Problem set 3 sample solutions

2024-09-27

In these exercises, we will explore a subset of the NHANES dataset to investigate potential differences in systolic blood pressure across groups defined by self reported race.

Instructions

- For each exercise, we want you to write a single line of code using the pipe (`|>`) to chain together multiple operations. This doesn't mean the code must fit within 80 characters or be written on a single physical line, but rather that the entire sequence of operations can be executed as one continuous line of code without needing to assign intermediate values or create new variables.

  For example, these are three separate lines of code:

```
x <- 100; x <- sqrt(x); log10(x)
```

`Whereas this is considered one line of code using the pipe:`

```
100 |>
  sqrt() |>
  log10()
```

- Generate an html document that shows the code for each exercise.
- For the exercises that ask to generate a graph, show the graph as well.
- For exercises that require you to display tabular results, use the `kable` function to format the output as a clean, readable table. Do not display the raw dataframe directly—only show the nicely formatted table using `kable`.
- Use only two significant digits for the numbers displayed in the tables.

- Submit both the html and the qmd files using Git.
- You will need the following libraries:

```
library(dplyr)
library(tidyr)
library(forcats)
library(ggplot2)
library(knitr)
library(NHANES)
options(digits = 2)
```

- The `.qmd` file must be able to render properly on the TFs' computers. They will already have the necessary packages installed, so there is no need to include code for installing packages. Just focus on writing the code that uses these packages.

Exercises

1. Filter the `NHANES` data to only include survey year 2011-2012. Save the resulting table in `dat`. This table should have 5,000 rows and 76 columns.

```
# Sample answer
dat <- NHANES |> filter(SurveyYr == "2011_12")
```

2. Compute the average and standard deviation (SD) for the *combined systolic blood pressure* (SBP) reading for males and females separately. Show us a data frame with two rows (female and male) and two columns (average and SD).

```
# Sample answer
dat |>
  group_by(Gender) |>
  summarize(average = mean(BPSysAve, na.rm = TRUE),
            standard_deviation = sd(BPSysAve, na.rm = TRUE)) |>
  kable()
```

Gender	average	standard_deviation
female	117	18
male	120	17

3. Because of the large difference in the average between males and females, we will perform the rest of the analysis separately for males and females.

Compute the average and SD for SBP for each race variable in column `Race3` for females and males separately. The resulting table should have four columns for sex, race, average, and SD, respectively, and 12 rows (one for each strata). Arrange the result from highest to lowest average.

```
# Sample answer
dat |>
  group_by(Gender, Race3) |>
  summarize(average = mean(BPSysAve, na.rm = TRUE),
            standard_deviation = sd(BPSysAve, na.rm = TRUE)) |>
  arrange(desc(average)) |>
  kable()
```

```
`summarise()` has grouped output by 'Gender'. You can override using the
`.groups` argument.
```

Gender	Race3	average	standard_deviation
male	Black	122	19
male	White	121	17
male	Other	121	14
female	Black	120	19
male	Hispanic	120	15
male	Asian	118	15
female	White	118	18
male	Mexican	116	15
female	Hispanic	114	19
female	Other	114	18
female	Asian	114	18
female	Mexican	111	15

4. Repeat the previous exercise but add two columns to the final table to show a 95% confidence interval. Specifically, add columns with the lower and upper bounds of the interval with names `lower` and `upper`, respectively. The formula for these values is

$$\bar{X} \pm 1.96\, s/\sqrt{n}$$

with $\bar{X}$ the sample average and s the sample standard deviation. This table will simply add two more columns to the table generated in the previous exercise: one column for the lower and upper bound, respectively.

```r
# Sample answer
dat |>
  group_by(Gender, Race3) |>
  summarize(n = n(),
            average = mean(BPSysAve, na.rm = TRUE),
            standard_deviation = sd(BPSysAve, na.rm = TRUE)) |>
   mutate(lower = average - 1.96*standard_deviation/sqrt(n),
          upper = average + 1.96*standard_deviation/sqrt(n)) |>
  arrange(desc(average)) |>
  kable()
```

```
`summarise()` has grouped output by 'Gender'. You can override using the
`.groups` argument.
```

