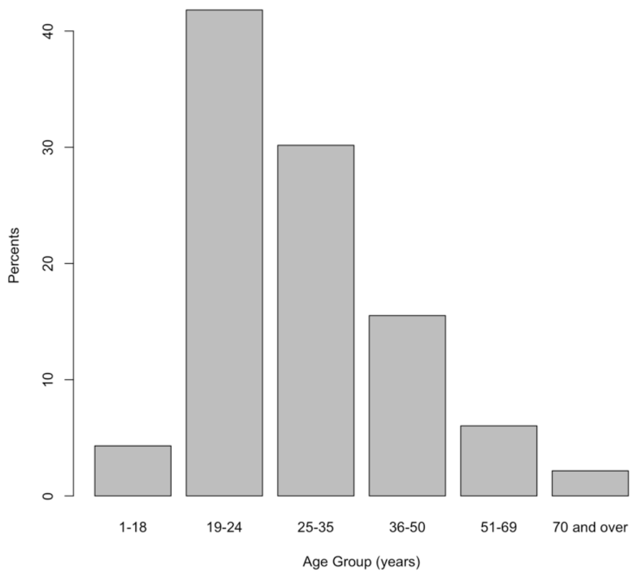
ST 305 HW1 Solution

Ch 1.1

14. (b) (1 point)
label: employee identification number
quantitative variable: number of years with the company, salary, age
categorical variable: last name, first name, middle initial, department, education

21.
(a) (1 point) 10 + 97 + 70 + 36 + 14 + 5 = 232.
Percents for each age group: 10/232 = 4.31%, 97/232 = 41.81%, 70/232 = 30.17%, 36/232 = 15.52%, 14/232 = 6.03%, 5/232 = 2.16%.

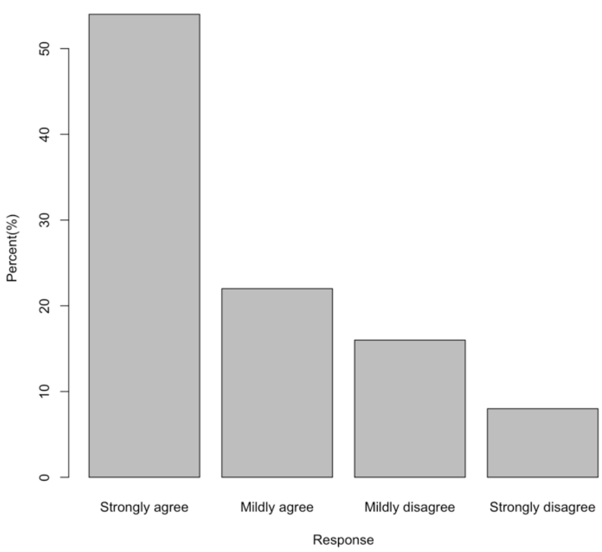
(b) (1 point)



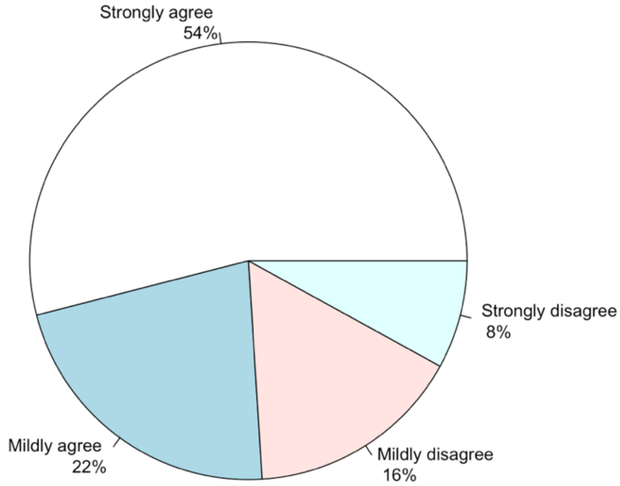
(c) (1 point) For example, 87.5% of the group were between 19 and 50.

(d) (1 point) The age-group classes do not have equal width.

