HW04 Solutions

Stat 145

Due February 6, 2024 at 11 pm

Problem set policies. Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving *R*, include the code in your solution, along with any plots. (Using R Markdown to write up your answers would be ideal.)

Please submit your homework set via MOM as a PDF, along with the R Markdown source file.

I encourage you to discuss problems with other students or me, but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution

Problem 1.

Since states with larger numbers of elderly residents would naturally have more nursing home residents, the number of nursing home residents in a state is often adjusted for the number of people 65 years or older (65+). That adjustment is usually given as the number of nursing home residents age 65+ per 1,000 members of the population age 65+. For example, a hypothetical state with 200 nursing home residents age 65+ and 50,000 people age 65+ would have the same adjusted number of residents as a state with 400 residents and a total age 65+ population of 100,000 – 4 residents per 1,000.

The data file found at the link

https://scofield.site/teaching/data/csv/harrington/nursingHome.csv

contains this adjusted number of residents for each state in the United States. The state names are saved under the variable name `state` and the adjusted number of residents under the variable name `resident`.[1]

a) Which state has the smallest number of nursing home residents per 1000 population 65 years of age and over? Which state has the largest number? Hint: use the R functions `min()` and `max()`, or perhaps `qdata()`. Alternatively, look directly at the data in *RStudio*.

Hawaii (13.6) has the smallest number of nursing home residents per 1000 population 65 years of age and over while South Datoka (74.9) has the largest.

```
#load the data
nursingHome = read.csv("https://scofield.site/teaching/data/csv/harrington/nursingHome.csv")

# get the 5-number summary?  It includes the min/max
qdata(~resident, data=nursingHome)
```

```
##    0%   25%   50%   75%  100%
## 13.60 32.85 44.20 54.30 74.90
```

[1] The data originally appeared in Chapter 12 of *Case Studies in Biometry*, 1994, by Lange et al.

```
# Or, get the min/max explicitly
min(~resident, data=nursingHome)
```

```
## [1] 13.6
```

```
max(~resident, data=nursingHome)
```

```
## [1] 74.9
```

So, '13.6' is the smallest value, while '74.9' is the largest. We might filter the data to determine which states have these values:

```
filter(nursingHome, resident < 13.7)
```

```
##    X  state resident
## 1 12 Hawaii     13.6
```

```
filter(nursingHome, resident > 74.8)
```

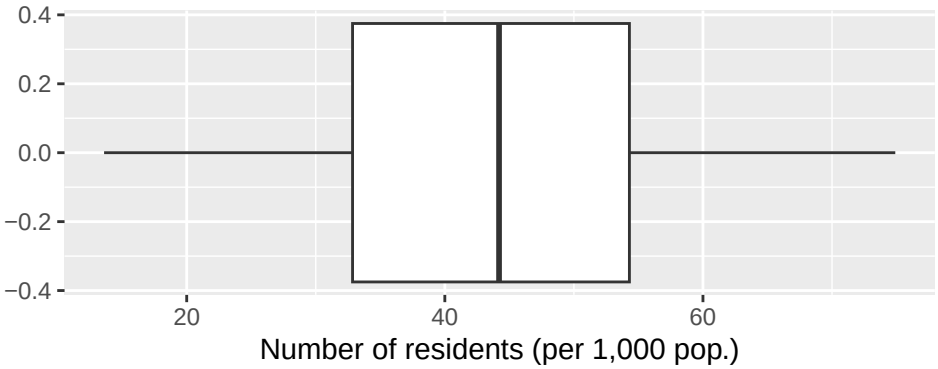
```
##    X        state resident
## 1 42 South Dakota     74.9
```

b) What factors might influence the substantial amount of variability among different states? This question cannot be answered from the data; speculate using what you know about the demographics of the United States.