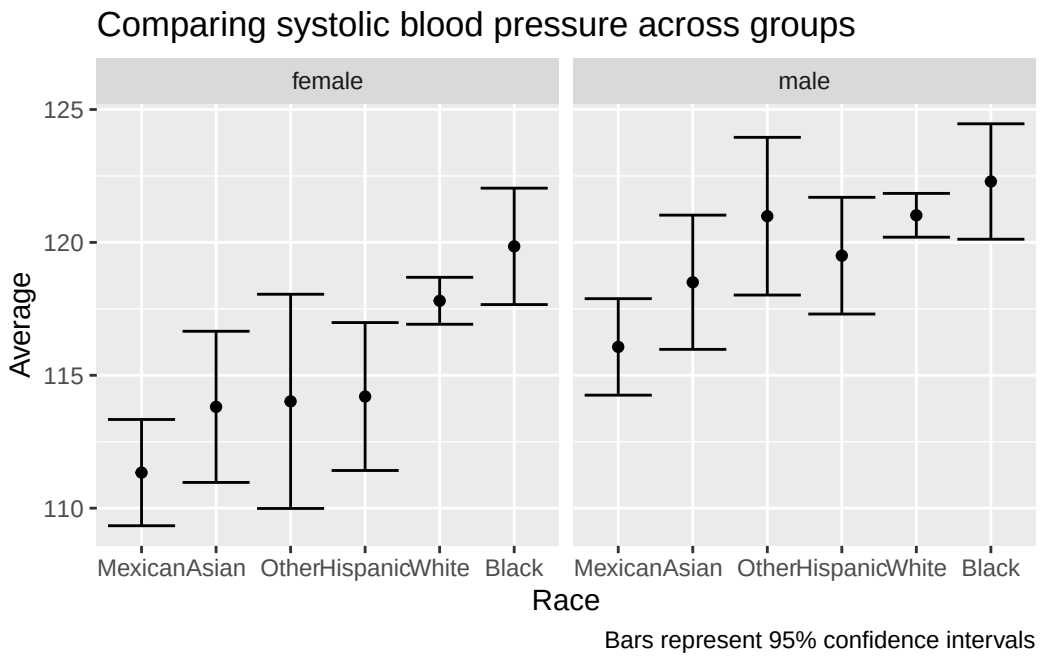
Gender	Race3	n	average	standard_deviation	lower	upper
male	Black	297	122	19	120	124
male	White	1547	121	17	120	122
male	Other	85	121	14	118	124
female	Black	292	120	19	118	122
male	Hispanic	168	120	15	117	122
male	Asian	141	118	15	116	121
female	White	1588	118	18	117	119
male	Mexican	267	116	15	114	118
female	Hispanic	182	114	19	111	117
female	Other	73	114	18	110	118
female	Asian	147	114	18	111	117
female	Mexican	213	111	15	109	113

5. Make a graph of showing the results from the previous exercise. Specifically, plot the averages for each group as points and confidence intervals as error bars (use the geometry `geom_errorbar`). Order the groups from lowest to highest average (the average of the males and females averages). Use `facet_wrap` to make a separate plot for females and males. Label your axes with *Race* and *Average* respectively, add the title *Comparing systolic blood pressure across groups*, and the caption *Bars represent 95% confidence intervals.*

```r
# Sample answer
dat |>
  group_by(Gender, Race3) |>
```

```r
  summarize(n = n(),
            average = mean(BPSysAve, na.rm = TRUE),
            standard_deviation = sd(BPSysAve, na.rm = TRUE)) |>
   mutate(lower = average - 1.96*standard_deviation/sqrt(n),
          upper = average + 1.96*standard_deviation/sqrt(n)) |>
  mutate(Race3 = reorder(Race3, average)) |>
  ggplot(aes(Race3, average, ymin = lower, ymax = upper)) +
  geom_point() +
  geom_errorbar() +
  labs(title = "Comparing systolic blood pressure across groups", x = "Race",
   ↪  y = "Average",
       caption = "Bars represent 95% confidence intervals") +
  facet_wrap(~Gender)
```

```
`summarise()` has grouped output by 'Gender'. You can override using the
`.groups` argument.
```