22. (a) (1 point)



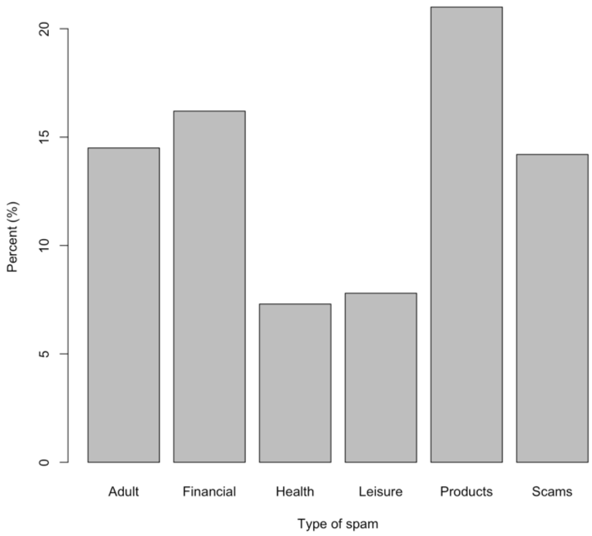
(b) (1 point)

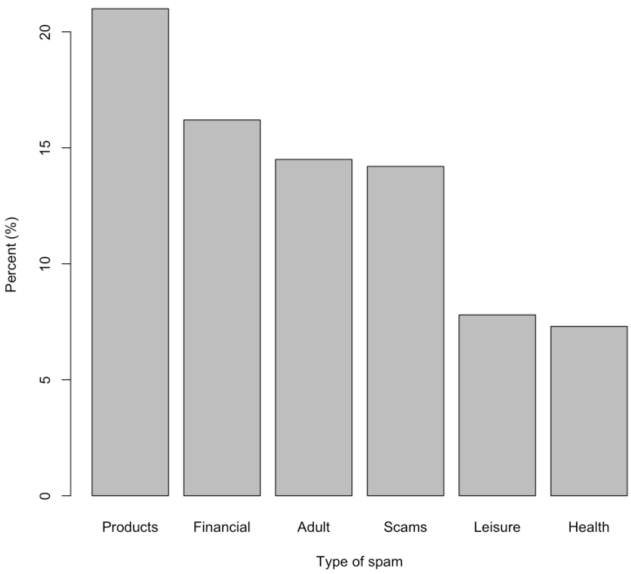


(c) (1 point) More than half of the responded users strongly agree with the statement.

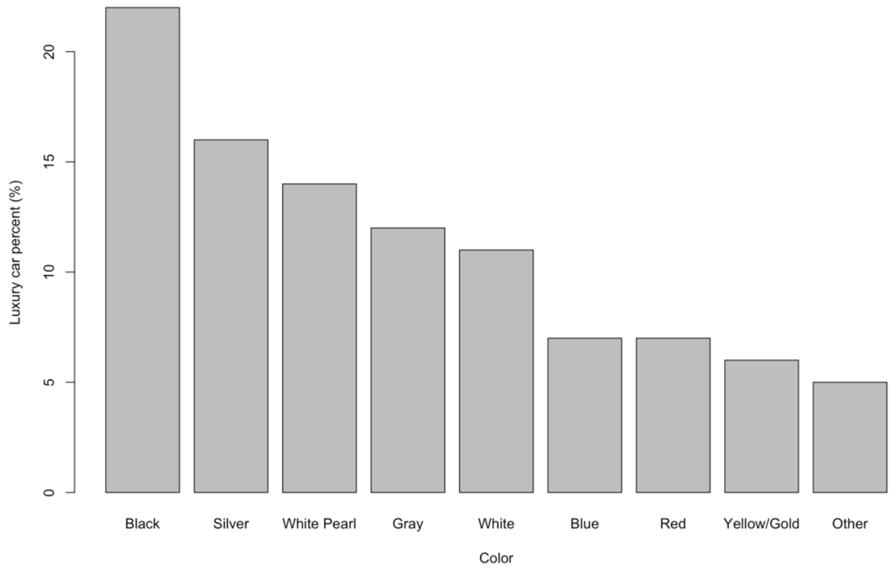
(d) (1 point) I prefer the bar chart for this data set since it is easier to compare the percent of each response.

