Biostat
Homework 2

Q1 – Roles and types of measurements

Do Vittinghoff *et al*, Problem 2.1. (If you haven't already downloaded this book, recall it is available with your UWNetID login on Springer Link.)

> **Problem 2.1.** Classify each of the following variables as numerical or categorical. Then further classify the numerical variables as continuous or discrete, and the categorical variables as ordinal or nominal.
>
> (1) Gender
> (2) Race
> (3) Age (in years)
> (4) Age in categories (0–20, 21–35, 36–45, 45–60, 60–85, 85+)
> (5) Zipcode
> (6) Toxicity (mild, moderate, life-threatening, dead)
> (7) Number of hospitalizations in the past year
> (8) Change in HIV-RNA
> (9) Weeks on treatment
> (10) Treatment (placebo versus estrogen)

(1) Gender
Categorical variables, nominal
(2) Race
Categorical variables, nominal
(3) Age (in years)
Numerical variables, discrete
(4) Age in categories (0-20, 21-35, 36-45, 45-60, 60-85, 85+)
Categorical variables, ordinal
(5) Zipcode
Categorical variables, nominal
(6) Toxicity (mild, moderate, life-threatening, dead)
Categorical variables, ordinal
(7) Number of hospitalizations in the past year
Numerical variables, discrete
(8) Change in HIV-RNA
Numerical variables, continuous
(9) Weeks on treatment
Numerical variables, discrete
(10) Treatment (placebo versus estrogen)
Categorical variables, nominal

Q2 – Roles of measurements, connecting science and statistics

The class site contains the dataset on salaries, and a description of what it contains and why it was collected. For each of the scientific questions in that description, briefly characterize the type of statistical question as clustering cases, clustering variables, quantifying distributions within groups, comparing distributions across groups, or prediction, identifying any variable whose distribution is of interest and any groups that might be being compared. In your answer for each scientific question, state which variable(s) you would use, and if necessary any subsets of the data you would consider for the analysis.

Note you should assume the data are error-free. Also note that you are not required to *do* any analysis, though you are welcome to examine the data if you find this helpful.

- Does sex bias exist at the university in the most current year available (1995)?
  Comparing distributions across groups, as the statistical question is trying to summarize the difference between tendencies across female and male faculties of the university in 1995 to access sex bias.
  Variables of interest includes rank, year, salary, and admin (faculty members had administrative duties might mean working as chair or higher-level positions)
  Groups to compared: sex (female vs. male)
  I would mainly compare the salary, rank, year works (year) in the university between female and male. With these data, I could access the potential effect of sex on salary level, years employed, and ranked positions.

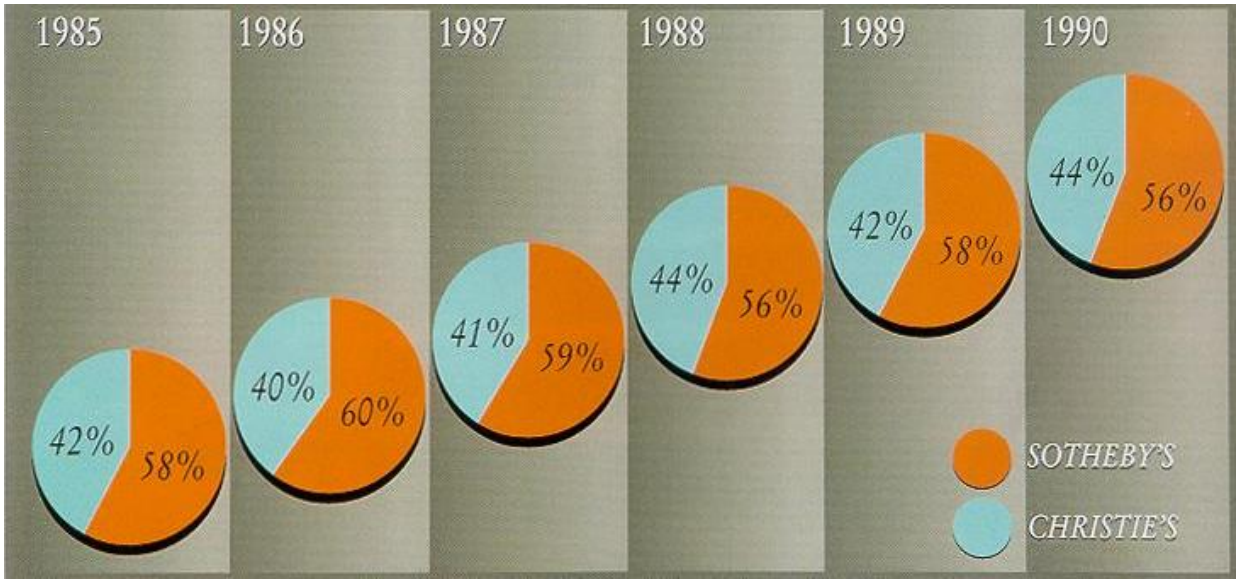
- Has sex bias existed in the starting salaries of faculty members (i.e., salaries in the year hired)
  Comparing distributions across groups, as the statistical question is trying to summarize the difference between tendencies across female and male faculty members to access sex bias in the starting salaries.
  Variables of interest includes year started (startyr) and salary
  Groups to compared: sex (female vs. male)
  I would compare the salary and year started in the university between female and male to observe any differences due to sex bias.