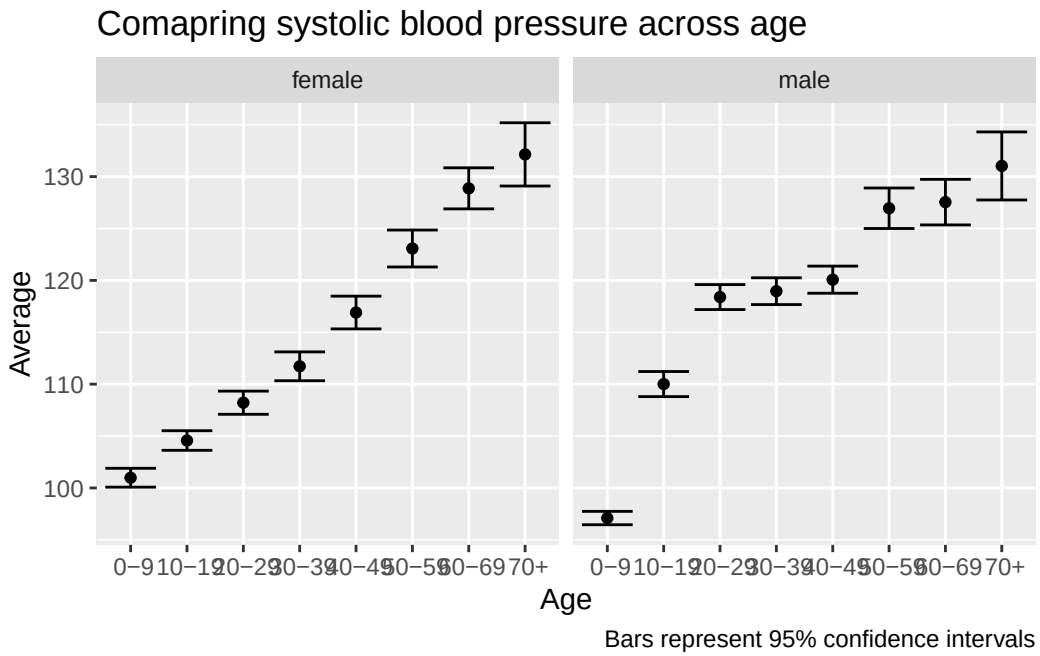


6. In the plot above we see that the confidence intervals don't overlap when comparing the `White` and `Mexican` groups. We also see a substantial difference between `Mexican` and `Hispnanic`. Before concluding that there is a difference between groups, we will explore if differences in age, a very common *confounder*, explain the differences.

Create table like the one in the previous exercise but show the average SBP by sex and age group (`AgeDecade`). The the groups are order chronologically. As before make a separate plot for males and females. Make sure to filter our observations with no `AgeDecade` listed.

```
# Sample answer
dat |>
  filter(!is.na(AgeDecade)) |>
  group_by(Gender, AgeDecade) |>
  summarize(n = n(),
            average = mean(BPSysAve, na.rm = TRUE),
            standard_deviation = sd(BPSysAve, na.rm = TRUE)) |>
  mutate(lower = average - 1.96*standard_deviation/sqrt(n),
         upper = average + 1.96*standard_deviation/sqrt(n)) |>
  ggplot(aes(AgeDecade, average, ymin = lower, ymax = upper)) +
  geom_point() +
  geom_errorbar() +
  labs(title = "Comapring systolic blood pressure across age", x = "Age", y =
   ↪  "Average",
       caption = "Bars represent 95% confidence intervals") +
  facet_wrap(~Gender)
```

```
`summarise()` has grouped output by 'Gender'. You can override using the
`.groups` argument.
```