27. (1 point)

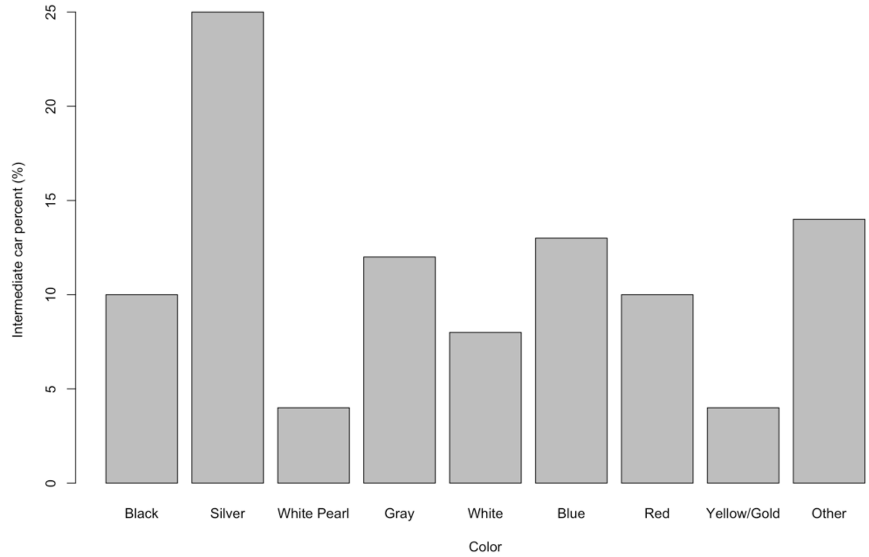




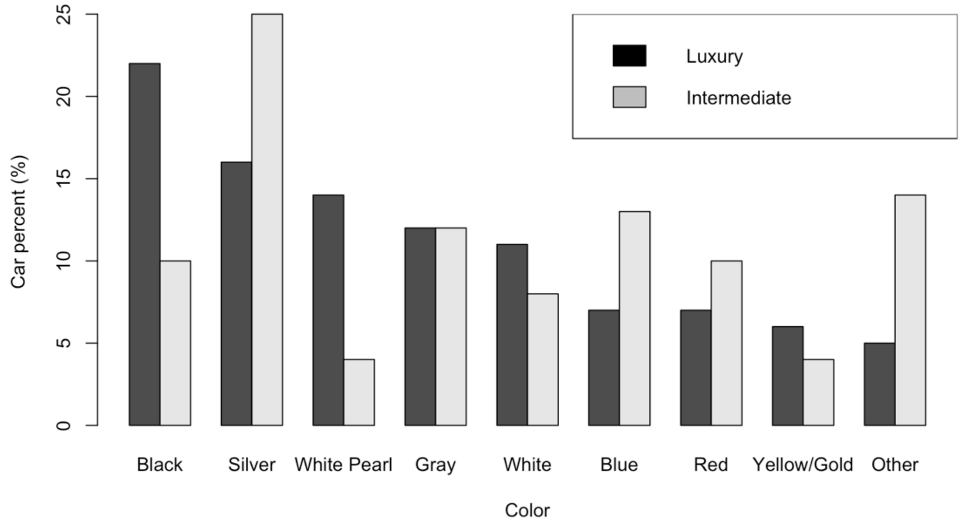
31. (a) (1 point)



(b) (1 point)



(c) (1 point)



33. (1 point)
I trimmed the numbers by removing the last digit and splitted each stem into two.

```
0 | 799
1 | 0134444
1 | 5577
2 | 0
2 | 57
3 |
3 | 5
```

The distribution is skewed to the right. 359 mg/dl is an outlier; only four are in the desired range.

