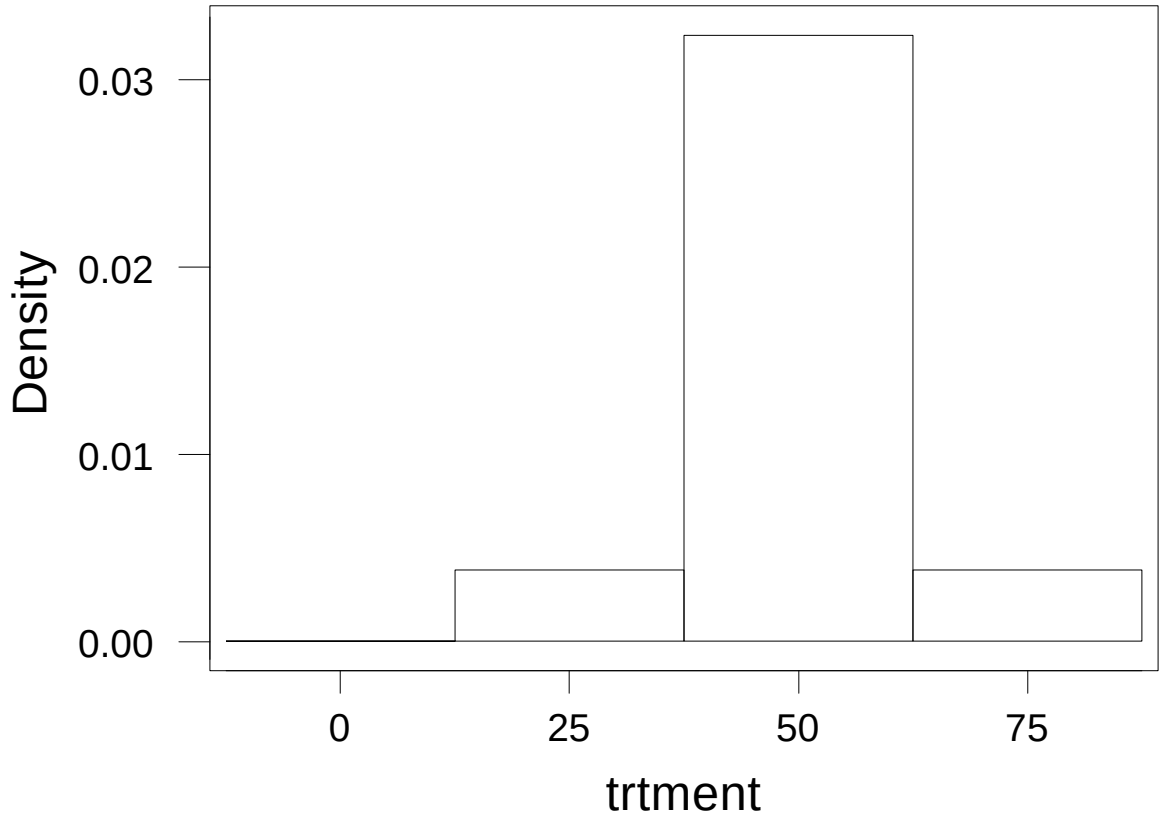
Our eye responds to *area*; now the percents are the *areas* of the bars.