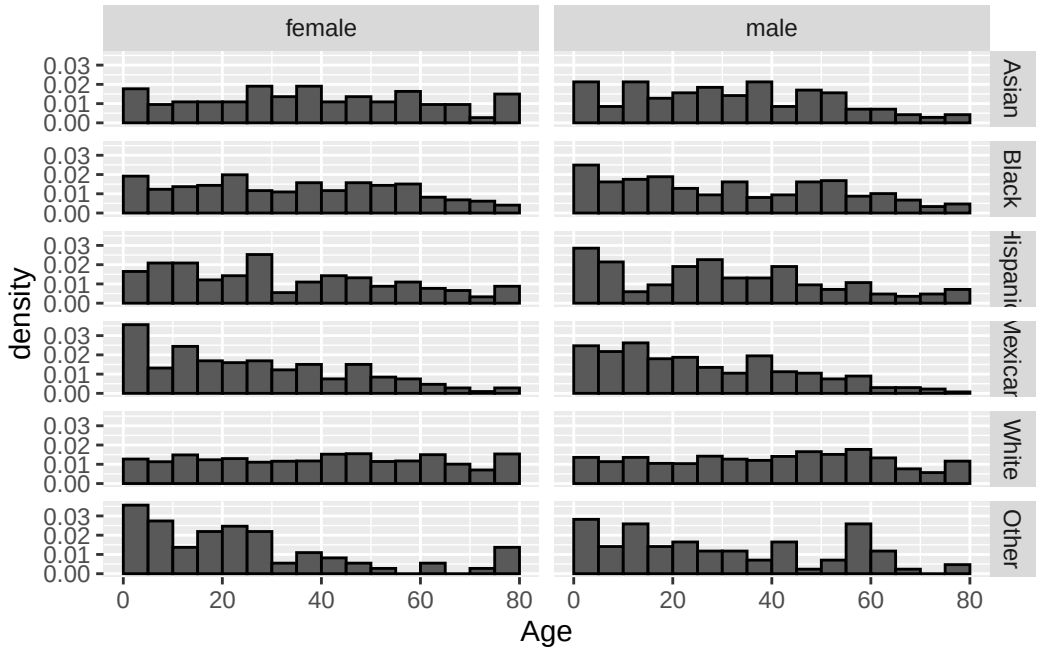


7. We note that for both males and females the SBP increases with age. To explore if age is indeed a confounder we need to check if the groups have different age distributions.

Explore the age distributions of each `Race3` group to determine if the groups are comparable. Make a histogram of `Age` for each `Race3` group and stack them vertically. Generate two columns of graphs for males and females, respectively. In the histograms, create bins increments of 5 years up to 80.

Below the graph, comment on what notice about the age distributions and how this can explain the difference between the *White* and *Mexican* groups.

```
# Sample answer
dat |>
  ggplot(aes(x = Age, y = after_stat(density))) +
  geom_histogram(breaks = seq(0, 80, 5), color = "black") +
  facet_grid(Race3 ~ Gender)
```



8. Summarize the results shown in the graph by compute the median age for each `Race3` group and the percent of individuals that are younger than 18. Order the rows by median age. The resulting data frame should have 6 rows (one for each group) and three columns to denote group, median age, and children respectively.

```r
# Sample answer
dat |> group_by(Race3) |>
  summarize(median_age = median(Age),
            children = mean(Age < 18)*100) |>
  arrange(median_age) |>
  kable()
```

Race3	median_age	children
Mexican	22	40
Other	22	38
Hispanic	28	31
Black	33	29
Asian	35	25
White	41	22

Given these results provide an explanation for the difference in systolic pressure is lower when comparing the `White` and `Mexican` groups.

9. When the age distribution between two populations we can't conclude that there are differences in SBP based just on the population averages. The observed differences are likely due to age differences rather than genetic differences. We will therefore stratify by group and then compare SBP. But before we do this, we might need redefine `dat` to avoid small groups.