34. (1 point)
I trimmed the numbers by removing the last digit and splitted each stem into two.

```
Individual   Class
          | 0 |
          | 0 |799
        22| 1 |0134444
  99866655| 1 |5577
     22222| 2 |0
         8| 2 |57
          | 3 |
          | 3 |5
```

The distribution of fasting plasma glucose of diabetics in the individual instruction groups is roughly symmetric and there is no outlier. The distribution of fasting plasma glucose of diabetics in the class is skewed to right. By comparison, the fasting plasma glucose of diabetics in the individual instruction group has larger value and smaller spread than that of diabetics in the class.

41. (a) (1 point) Most people will "round" their answers when asked to given an estimate like this, and some way exaggerate. I think the response that he studies zero minute per night is suspicious.

(b) (1 point) I eliminated one student who claimed to study zero minutes per night and trim the numbers by removing the last digit and split each stem into two.

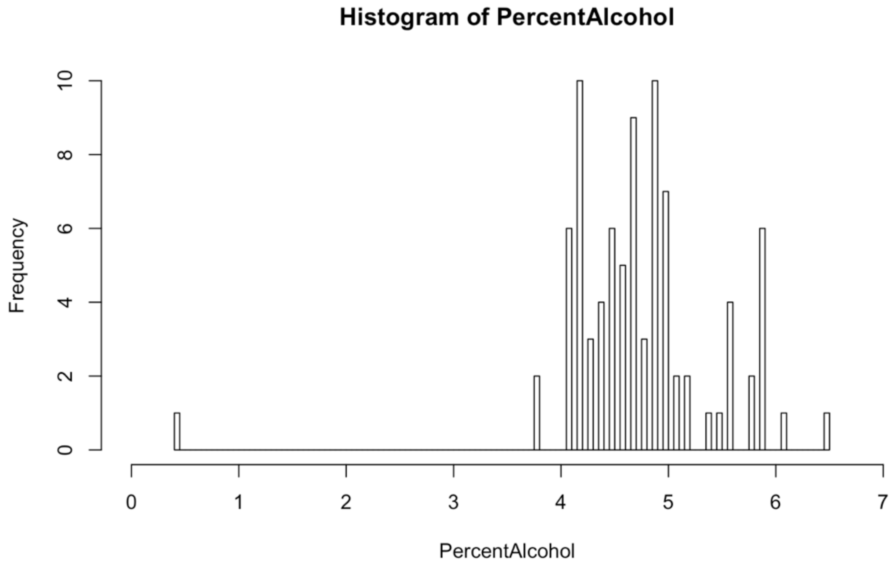
```
             Women      Men
                  | 0 |33334
                96| 0 |66679999
          22222221| 1 |2222222
    88888888875555| 1 |558
              4440| 2 |00344
                  | 2 |
                  | 3 |0
                 6| 3 |
```

The stemplots and midpoints (175 for women, 120 for men) suggest that women (claim to) study more than men.

Ch 1.2

62

(a) (1 point) For graphical summary, I use histogram since percent alcohol is a quantitative variable.



For numerical summary, I choose the five-number summary, which is 0.4, 4.325, 4.7, 5, 6.5.

(b) (1 point) The minimum 0.4 is an outlier. It is a beer of the brand O'Doul's which is famous for non-alcohol.

63. (a) (1 point) $\bar{x}$ changes from 4.76% (with) to 4.81% (without); the median (4.7%) does not change.

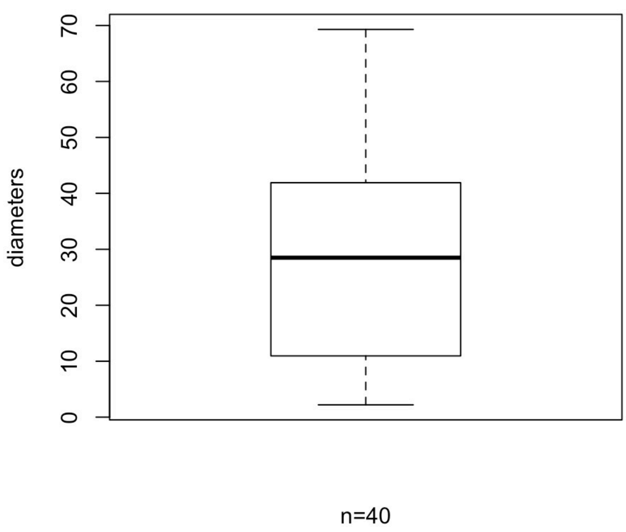
(b) (1 point) s changes from 0.7523% to 0.5864%; Q_1 changes from 4.3% to 4.35%, while $Q_3 = 5\%$ does not change.