Various factors could influence the number of elderly people in a given state who choose to live in nursing homes. In states that are more sparsely populated, more residents might choose to stay in nursing homes because of the difficulties associated with living in rural areas, such as absence of easily accessible medical care. In wealthier states, however, people are less likely to live in nursing homes because of their ability to afford the costs of staying home (e.g. in-home nursing care); Social Security, Medicare, and Medicaid pay quite a lot of nursing home costs.

c) Construct a boxplot for the number of nursing home residents per 1,000 population.

```
#construct a boxplot
gf_boxplot(~ resident, data=nursingHome) |>
  gf_labs(x="Number of residents (per 1,000 pop.)")
```



d) Is the distribution of nursing home resident per 1000 population symmetric or skewed? Are there any states that could be considered outliers?

The distribution looks symmetric and the boxplot does not indicate any outliers.

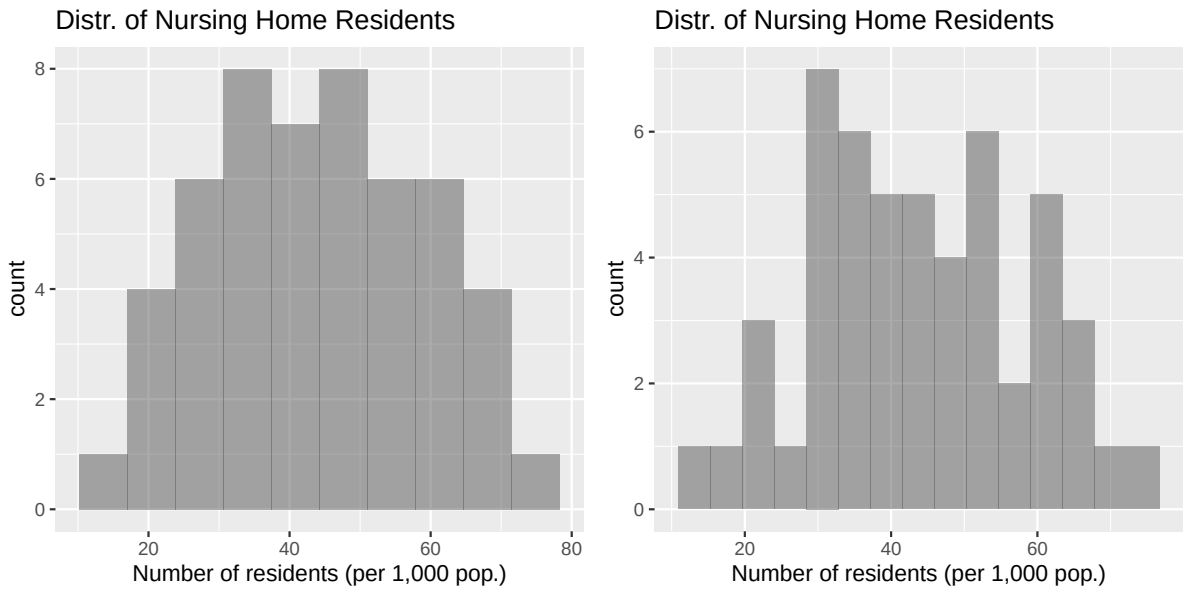
e) Display the number of nursing home residents per 1000 population using a histogram. Do you find this graph to be more or less informative than the box plot? Explain your answer.

The histogram is more informative than the boxplot because it provides a more detailed picture of how the data are distributed. The boxplot does not provide information about the number of modes in the data or the overall data shape outside of the interquartile range.

A histogram with more bins *might*(?) reveal additional detail; the version shown on the left has 10 bins, and the one on the right has 15.

```
# The package gridExtra supplies the grid.arrange() command

# construct a histogram with 10 bins
p1 = gf_histogram(~resident, data=nursingHome,
                  bins=10) |>
  gf_labs(title = "Distr. of Nursing Home Residents",
 x = "Number of residents (per 1,000 pop.)")
# construct another histogram with 15 bins
p2 = gf_histogram(~resident, data=nursingHome,
                  bins=15) |>
  gf_labs(title = "Distr. of Nursing Home Residents",
 x = "Number of residents (per 1,000 pop.)")
grid.arrange(p1, p2, ncol=2)
```



Problem 2.

