Dehshini Ali

R version 4.3.1 (2023-06-16) -- "Beagle Scouts"

> baltps1 <- read.csv("~/Library/CloudStorage/OneDrive-JohnsHopkins/JHU/1st Term/statistical methods/week1/baltps1.csv")
> View(baltps1)

Question1

1. a

> baltps1.1 = filter(baltps1, group==1)
> baltps1.2 = filter(baltps1, group==2)

> stem(baltps1.1$deaths, scale=1)

The decimal point is 1 digit(s) to the right of the |

1 | 44
1 | 567778899
2 | 334
2 | 5678
3 | 3
3 | 7

> stem(baltps1.1$deaths, scale=3)
The decimal point is at the |

14 | 000
16 | 0000
18 | 0000
20 |
22 | 00
24 | 00
26 | 00
28 | 0
30 |
32 | 0
34 |
36 | 0

> stem(baltps1.2$deaths, scale=1)

The decimal point is 1 digit(s) to the right of the |

1 | 134
1 | 5666667899
2 | 0134
2 | 577

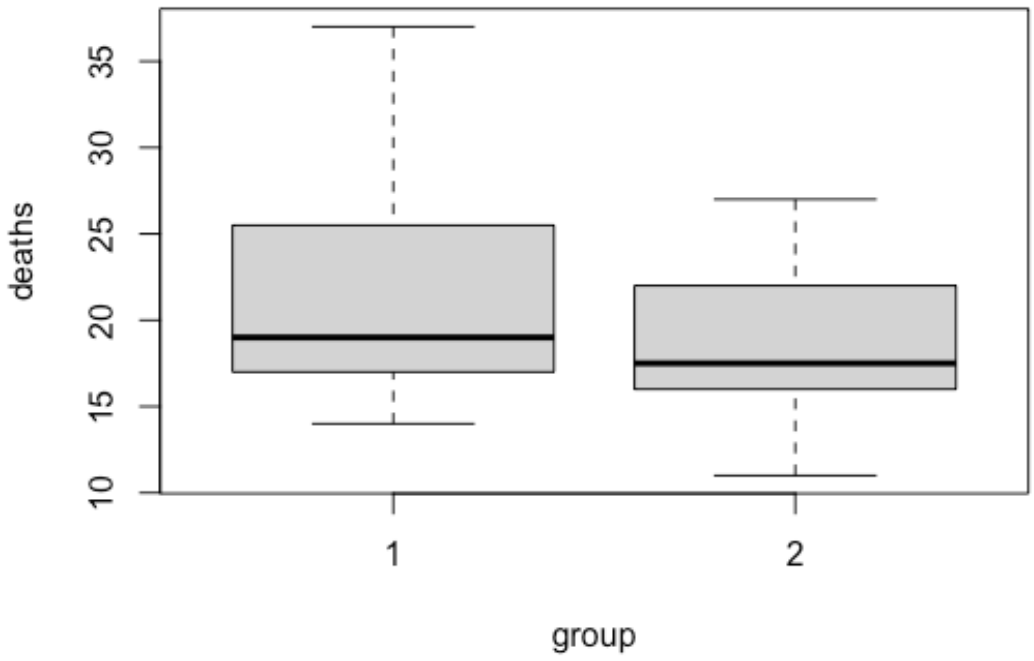
> stem(baltps1.2$deaths, scale = 3)

The decimal point is at the |

10 | 0
12 | 0
14 | 00
16 | 000000
18 | 000
20 | 00
22 | 0
24 | 00
26 | 00

1.b

> boxplot(deaths ~ group, data=baltps1)



1.c.

I obtained similar results by hand

1.d.

The number of deaths a day after high particulate air pollution days were higher (range 14 – 37 deaths, median 19 deaths, IQR 9) as compared to the number of deaths a day after low particulate air pollution days (range 12 – 27 deaths, median 17.5 deaths, IQR 7).

1.e

Script file saved

Question 2

2.i

```
>   View(ce621)
> ce621.male = filter(ce621, sex=="Male")
> ce621.female = filter(ce621, sex=="Female")
> stem(ce621.male$totchg)
```

The decimal point is 3 digit(s) to the right of the |

```
 2 | 599002344456677999
 4 | 011233334444444457900012223444566999
 6 | 00122444556777899900223 49
 8 | 577891357
10 | 22678
12 | 288
14 | 7
16 | 1
18 | 44
```

```
> stem(ce621.female$totchg)
```

The decimal point is 4 digit(s) to the right of the |

```
0 | 023333333333444444444444444
0 | 55555555555555555556666666666666777777778999999
1 | 0000112223334
1 | 778889
2 | 00
2 | 5
3 | 13
3 |
4 | 0
```