(c) (1 point) A low outlier decreases $\bar{x}$; any kind of outlier increases s. Outliers have little or no effect on the median and quartiles.

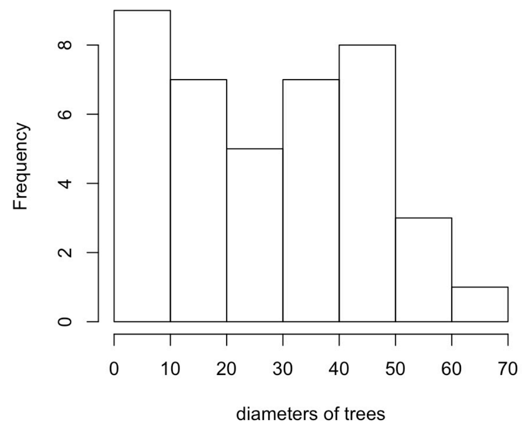
65. (1 point) For example, 0, 1, 2, 998, 1000; the median changes from 2 to 500.

68. (a) (1 point) 2.20, 11.18, 28.50, 41.20, 69.30.

(b) (1 point)

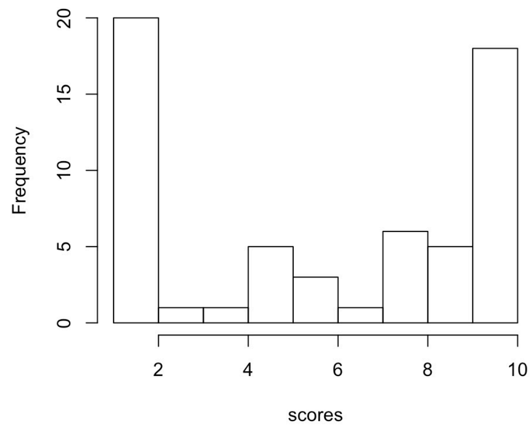


(c) (1 point)



(d) (1 point) The distribution is skewed to right. I prefer the histogram for these data since it reveals much detail about the shape. Boxplots are mainly good when comparing two or more distributions.

72. (1 point)



The five-number summary is 1, 1, 6.5, 10, 10.

The distribution is bimodal.

73. (1 point) This distribution would almost surely be strongly skewed to the right.

78. (1 point) $1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439 = 11200$.
$\bar{x} = 11200/7 = 1600$.
$1792 - 1600 = 192, 1666 - 1600 = 66, 1362 - 1600 = -238, 1614 - 1600 = 14$,
$1460 - 1600 = -140, 1867 - 1600 = 267, 1439 - 1600 = -161$.

$192 + 66 - 238 + 14 - 140 + 267 - 161 = 0$.
$s^2 = [192^2 + 66^2 + (-238)^2 + 14^2 + (-140)^2 + 267^2 + (-161)^2]/6 = 35811.67$.
$s = \sqrt{35811.67} = 189.24$.

99. (1 point) Full data set: $\bar{x} = 196.575$ and $M = 103.5$ minutes. The 10% and 20% trimmed means are $\bar{x}^* \approx 127.734$ and $\bar{x}^{**} \approx 111.917$.

Ch 1.3

114. (a) (1 point) Using the formula of standardized value of x: $z = \frac{x-\mu}{\sigma}$, the standardized scores are $-0.5, -1.6, 2.2, 0.5, 0.3, 2.8, -0.6, -1.5, 1, 0$

(b) (1 point) Using Table A, the cut-off value is 1.04.

(c) (1 point) The students with original scores equal to 92, 98 earned a grade of A.

