Chapter 1 - Introduction to Data

Binish Kurian Chandy

February 5, 2018

1.8
a. Each row represent a case (smoking habit of each UK resident who participated in the survey).
b. 1691
c. sex : nominal categorical
age : continuous numerical
marital : nominal categorical
grossIncome : ordinal categorical
smoke: nominal categorical
amtWeekends: discrete numerical
amtWeekdays: discrete numerical

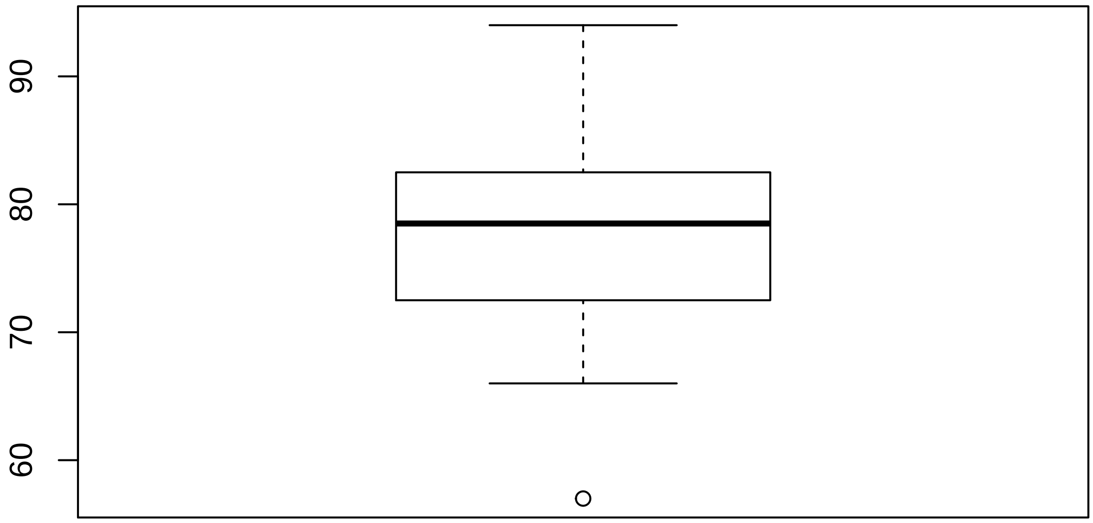
1.10
a. Population of interest : all children between age of 5 and 15 Target population : 160 children between the age of 5 and 15
b. If the students in this sample, who are likely not randomly sampled, can be considered to be representative of all children between 5 and 15, then the results are generalizable to the population defined above. Additionally, since the study is experimental, the findings can be used to establish causal relationships.

1.28
a. No. The participants were not randomly sampled (voluntary sample). Under coverage bias, the sample may not represent the population as it is voluntary.
b. No. Observational studies are generally sufficient to show assosciation only not conclusion. The best conclusion that can be drawn is there can be a confounding variable (eg: general or mental health of student) that can cause sleep disorder and behavioral issues.

1.36
a. Randomized experiment
b. Treatment group : group that exercised twice a week.
Control group : group that didn't exercise
c. Age
d. No
e. Since this is a randomized experiment, we can make a causal relationship
f. Only reservation I have is about the bias that can happen because of not blinding the control group and this can affect the end result of the study.

1.48

```
data <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)
boxplot(data, xlab = "final exam score")
```



final exam score

1.50
a. 2
b. 3
c. 1

1.56
a. The distribution of houses is rightly skewed as there are handful of house prices that are outliers. Therefore the center would be best described by the median and the variability would be describe by IQR.
b. The distribution is symmetrical even though some outliers are present. Mean and SD are robust statistics.
c.The distribution of number of alcoholic drinks by college students is likey right skewed as there is a boundary at 0 (no drinks since under 21 years old) and a few people drink excessively. Therefore the center would be best described by the median, and variability would be best described by the IQR.
d. The distribution of salary is likely right skewed. That is, most employees make a small amount of money while a few executives earn much higher salaries. Therefore the center would be best described by the median, and variability would be best described by the IQR.

1.70 a. The vertical locations at which the control and treatment groups break into alive and dead categories differ, suggests that two variables may be dependent.
b. The median survival time of treatment group is 207 days which is much higher than control group which is 21 days. Also survival time of first and third quartile of treatment group is higher than control group. So by having heart transplant patients may be surviving longer.
c.

```
treatment_grp_death <- 45/69
print(treatment_grp_death)
```

```
## [1] 0.6521739
```

```
control_grp_death <- 30/34
print(control_grp_death)
```

```
## [1] 0.8823529
```

d.
i. H0: The heart transplant and increased lifespan are independent. They have no relationship, and the

increased lifespan form heart transplant is due to chance.

HA: The heart transplant and increased lifespan are not independent. The increase in lifespan is due to heart transplant.

ii.

```
28
75
69
34
0
24/69 - 4/34
```

iii. Zero out of 100 simulations show atleast 23% difference. We conclude that the evidence is strong to reject H0 null hypthesis. In this case we accept the alternative model ie increased life span and heart transplant are dependent.