The data file found at

https://scofield.site/teaching/data/csv/harrington/adolescentFertility.csv

contains data on the number of children born to women aged 15-19 from 189 countries around the world for the years 1997, 2000, 2002, 2005, and 2006.[2] The data are defined using a scaling similar to that used in the nursing home data. The values for the annual adolescent fertility rates represent the number of live births among women aged 15-19 per 1,000 women members of the population of that age.

For the years 2000-2006, the adolescent fertility rate for Iraq is coded `NA`, or missing. When calculating a mean or standard deviation in R for a variable `x` which has missing data, add `na.rm=TRUE` to the argument to perform the calculations without the missing observations:

```
mean(~x, na.rm=TRUE)
sd(~x, na.rm=TRUE)
```

a) Calculate the mean, standard deviation, and five-number summary for the distribution of adolescent fertility in 2006 (`fert_2006`).

The mean is 53.58; the standard deviation is 46.98. The five-number summary consists of the minimum, first quartile, median, third quartile, and maximum: 1.45, 17.88, 40.07, 75.73, and 223.80. (All numbers in units of live births per 1,000 women aged 15-19.)

```
#load the data
teenpreg = read.csv("https://scofield.site/teaching/data/csv/harrington/adolescentFertility.cs

#calculate numerical summaries
mean(~fert_2006, data=teenpreg, na.rm=TRUE)
```

```
## [1] 53.58395
```

```
sd(~fert_2006, data=teenpreg, na.rm=TRUE)
```

```
## [1] 46.97848
```

```
qdata(~fert_2006, data=teenpreg)
```

```
##       0%      25%      50%      75%     100%
##   1.4534  17.8759  40.0682  75.7267 223.8336
```

b) What is the 70^{th} percentile of the distribution? Write a sentence explaining the 70^{th} percentile in the context of this data.

The 70^{th} percentile is 64.77. 70% of the countries have an adolescent fertility rate less than or equal to 64.77 births per 1,000 adolescents.

c) Suppose those observations for Iraq between 2000 and 2006 could be added. Is it likely the addition of these numbers would have notable effect on the five-number summary?

It is unlikely that the five-number summary would have been affected very much, even if the values were extreme; the median and IQR are robust/resistant estimates, and the dataset is

[2] Data from the CIA World Factbook

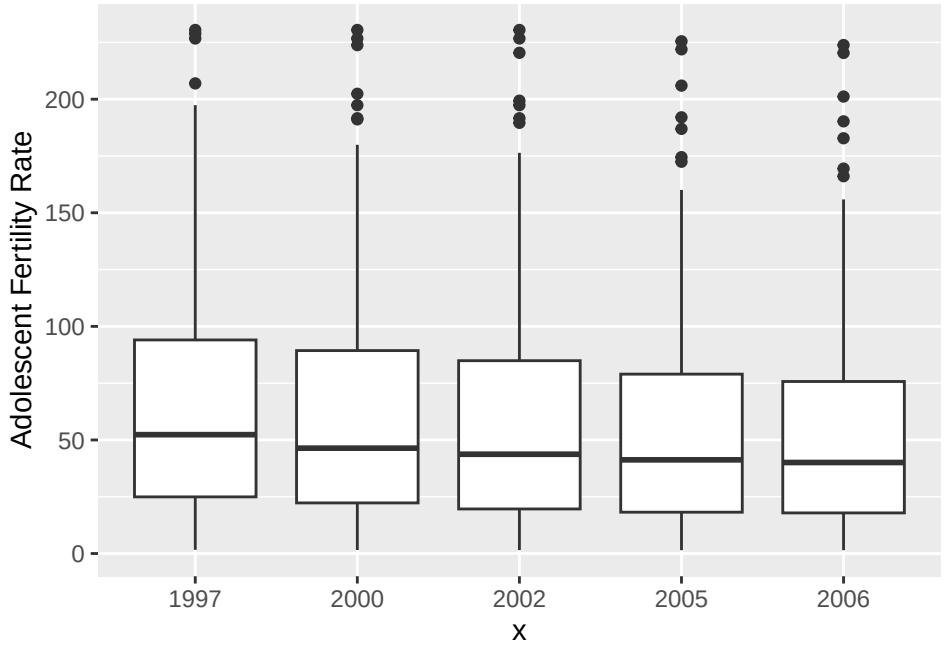
relatively large, with data from 188 other countries.