Compute the number of observations in each gender, age group and race combination. Show the groups with less than five observations. Make sure to remove the rows with no BPSysAve measurments before calculating the number of observations. Show a table with four columns representing gender, age strate, group, and the number of individuals in that group. Make sure to include combinations with 0 individuals (hint: use `complete`).

```r
# Sample answer
dat |> filter(!is.na(BPSysAve) & !is.na(AgeDecade)) |>
  count(Gender, AgeDecade, Race3) |>
  complete(Gender, AgeDecade, Race3, fill = list(n = 0)) |>
  filter(n < 5) |>
  kable()
```

Gender	AgeDecade	Race3	n
female	0-9	Asian	2
female	40-49	Other	4
female	50-59	Other	2
female	60-69	Other	2
female	70+	Mexican	3
male	0-9	Asian	2
male	0-9	Other	2
male	70+	Asian	3
male	70+	Mexican	3
male	70+	Other	0

10. Based on the observations made in the previous exercise, we will redefine `dat` but with the following:

    - As before, include only survey year 2011-2012.
    - Remove the observations with no age group reported.
    - Remove the 0-9 age group.
    - Combine the 60-69 and 70+ ageroups into a 60+ group.
    - Remove observations reporting `Other` in `Race3`.
    - Rename the variable `Race3` to `Race`.

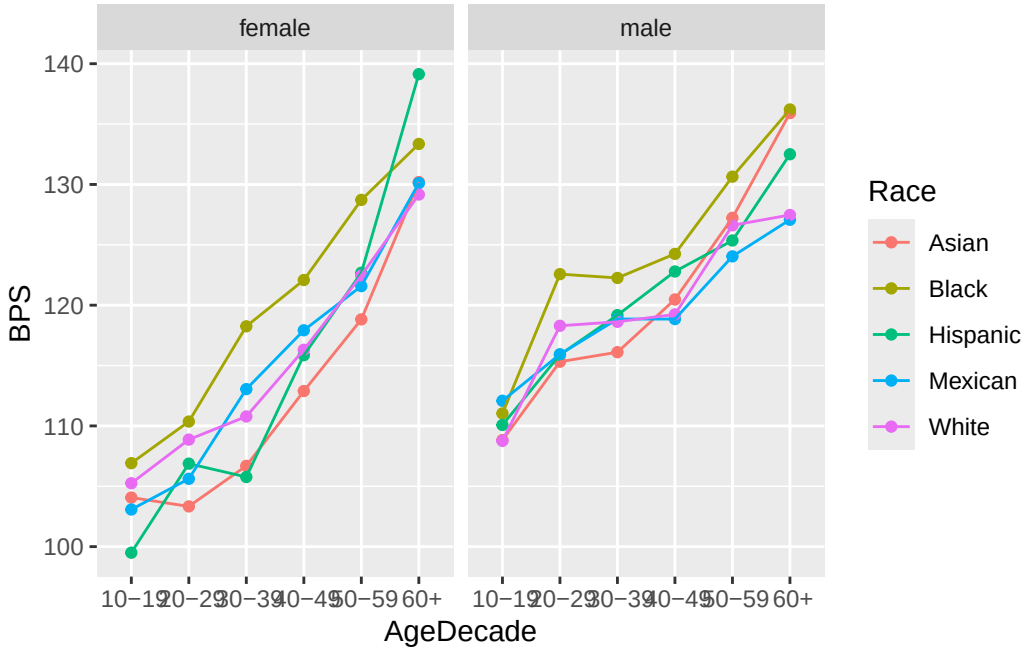
    Hints:

    - Note that the levels in `AgeDecade` start with a space.
    - You can use the `fct_collapse` function in the **forcats** to combine factors.

```
# Sample answer
dat <- NHANES |> filter(SurveyYr == "2011_12" &
                          AgeDecade != " 0-9" &
                          Race3 != "Other") |>
  mutate(AgeDecade = fct_collapse(AgeDecade, " 60+" = c(" 60-69", " 70+"))) |>
  rename(Race = Race3)
```

11. Crete a plot that shows the averege BPS for each age decade. Show the different race groups with color and lines joining them. Generate a two plots, one for males and one for females.