- Has sex bias existed in granting salary increases between 1976 and 1995?
  Comparing distributions across groups, as the statistical question is trying to summarize the difference in salary increases between 1976 and 1995 between female and male faculty members to access sex bias in granting salary increases.
  Variables of interest includes year works (year) and salary differences between 1976 and 1995
  Groups to compared: sex (female vs. male)
  I would compare the salary growth (salary difference between 1976 and 1995) in the university between female and male faculties to access any effect of sex bias on granting salary increases.

- Has sex bias existed in granting promotions from Associate Professor to Full Professor?
  Comparing distributions across groups, as the statistical question is trying to summarize the difference between promotions from Associate Professor to Full Professor across female and male faculty members to access sex bias. First would need to cluster groups of Associate Professor promoted to Full professor.
  Variables of interest includes year works, rank, and admin (faculty members had administrative duties might mean working as chair or higher-level positions)
  Groups to compared: sex (female vs. male)
  I would mainly compare the year works and rank (looking at Associate Professor and Full Professor and year works before promotion) between female and male to observe any differences due to sex bias.

Q3 – Displaying data

The following display is an early *infographic* – a display of data that is meant to be eye-catching as well as informative. It displays the 1985-1990 market share of Sotheby's and Christie's, two competing auction houses;



a. What information is conveyed by the horizontal positions of the pie charts?

Years from 1985 – 1990 (ascending from left to right)

b. What information is conveyed by the vertical positions of the pie charts?

Two auction houses Sotheby's and Christie's and their percentages in each year, lower data in earlier time (as the pie charts ascend)

c. Suggest a more straightforward graph, that would convey exactly the same information but with less redundancy. You may give a written description of the graph you would plot, or actually plot it; either earns full credit.

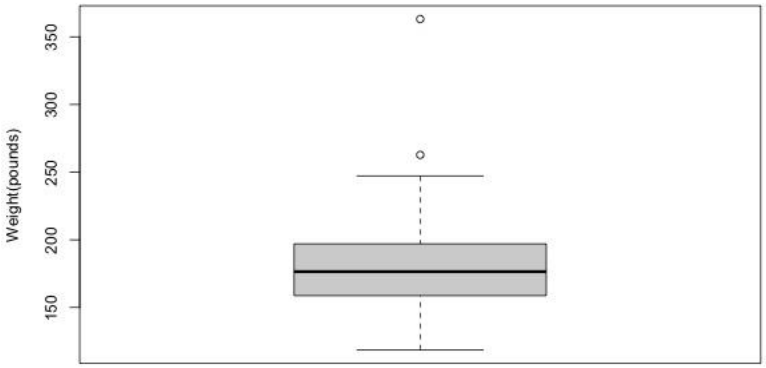
I would use a histogram to show the data. On a big histogram, x-axis would be year (1985, 1986, 1987, 1989, and 1990) and each year would have two bars (blue for Christie's and orange for Sotheby's). The y-axis would be percentage occupied by the auction houses (from 0 – 100%). With this histogram, all the data are represented in an easier way to read and be interpreted.

Q4 – 1D Summaries

Note: before you begin, please read the bullet point above about including your code (e.g. your .R script) as an appendix to your answers. Raw code and output should not appear in your written solutions to a, b or c.

