Exam 01

STA 198 Spring 2020 (Jiang)

Wednesday, June 10, 2020

Question 2

2.1

Categorical ordinal.

2.2

These data came from an observational study; the researchers did not have the ability to manipulate any of the variables in the dataset, but rather only examined associations.

2.3

Since we are looking for the number of successes (dangerous or hazardous pollution days) in a fixed number of trials (the number of days in a given month), a binomial distribution might be the most appropriate.

2.4

There is indeed a fixed number of fixed trials, each of which is a Bernoulli random variable. However, the outcomes of the trials are not independent; if one day is very polluted, then it is likely that neighboring days will be polluted as well. Thus, each day is not independent. As well, we also see that the probability of success might not be the same for each one of the trials (it may fluctuate depending on neighboring days).

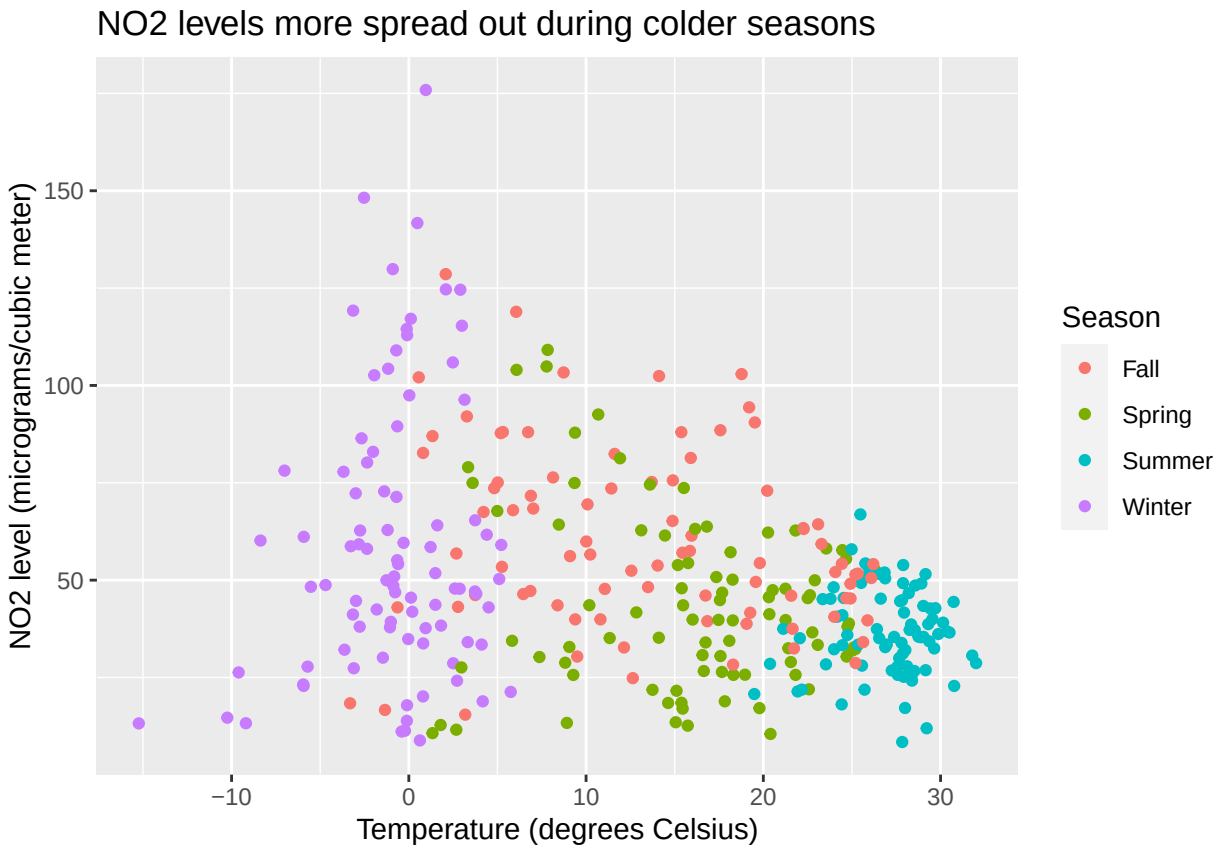
2.5

Boxplots and histograms can both be used to summarize the distribution of a single continuous variable. A histogram readily captures the shape of the entire distribution (for instance, skewness, spread, number of modes, etc.), but this information is harder to come by in a boxplot (in fact, it is impossible to tell how many modes a distribution has from simply its boxplot). On the other hand, the boxplot readily displays the five number summary and specifically calls out any outlying values, which may not be immediately apparent in a histogram.