d) Produce side-by-side boxplots of the fertility rates for each of the five years in the dataset. What pattern do you see?

The median and IQR decreases each year, with Q1 and Q3 also decreasing.

```
teenpreg |> gf_boxplot(fert_1997 ~ "1997") |>
   gf_boxplot(fert_2000 ~ "2000") |>
   gf_boxplot(fert_2002 ~ "2002") |>
   gf_boxplot(fert_2005 ~ "2005") |>
   gf_boxplot(fert_2006 ~ "2006") |>
   gf_labs(y = "Adolescent Fertility Rate")
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
## Removed 1 rows containing non-finite values (`stat_boxplot()`).
## Removed 1 rows containing non-finite values (`stat_boxplot()`).
## Removed 1 rows containing non-finite values (`stat_boxplot()`).
```



Problem 3.

A recently published analysis examined 10 studies that measured optimism and pessimism by asking participants about their level of agreement with statements like "In uncertain times, I usually expect the best," or "I rarely expect good things to happen to me". Optimistic people tend to expect that they will encounter favorable outcomes, whereas less optimistic people tend to expect that they will encounter unfavorable outcomes.[3]

These studies also measured other variables on participants, including factors related to heart disease. The analysis found that compared with pessimists, people with the most optimistic outlook had a 35% lower risk for cardiovascular events (e.g., heart attacks). The studies, on average, observed people over a 14-year period and compared the rate of cardiovascular events between those classified as optimists versus pessimists.

a) A popular newspaper reports on the analysis with the headline "Thinking Positively Improves Cardiovascular Health". Write a short response to the editor explaining clearly why the headline is potentially misleading. Be sure to use language accessible to a general audience without a statistics background. Limit your answer to at most five sentences.

   The headline is potentially misleading because it interprets evidence for an association between optimism and lower risk for cardiovascular events as a causal relationship. It is not prudent to make causal claims from observational studies since there could be unmeasured variables (i.e., confounding variables) obscuring the true causal relationship. For example, perhaps optimists tend to lead lifestyles that promote cardiovascular health (e.g., exercising more, eating healthier diets), and it is these lifestyle factors that actually have a causal effect on cardiovascular health. The analysis only demonstrates evidence that optimistic people tend to have lower risk for cardiovascular health; it does *not* demonstrate evidence that optimism improves cardiovascular health.

b) Briefly describe a plausible study design that has the potential to demonstrate the effect of thinking positively on cardiovascular health.

   Randomly select a study sample from the population of interest. Randomize half the participants to the control group and half the participants to the treatment group. The control group will receive typical advice about behaviors that promote cardiovascular health. The treatment group will receive typical advice in addition to attending mindfulness sessions that teach strategies for thinking positively. Observe participants over a set period of time, then compare rate of cardiovascular events between the groups.

c) Suppose someone who is very optimistic reads about the analysis and concludes that the findings suggest he has a 35% lower risk for cardiovascular events than his friend who is extremely pessimistic. Explain why this is not necessarily the case.

   This is not necessarily the case because "the individual is not the average". Each person's individual risk for cardiovascular events is influenced by a myriad of factors specific to that person, such as diet and family history. The 35% risk figure is an overall average calculated using data from many individuals; it would be flawed to assume that all individuals have the same risk as the average risk. In mathematical terms, for example, if $\bar{x} = 5$, this does not imply that $x_1 = x_2 = ... = x_n = 5$.