2.ii

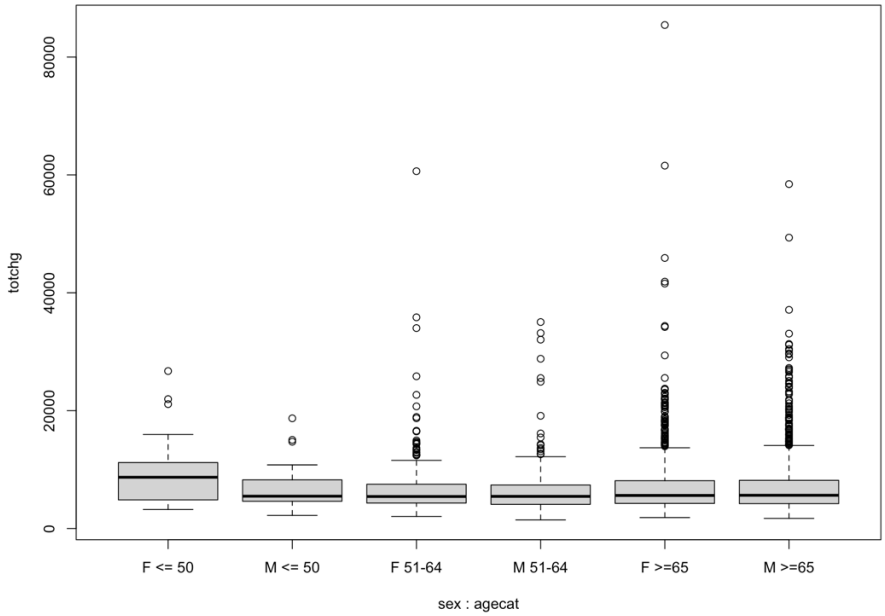
```
> summary(ce621.male$totchg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2528    4328    5598    6484    7189   19427
> sd(ce621.male$totchg)
[1] 3277.939
> quantile(ce621.male$totchg, c(0,.1,.25,.5,.75,.9,1))
     0%     10%     25%     50%     75%     90%    100%
 2528.0  3528.6  4327.5  5597.5  7189.0 11231.9 19427.0
> summary(ce621.female$totchg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    200    4359    5644    8099    9546   40083
> sd(ce621.female$totchg)
[1] 6679.283
> quantile(ce621.female$totchg, c(0,.1,.25,.5,.75,.9,1))
     0%     10%     25%     50%     75%     90%    100%
 200.00  3197.70  4359.25  5643.50  9546.25 16817.60 40083.00
```

Statistics	Male	Female
Typical value		
Mean	6484	8099
Median	5598	5644
Variability		
Interquartile range IQR	7189.0 - 4327.5 = 2861.5	9546.25 - 4359.25 = 5187
Standard deviation s	3277.939	6679.283

2. iii

The mean cost for males in this study is 6484, whereas the mean cost for females is 8099. The IQR for males (2861.5) is lower compared to that of females (5187) with median values of 5598 and 5644 respectively. The greatest proportion of the sample population was at the lower end of the cost graph (spent relatively less) since both the male and female cost distributions are skewed to the right.

2. iv

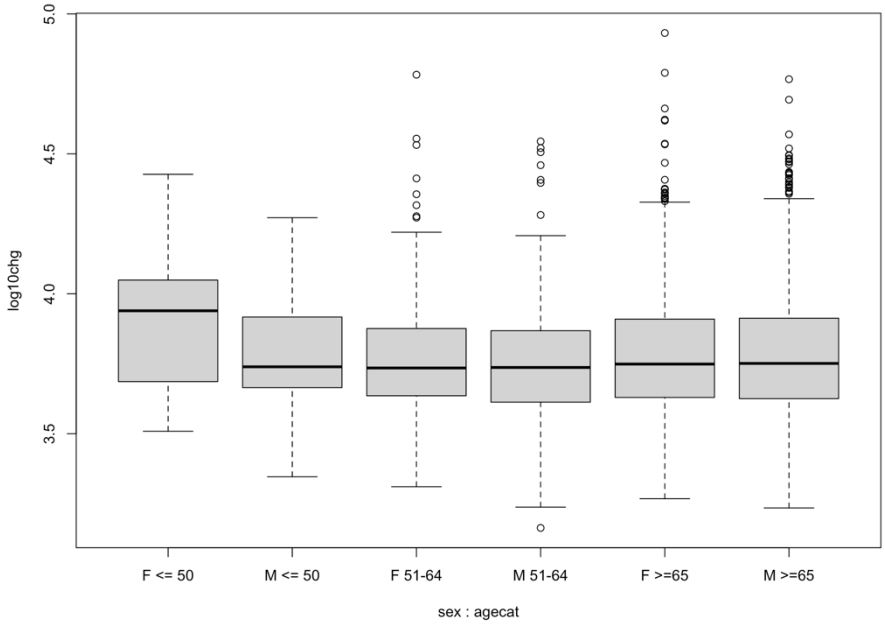


2. v

With the exception of the age group under 50, where the cost for males is comparatively cheaper than the cost for females, the cost of CE for males is equivalent to the cost for women.

While the cost of CE for men is largely constant regardless of age, the cost of CE for females marginally reduces as age increases.

2. vi



- Comparing the arithmetic scale to the log scale, the arithmetic scale for total charge displays broader boxes and longer whiskers.
- It has become evident that the minimum value for males in any given age group is actually lower than the minimum value for females in that same age group.
- Males in the 51–64 age group now have a lower outlier visible.
- There are fewer upper outliers overall, with the <=50 age groups for both sexes now being free of outliers.