Question 3

3.1

An example plot is below. Note that all plots must have summarized the three-way relationship between NO2, temperature, and season. NO2 and temperature should have been continuous variables, and season should have been a categorical variable. The plot should have had well-labeled axes and a meaningful title. Finally, the season variable should have been correctly created from the month variable, with -99 values excluded from the plot.



3.2

Code is given below to recreate the previous plot:

```
library(tidyverse)
pollution <- read_csv("pollution.csv")

pollution <- pollution %>%
  filter(NO2 != -99) %>%
  mutate(Season = case_when(month %in% c(12, 1, 2) ~ "Winter",
                            month %in% c(3:5) ~ "Spring",
                            month %in% c(6:8) ~ "Summer",
                            month %in% c(9:11) ~ "Fall"))

ggplot(data = pollution, mapping = aes(x = TEMP,
                                       y = NO2,
                                       color = Season)) +
  geom_point() +
  labs(x = "Temperature (degrees Celsius)",
       y = "NO2 level (micrograms/cubic meter)",
       title = "NO2 levels more spread out during colder seasons")
```

3.3

As expected, winter had the lowest temperatures, summer had the highest, and spring and fall were in the middle. Colder seasons had a much higher spread of NO2 level, and were also associated with relatively

higher NO2 levels in general.

Question 4

4.1

$$\begin{aligned} P(T^+) &= P(D^+) \times P(T^+|D^+) + P(D^-) \times P(T^+|D^-) \\ &= prev. \times sens. + (1 - prev.) \times (1 - spec.) \\ &= 0.06 \times 1 + (1 - 0.06) \times (1 - 0.99) \\ &= 0.0694 \end{aligned}$$

4.2

The number of patients who are actually negative out of 20 positive tests is given a binomial distribution, with $n = 20$ and $p = P(D - |T+)$. Note that this probability is **not** the false positive rate, $P(T^+|D^-)$.

To calculate the probability of interest, we have

$$\begin{aligned} P(D^-|T^+) &= \frac{P(T^+|D^-)P(D^-)}{P(T^+)} \\ &= \frac{(1 - P(T^-|D^-))P(D^-)}{P(T^+|D^-)P(D^-) + P(T^+|D^+)P(D^+)} \\ &= \frac{(1 - spec.) \times (1 - prev.)}{(1 - spec.) \times (1 - prev.) + sens. \times prev.} \\ &= \frac{(1 - 0.99) \times (1 - 0.06)}{(1 - 0.99) \times (1 - 0.06) + 1 \times 0.06} \\ &= \frac{47}{347} \end{aligned}$$

Thus, letting X be the random variable in question, $X \sim Binom(20, 47/347)$. The probability that at least two are negative is given by $P(X \geq 2) = 1 - P(X \leq 1)$ which we find in R using `1 - pbinom(1, 20, 47/347)`. This is approximately 0.7750.

4.3

Note the following:

$$\begin{aligned} P(D^+|T^+) &= \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} \\ &= \frac{sens. \times prev.}{sens. \times prev. + (1 - spec.) \times (1 - prev.)} \end{aligned}$$

Plugging in for the sensitivity and specificity, we have the boundary at

$$0.9 = \frac{prev.}{prev. + 0.01 \times (1 - prev.)}$$

Solving for the prevalence, we find it equal to 9/109 (or approximately 0.0826). Thus, any prevalence equal to approximately 0.0826 or greater would result in positive predictive values of 90% or greater.

Question 5

5.1

```
qnorm(0.95, 2.7, 0.39)
```

```
## [1] 3.341493
```

Lead levels of approximately 3.3415 log micrograms per deciliter would be considered at risk prior to lead-reducing legislation.

5.2

```
pnorm(3.0, 2.7, 0.39) - pnorm(2.5, 2.7, 0.39)
```

```
## [1] 0.4750834
```

0.4751.

5.3

```
1 - pnorm(1.61, 2.7, 0.39)
```

```
## [1] 0.997404
```

0.9974 (whoa!).

Question 6

6.1: FALSE

6.2: TRUE

6.3: FALSE

6.4: TRUE

6.5: TRUE

6.6: TRUE

6.7: FALSE

6.8: FALSE

6.9: TRUE

6.10: FALSE