[3] Alan Rozanski, MD, et al. Association of optimism with cardiovascular events and all-cause mortality. *JAMA Network Open* 2019; 2(9):e1912200.

Problem 4.

Suppose that you are interested in determining whether a relationship exists between the fluoride content in a public water supply and the dental caries experience of children using this water. The file

https://scofield.site/teaching/data/csv/harrington/water.csv

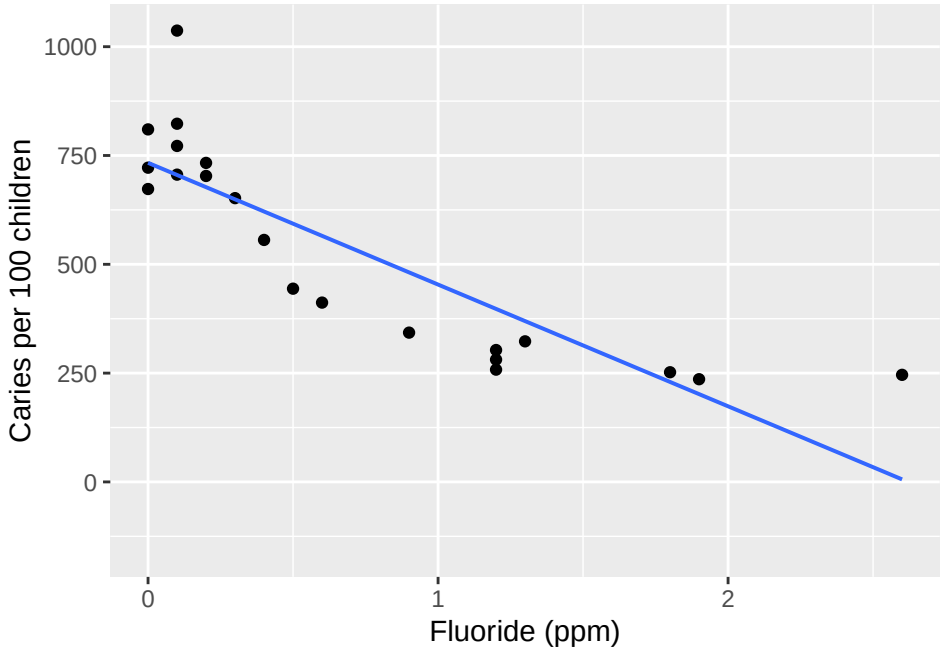
contains the data from a study examining 7,257 children in 21 cities from the Flanders region in Belgium.

The fluoride content of the public water supply in each city, measured in parts per million (ppm), is saved under the variable name `fluoride`; the number of dental caries per 100 children examined is saved under the name `caries`. The total dental caries number is obtained by summing the numbers of filled teeth, teeth with untreated dental caries, teeth requiring extraction, and missing teeth.[4]

a) Construct a scatterplot for these data, with `fluoride` as the x-variable and `caries` as the y-variable.

```
#load the data
water = read.csv("https://scofield.site/teaching/data/csv/harrington/water.csv")

#make a scatterplot
gf_point(caries ~ fluoride, data=water) |>
 gf_labs(x="Fluoride (ppm)",
         y="Caries per 100 children") |>
  gf_lm()
```



b) Do `fluoride` and `caries` appear to be positively or negatively associated? Explain your answer.

[4]These data appear in Table B21 in *Principles of Biostatistics*, 2nd ed. by Pagano and Gauvreau.

The two variables seem to be negatively associated, because the number of caries decreases as fluoride levels increase.

c) Later in the course, we will study methods for fitting a straight line to data.

i. If you were to add a straight line to the plot that you think best fits the data, what would be its x-intercept and y-intercept? (*Hint*: Be sure to look at the limits on the axes...)