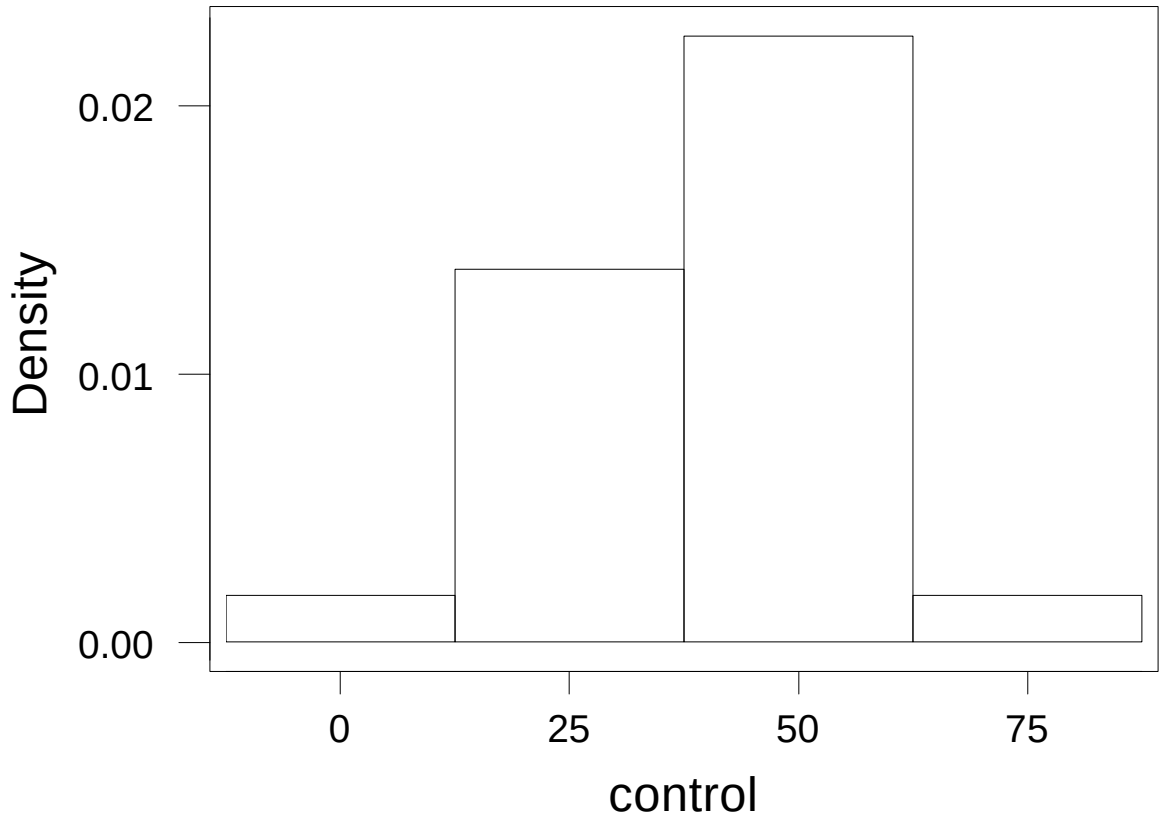

Frequency
0
1
2
3
4
5
6
7
10
20
30
40
50
60
70
80
90
trtment

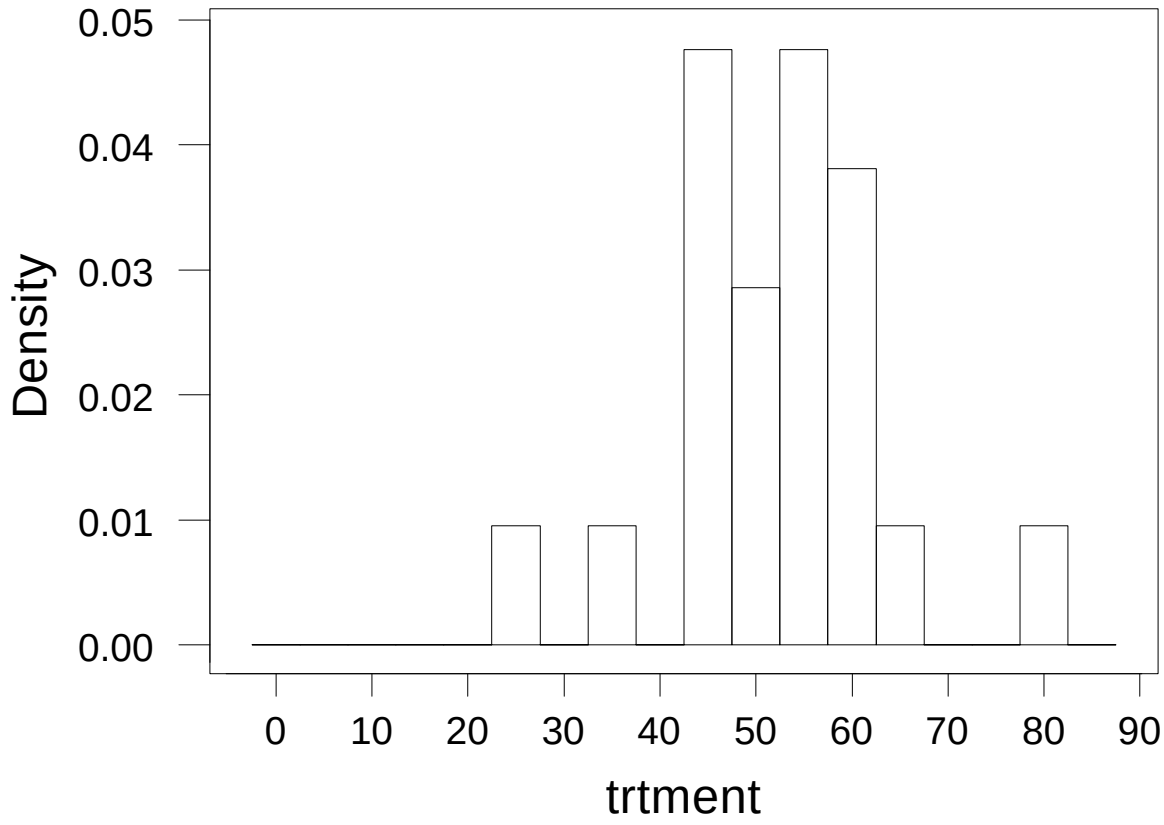

Frequency
0
1
2
3
4
5
6
7
10
20
30
40
50
60
70
80
90
control

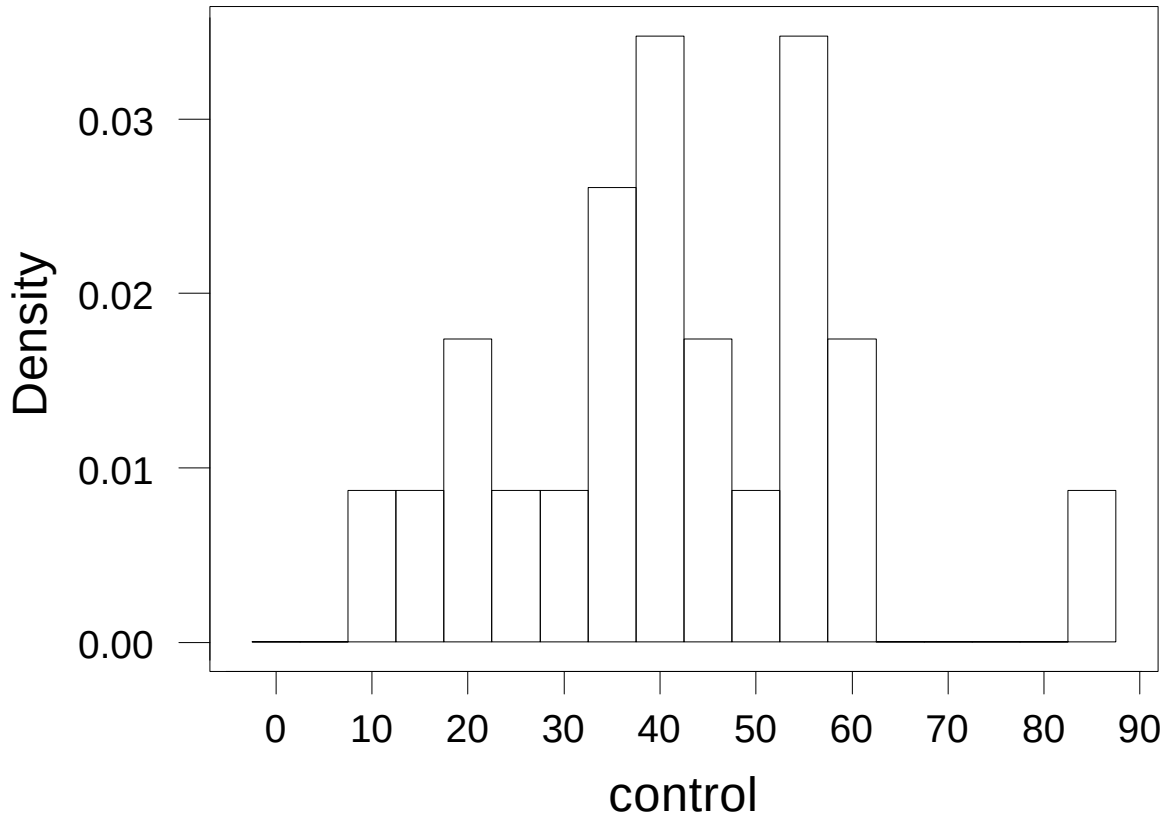

Density
0.03
0.02
0.01
0.00
10
20
30
40
50
60
70
80
90
trtment


Density
0.03
0.02
0.01
0.00
10
20
30
40
50
60
70
80
90
control


0.03
0.02
0.01
0.00
Density
0
