1.2

(a)

$$P(\mathrm{A}) = \frac{66}{85} \approx 0.776 = 77.6\% \qquad P(\mathrm{B}) = \frac{65}{81} \approx 0.802 = 80.2\%$$

The percent of patients in the treatment group experienced a significant improvement is about 77.6%.

The percent of patients in the control group experienced a significant improvement is about 80.2%.

(b) Since $P(\mathrm{A}) < P(\mathrm{B})$ a little bit, the control group which uses placebo consisted of symptomatic treatments appears to be more effective.

(c) The data may not be regarded as convincing evidence. Whether or not getting improvements were self-reported, so it is not exactly right. And the number of total participants is not large enough. The extent that they got sick may also be different. Also, the difference between the two probabilities is not large. So the difference may be due to chance.

1.4

(a) cases: eligible study subjects that consist of 600 asthma patients aged 18 to 69 who relied on medication for asthma treatment

(b) variables and types: scores on quality of life, activity, asthma symptoms, and medication reduction given by patients on a scale from 0 to 10. I suppose the scores are discrete numerical variables (consecutive integers).

(c) the main research question: Is it effective or not by adopting Buteyko method to reduce asthma symptoms and to improve quality of life?

1.6

(a) cases: 129 University of California undergraduates at Berkeley

(b) variables and types: low or high socio-economic class (categorical), most or least money (categorical), most or least education (categorical), most or least respected jobs (categorical), the number of candies undergraduates had taken (discrete numerical, consecutive integer)

(c) the main research question: What's the relationship between socio-economic class and unethical behavior?

1.8

(a) Each row represents the data of an individual UK resident. A single case.

(b) 1691 participants.

(c) sex (categorical, not ordinal), age (numerical, discrete when only write down the integers), marital (categorical, not ordinal), grossIncome (Since there is not enough information of the whole data, I suppose there is no overlapping between different intervals. It should be categorical, ordinal), smoke (categorical, not ordinal), amtWeekends (numerical, discrete, integer), amtWeekdays (numerical, discrete, integer)

1.10

(a) Population: children between the ages of 5 and 15. Sample: 160 such children

(b) If children in this sample can be considered to be representative of all children between the ages of 5 and 15, then the results are generalizable to the population defined above. Since the study is experimental, the findings can be used to establish casual relationships.

1.12

(a) Population: undergraduates of different socio-economic classes. (While the description did not convey that information. I suppose it should be.) Sample: 129 University of California Berkeley such undergraduates

(b) If undergraduates in this sample can be considered to be representative of all undergraduates of high and low socio-economic classes, then the results are generalizable to the population defined above. I suppose it is not representative. Since the study is experimental, the findings can be used to establish casual relationships.

1.18

(a) Observational

(b) Use stratified sampling to randomly sample a fixed number of students, say 5% of all freshmen and 5% of all sophomores who live in the eastern part, and 5% of all juniors and 5% of all seniors who live in the western part. Also the number of students in the four grades should be approximate.

1.22

Each individual in the population has the same probability to be selected as sample. Random digit dialing includes unlisted numbers that would be missed if the numbers were selected from a phone book. In other words, it is likely to get information from people who are not easy to contact.

1.24

Asking elementary school students about their family sizes is an observational study. It is not a good measure of household size. Because students may not correctly define the real size of household and they may not tell the truth. Students may also get influenced by others' answers if possible. Family size may not be the actual household size. Random sampling is good theoretically of no bias. The true value of family size may be underestimated because students might miss some members who are living together in the same dwelling and sharing living accommodations.

1.26

(a) Simple random sample. It may not be effective because there are diverse kinds of housing in this area. Random sampling may not cover all the types.

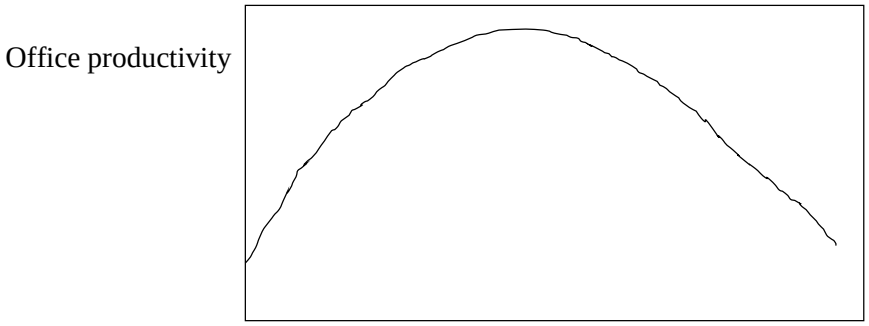
(b) Stratified sampling. For this kind of sampling, similar cases should be classified as one strata. But there are diverse types with unclear boundary of each other. That is to say, there are mixtures. So this method may not be effective enough.