115. (a) (1 point) -1.64, -1.04, 0.13 and 1.04

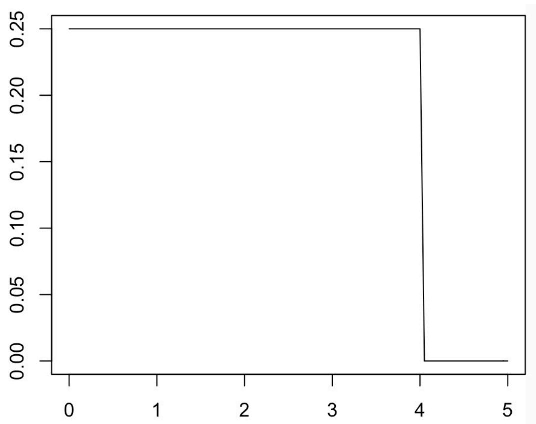
(b) (1 point) 53.6, 59.6

116. (a) (1 point) Since the figure is a defined as a density curve, then by definition it has a total area of 1 square unit. The area represents 100% of the population.

(b) (1 point) (area left of 0.35) = $1 \times 0.35 = 0.35$

(c) (1 point)
(area between 0.35 and 0.65) = (area left of 0.65) - (area left of 0.35)
$= 1 \times 0.65 - 1 \times 0.35 = 0.3$

117. (a) (1 point) $1/4 = 0.25$



(b) (1 point) (area left of 1) = $1 \times 0.25 = 0.25$

(c) (1 point) (area between 0.5 and 2.5) = (area left of 2.5) - (area left of 0.5)
$= 0.25 \times (2.5 - 0.5) = 0.5$

118. (1 point) $\bar{x} = 0.5$, median = 0.5, $Q_1 = 0.25, Q_3 = 0.75$.

132. (1 point)
For Tonya, $z_1 = \frac{1820-1509}{321} = 0.9688$. For Jermaine, $z_2 = \frac{29-21.5}{5.4} = 1.3889$.
Since $z_1 < z_2$, then we conclude that Jermaine has a higher score.

134. (1 point) $z = \frac{2080-1509}{321}$, $5.4z + 21.5 = 31.11$.

136. (1 point) $z = \frac{2090-1509}{321} = 1.81$. Using Table A, the area left to 1.81 in the standard normal distribution is 96.49%.

139. (1 point) According to Table A, the 20% quantile is around -0.84.
$-0.84 \times 321 + 1509 = 1239$.
The SAT scores which are less than or equal to 1239 make up the bottom 20%.

145. (a) (1 point) $z_1 = \frac{240-266}{16} = -1.625$.
According to Table A, (area left of -1.625) = 0.0526. The percent of pregnancies less than 240 days is about 5.26%.

(b) (1 point) $z_2 = \frac{270-266}{16} = 0.25$. According to Table A, (area between -1.625 and 0.25) =-(area left of 0.25) - (area left of -1.625) = 0.547.
The percent of pregnancies between 240 and 270 days is about 54.7%.

(c) (1 point) The 80% percentile of standard normal distribution is 0.84.
$0.84 \times 16 + 266 = 279$. The longest 20% of pregnancies last more than 279 days.

147. (a) (1 point) $Q_1 = -0.675, Q_3 = 0.675, IQR = Q_3 - Q_1 = 1.35$.

(b) (1 point) For $N(\mu, \sigma)$ distribution, the quartiles are $Q_1^* = \mu + 0.675\times\sigma$, $Q_3^* = \mu + 0.675\times\sigma$.

$IQR^* = Q_3^* - Q_1^* = \mu + 0.675\times\sigma - \mu - 0.675\times\sigma = 1.35\sigma$. Thus, $c = 1.35$.

148. (1 point)

Since the tail area for the lowest 25% has a $z = -0.67$ and the z for the upper 25% is $z = 0.67$, the interquartile range is 1.34.

$1.5\times1.34 = 2.01$. So we need to find the percent that falls above $0.67 + 2.01 = 2.68$ and the percent that falls below -2.68. The tail area is $2\times0.0037 = 0.0074 = 0.74\%$.