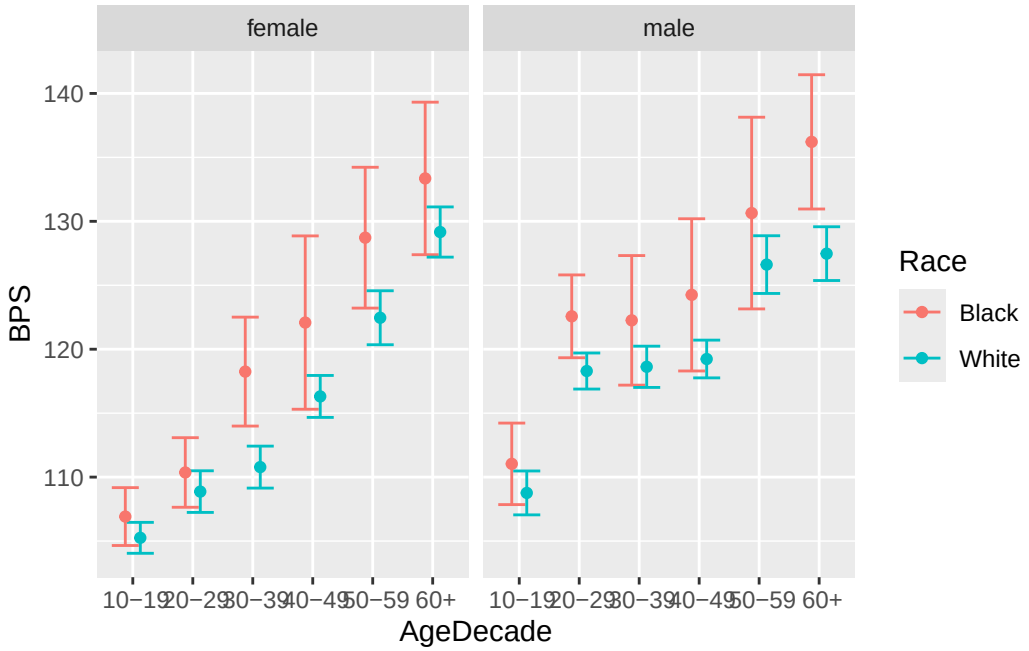
```r
# Sample answer
dat |>
  group_by(Gender, AgeDecade, Race) |>
  summarize(BPS = mean(BPSysAve, na.rm = TRUE), .groups = "drop") |>
  ggplot(aes(AgeDecade, BPS, color = Race, group = Race)) +
  geom_point() +
  geom_line() +
  facet_wrap(~Gender)
```



12. Based on the plot above pick two groups that you think are consistently different and remake the plot from the previous exercise but just for these two groups, add confidence intervals, and remove the lines. Put the confidence intervals for each age strata next to each other and use color to represent the two groups. Comment on your finding.

```r
# Sample answer
dat |>
  filter(Race %in% c("Black", "White")) |>
  group_by(Gender, AgeDecade, Race) |>
  summarize(n = n(),
            BPS = mean(BPSysAve, na.rm = TRUE),
            standard_deviation = sd(BPSysAve, na.rm = TRUE), .groups =
              ↪  "drop") |>
  mutate(lower = BPS - 1.96*standard_deviation/sqrt(n),
```

```
        upper = BPS + 1.96*standard_deviation/sqrt(n)) |>
  ggplot(aes(AgeDecade, BPS, color = Race, group = Race)) +
  geom_errorbar(aes(ymin = lower, ymax = upper), position =
   ↪  position_dodge(.5)) +
  geom_point(position = position_dodge(.5)) +
  facet_wrap(~Gender)
```



13. For the two groups that you selected above compute the difference in average BPS between the two groups for each age strata. Show a table with three columns representing age strata, difference for females, difference for males.

```
# Sample answer
dat |>
  filter(Race %in% c("Black", "White")) |>
  group_by(Gender, AgeDecade, Race) |>
  summarize(BPS = mean(BPSysAve, na.rm = TRUE), .groups = "drop") |>
  pivot_wider(names_from = Race, values_from = BPS) |>
  mutate(difference = Black - White) |>
  select(Gender, AgeDecade, difference) |>
  pivot_wider(names_from = Gender, values_from = difference) |>
  kable()
```

AgeDecade	female	male
10-19	1.7	2.3
20-29	1.5	4.3
30-39	7.5	3.6
40-49	5.8	5.0
50-59	6.3	4.0
60+	4.2	8.7