Any reasonable answers for the x-intercept and y-intercept are acceptable, e.g. an x-intercept of (2.5, 0) and y-intercept of (0, 800). Note that R centers the plot around the data points. It would then appear that the x-intercept of a reasonable line is around 1.8.

ii. Based on the appearance of the plot, do you think that a straight line would be a reasonable way to represent these data? Explain your answer.

A straight line does not seem to be a good fit for the data. At low and high levels of fluoride, the data points are above the line; for intermediate levels, they are below the line. This suggests a non-linear association. In particular, it seems like a curve could fit the data much better.

Problem 5.

This problem features data from the *famuss* (*Functional SNPs Associated with Muscle Size and Strength*) study discussed in Chapter 1. The study examined the possible genetic determinants of skeletal muscle size and strength, before and after training.

This problem uses the following variables from the FAMuSS data:

- `ndrm.ch`: the percent change in strength in a participant's non-dominant arm, from before training and after.
- `drm.ch`: the percent change in strength in a particpant's dominant arm.
- `actn3.r577x`: the genotype at residue *r577x* within the *ACTN3* gene.
- `race`: race of the participant, with values stored as text strings.

The `famuss` dataset is in the `oibiostat` package.

a) Make a table of the genotypes for the SNP `actn3.r577x`.

```
#load the data
library(oibiostat)
data("famuss")

#make a table
tally(~actn3.r577x, data=famuss)
```

```
## actn3.r577x
##  CC  CT  TT
## 173 261 161
```

b) Construct a two-way table of `actn3.r577x` by race, with the genotypes in the columns of the table and races in the rows.

```r
#make a two-way table
tally(race ~ actn3.r577x, data=famuss)
```

```
##              actn3.r577x
## race           CC  CT  TT
##   African Am   16   6   5
##   Asian        21  18  16
##   Caucasian   125 216 126
##   Hispanic      4  10   9
##   Other         7  11   5
```

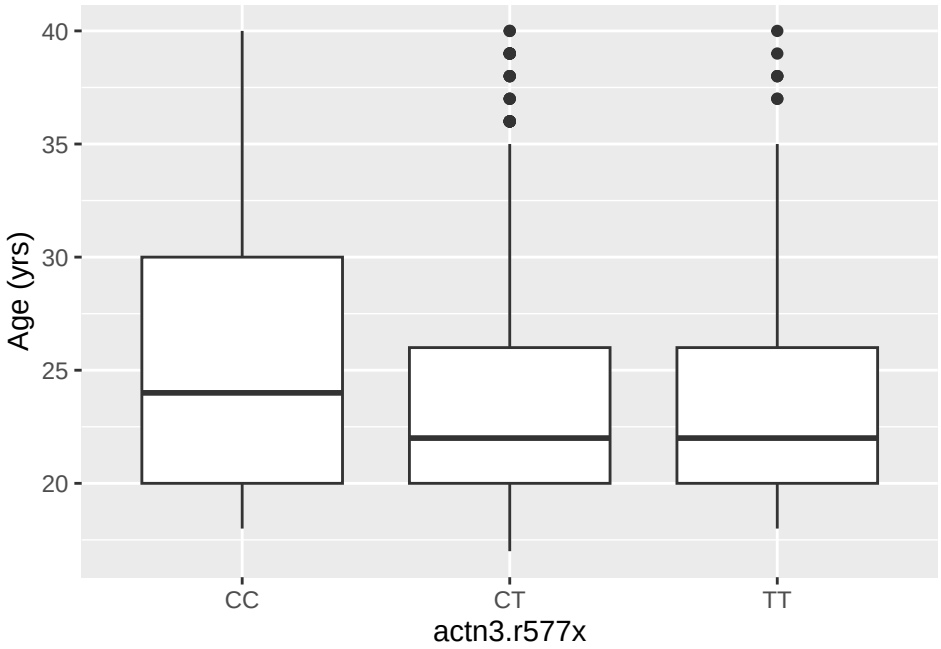
c) If you were to use numerical summaries to describe the `ndrm.ch` variable, would you prefer the mean and standard deviation or the five-number summary? Why?