(c) Cluster sampling. For diverse and mixture types of cases, this method is effective in this setting.

(d) Multistage sampling. For diverse and mixture types of cases, this method is effective in this setting.

(e) Convenience sampling. It is not effective with bias of certain types of neighborhoods.

1.40



1.44

(a) The score of the student who took the make-up exam will decrease the average score.

(b) $average_{new} = \frac{74 \times 24 + 64}{25} = 73.6$ The new average score is 73.6 points.

(c) The new score of the student will increase the standard deviation because it is much lower than the mean.

1.46

(a) The medians of (1) and (2) are identical, that is 6, because there is only one number 20 that is different. The IQR of (2) is greater than the IQR of (1), because both values of Q1 are the same and the Q3 of (2) is obviously larger because number 20.

(b) The median of (1) is 6, while the median of (2) is 7. So the median of (1) is smaller. Both (1) and (2) have the same Q1, and the Q3 of (2) is obviously higher than that of (1). So the IQR of (1) is smaller than the IQR of (2).

(c) The median of (1) is smaller than the median of (2), because all the numbers in (1) are smaller than those in (2). The IQR of (1) and (2) are identical, because both (1) and (2) are five consecutive integers.

(d) The median of (1) which is 50, is smaller than the median of (2) which is 500. The IQR of (1) is smaller than the IQR of (2) because all the numbers of (2) are ten times of those of (1).

1.47

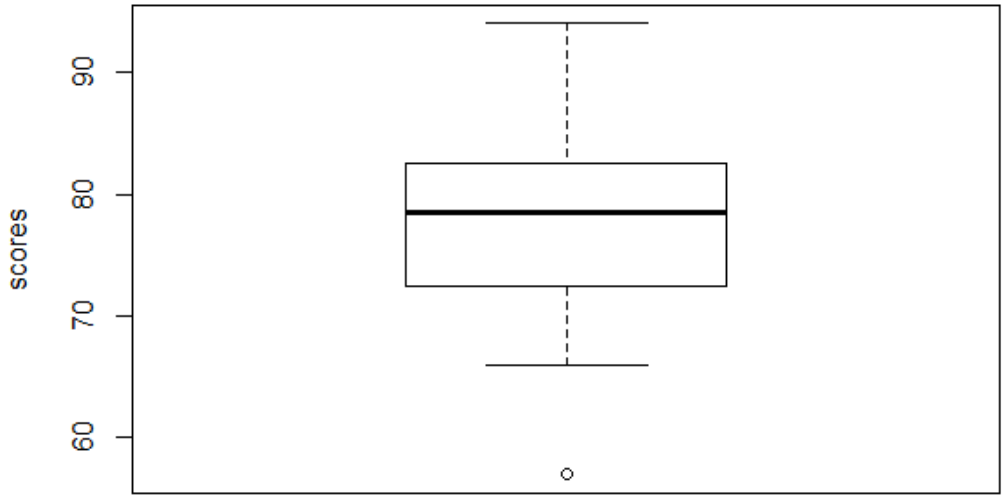
(a) The mean of (2) is higher because 20>13. The standard deviation of (2) is also higher because 20 is further from the rest of the numbers than 13.

(b) The mean of (1) is higher because -20>-40. The standard distribution of (2) is higher because -40 is further from the rest of the numbers than -20.

(c) The mean of (2) is higher because all values in (2) are higher than those of (2). Both (1) and (2) have the same standard deviation since they are equally variable around their respective means.

(d) Both (1) and (2) have the same mean, which is 300. (2) has a higher standard deviation since the observations are farther from the mean than in (1).

1.48



The circle in the box plot means outlier. That's the minimum which is smaller than Q1-1.5*IQR.

1.52

The estimated median is between 80 and 85. I expect the mean to be lower than the median since there are some extreme small numbers compared to the median. It seems that there are more numbers that are lower than the median.

1.56

(a) I expect the distribution to be right skewed. The median would best represent a typical observation since most prices are relatively low, and there are some extreme high prices that will increase the mean a lot. The variability of observations would be best represented using IQR because the extreme high values will increase the standard deviation a lot.

(b) I expect the distribution to be symmetric. The mean would best represent a typical observation, and the variability of observations would be best represented using standard deviation. Because the distribution seems normal without many extreme values.

(c) I expect the distribution to be left skewed. The median would best represent a typical observation since most students do not drink, and there are a few extreme high values that will increase the mean a lot. The variability of observations would be best represented using IQR because the extreme high values will increase the standard deviation a lot.

(d) I expect the distribution to be symmetric. The median would best represent a typical observation, and the variability of observations would be best represented using IQR. Because the distribution seems concentrated on middle level of salaries with a few high salaries.