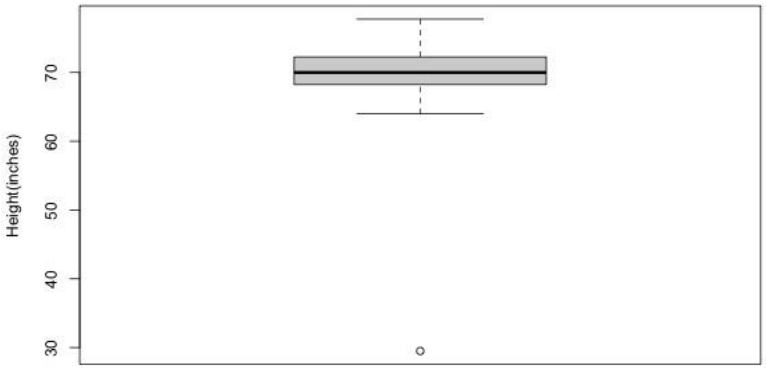
The bodyfat dataset on the class site contains data on the percent body fat, age, weight, height and ten body circumference measurements for 252 adult men.

a. Using appropriate graphical displays, examine the distribution of the men's body weights (in pounds). Are there unusual observations? If so, are they plausible values? Explain your answers.



From the boxplot, we see two outliers on the upper end of the plot, one at 262.75 pounds and the other at 363.15 pounds. These are still plausible values as there are heavy individuals weigh over 300 pounds in real life.

b. Using appropriate graphical displays, examine the distribution of the men's heights (in inches). Are there unusual observations? If so, are they plausible values? Explain your answers.



From the boxplot, we see one outlier on the lower end of the plot, which is at 29.5 inches. This is an unlikely value but still possible as the shortest men on earth ever is about 21 inches, so if the men in the data have certain health conditions, he might be considerably shorter than others in the dataset.

c. Give a table with at least three numeric measures of central tendency for the men's body weights and heights. If you found unusual observations in items a. and b., but they *were* plausible, keep them in the analysis. If any unusual observations were *not* plausible, discard them from the analysis. For both height and weight, your table should state how many unusual observations you identified and whether they were plausible.

I included the unusual observations in a. and b. as they are plausible.

Measure	Weight	Height
Mean	178.92	70.15
Median	176.50	70.00
Mode	184.25	71.50
# of unusual observations	2	1

Q5 – Looking ahead

In preparation for next week's lecture (Oct 18th-22nd) read slides 2.99-2.175 in Chapter 2. Also read to the end of Vittinghoff Chapter 2.

Done!

Appendix

```
### BIOST 514 HW2
### Q4
### part a

### reading in the data
bodyfat <- read_csv("~/Desktop/UW MS Biostat/BIOST FALL 2021/Biost 514/R
dataset/bodyfat.csv")

### plot a boxplot to show distribution of the men's body weights (in pounds)
jpeg("~/Desktop/UW MS Biostat/BIOST Fall 2021/Biost 514/R work/HW 2/HW
2jpg_weight.jpg", width=600,height=400)
weight <- bodyfat[,c("weight")]
boxplot(bodyfat$weight, ylab="Weight(pounds)")
dev.off()

### part b
### plot a boxplot to show distribution of the men's heights (in inches)
jpeg("~/Desktop/UW MS Biostat/BIOST Fall 2021/Biost 514/R work/HW 2/HW
2jpg_height.jpg", width=600,height=400)
height <- bodyfat[,c("height")]
boxplot(bodyfat$height, ylab="Height(inches)")
dev.off()

### calculate mean/median of height/weight
mean(weight)
median(weight)
mean(height)
median(height)

### generate the mode of height/weight
getmode <- function(v) {
        uniqv <- unique(v)
        uniqv[which.max(tabulate(match(v, uniqv)))]
}
w <- bodyfat[,c("weight")]
result <- getmode(w)
print(result)

h <- bodyfat[,c("height")]
result <- getmode(h)
print(result)

### create matrix with 3 rows
tab <- matrix(c(178.92,70.15,176.50,70.00,184.25,71.50), nrow = 3, byrow=TRUE)

### define column names and row names of matrix
colnames(tab) <- c('Weight', 'Height')
rownames(tab) <- c('Mean', 'Median', 'Mode')
### convert matrix to table
```

```
tab <- as.table(tab)

### view table
tab
```