The distribution of `ndrm.ch` is skewed right, which is (usually?) revealed from graphical displays (histogram or boxplot) or numerical summaries. The mean and standard deviation do not inherently convey information about skewness. Since a boxplot usually does convey skewness, the 5-number summary would as well.

d) Produce a graphical summary that shows the association between age and genotype at the SNP `actn3.r577x`. Describe what you see.

The median is highest in the CC group, which also has the largest range. Ages higher than 35 in the CT and TT groups are considered outliers, while those in the CC group are not. The median age in the CT and TT groups are similar.

```r
#make a boxplot
gf_boxplot(age ~ actn3.r577x, data=famuss) |>
  gf_labs(y="Age (yrs)", xlab = "Genotype at actn3.r577x")
```



Problem 6.

Does smoking have the same association with cardiovascular disease in women as it does in men? Epidemiologists typically use data from observational studies to investigate possible causes of disease.

Aortic stenosis is a narrowing or stricture of the aorta that impedes blood flow to the body.[5]

The dataset contains three variables, for 215 study participants:

- `disease`: coded `Yes` if stenosis is present, `No` if it is absent.
- `smoke`: coded `Smoker` if the participant is a current or former smoker, `NonSmoker` if the participant has never smoked.
- `sex`: coded as either `Male` or `Female`

Use the data from the file at

https://scofield.site/teaching/data/csv/harrington/stenosis.csv

to answer the following questions.

a) Construct a two-way table for smoking status and disease presence. What percentage of the 215 participants were both smokers and had aortic stenosis? This percentage is one component of the *joint distribution* of smoking and stenosis; what are the other three numbers of the joint distribution?

The percentage of participants that were both smokers and had aortic stenosis is 51/215 = 23.7%. The other three numbers are 67/215 = 31.1% (non-smoker, no stenosis), 54/215 = 25.1% (non-smoker, stenosis), and 43/215 = 20.0% (smoker, no stenosis).

```
#load the data
stenosis = read.csv("https://scofield.site/teaching/data/csv/harrington/stenosis.csv")

# the two-way table with marginal sums added
tally(disease ~ smoke, data=stenosis) |> addmargins()
```

```
##         smoke
## disease NonSmoker Smoker Sum
##     No         67     43 110
##     Yes        54     51 105
##     Sum       121     94 215
```

b) Among the smokers, what proportion have aortic stenosis? This number is a component of the conditional distribution of stenosis for the two categories of smokers. What proportion of non-smokers have aortic stenosis?

Among the smokers, 54.3% have aortic stenosis (51/94). Among the non-smokers, 44.6% have aortic stenosis (54/121).

c) Repeat part b) for males and females separately. To do this, first subset the data to create two datasets: one with only males, and one with only females. Include the tables in your solution. Are there any differences by sex in the proportion of smokers who suffer from aortic stenosis?

[5] The data appear in Table B20, *Principles of Biostatistics*, 2nd ed. by Pagano and Gauvreau.

Among female smokers, 42.4% have stenosis (14/33). Among female non-smokers, 38.2% have stenosis (29/76). Among male smokers, 60.7% have stenosis (37/61). Among male non-smokers, 55.6% have stenosis (25/45).

For both females and males, the percentage of smokers with stenosis is greater than the percentage of non-smokers with stenosis. However, the percentage of male smokers with stenosis is higher than the percentage of females with the disease. (This is also true for non-smokers.)

```
#subset males
females = subset(stenosis, sex == "Female")
males = subset(stenosis, sex == "Male")
tally(disease ~ smoke, data=females) |> addmargins()
```

```
##         smoke
## disease NonSmoker Smoker Sum
##      No         47     19  66
##      Yes        29     14  43
##      Sum        76     33 109
```

```
tally(disease ~ smoke, data=males) |> addmargins()
```

```
##         smoke
## disease NonSmoker Smoker Sum
##      No         20     24  44
##      Yes        25     37  62
##      Sum        45     61 106
```