25
50
75
trtment


0.02
0.01
0.00
Density
0
25
50
75
control


0.05
0.04
0.03
0.02
0.01
0.00
Density
0
10
20
30
40
50
60
70
80
90
trtment


0.03
0.02
0.01
0.00
Density
0
10
20
30
40
50
60
70
80
90
control

Other kinds of displays

- Pie Charts
- Pictograms

Graphical Clutter

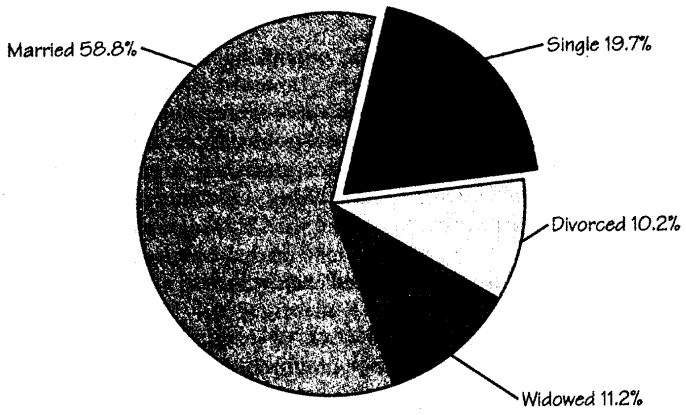


Figure 4-1 Pie chart of the marital status of adult women. The angle spanned by each wedge shows the percent of women having that marital status.

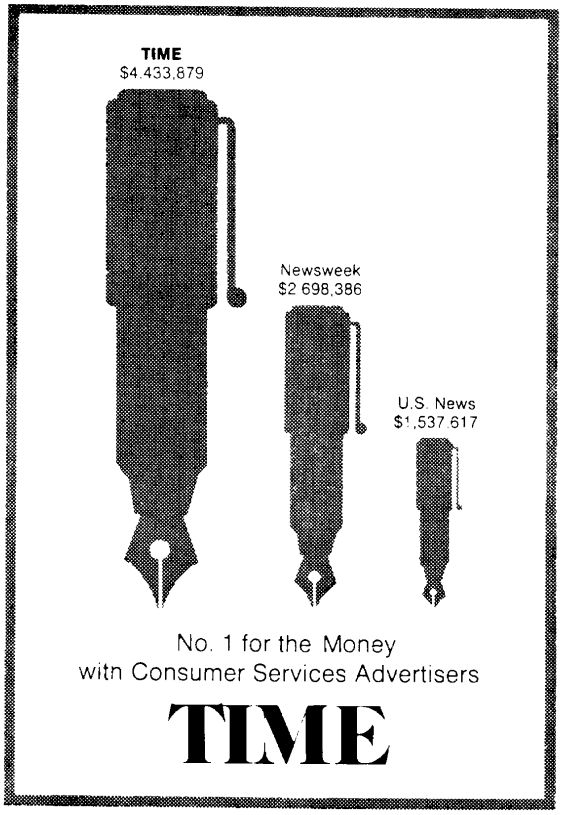


Figure 4-4 A pictogram. This variation of a bar graph is attractive but misleading. [Copyright © 1971 by Time, Inc. Reproduced by permission.]

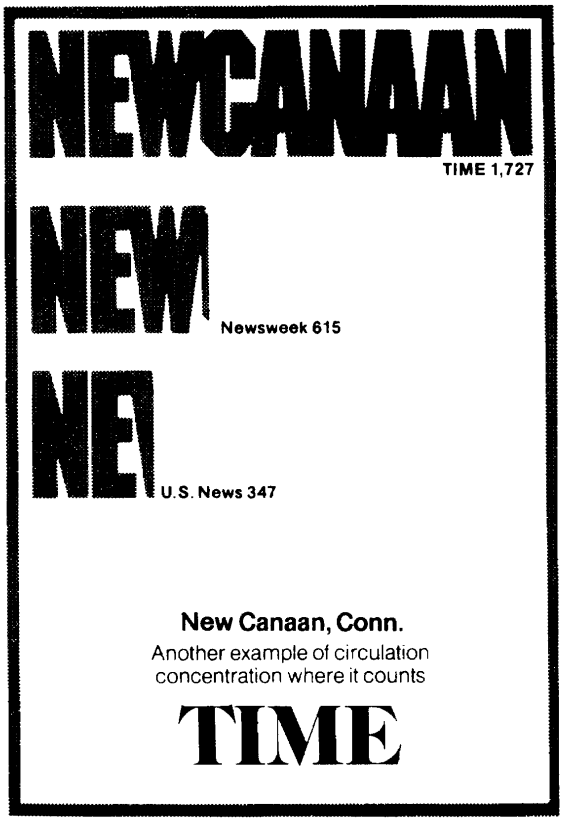


Figure 4-5 An attractive and accurate variation of a bar graph. [Copyright © 1972 by Time, Inc. Reproduced by permission.]