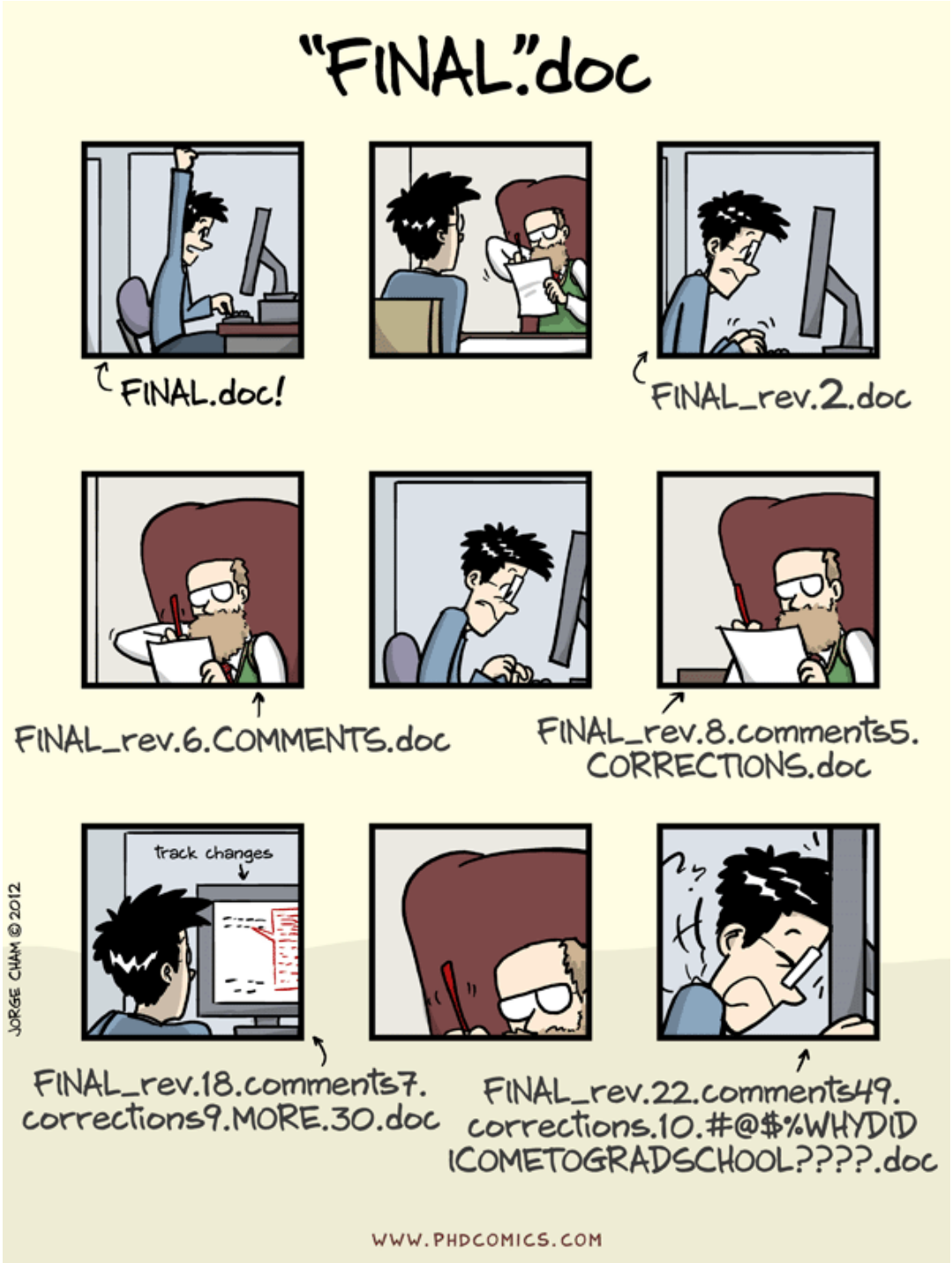
431 Class 05

thomaselove.github.io/431

2020-09-08

Today's Agenda

- Reviewing the Surveys from Lab 01
- A Few Thoughts from Leek *Elements of Data Analytic Style* (Chapters 5, 9 and 10)
- Getting Started with NHANES


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

To the Surveys from Lab 01

Dr. Love used R Markdown to explore your responses to the survey (part 4) of Lab 01, which we'll discuss at some length in class today.

https://github.com/THOMASELOVE/431-2020/blob/master/labs/lab01/survey-results-2020/README.md

Leek: Elements of Data Analytic Style

Leek *Elements of Data Analytic Style* (finish by Oct 2)

- Chapters 2-4 should be very helpful for project (Data analytic question, Tidying data, Checking data)
- 6-8 are more about inference and modeling
- 11-12 on Presenting Data and Reproducibility
- 13 touches on a few matters of form
- 14 is a Data Analysis Checklist

Leek Chapter 5: Exploratory Analysis

- EDA To understand properties of the data and discover new patterns
- Visualize and inspect qualitative features rather than a huge table of raw data

1. Make big data as small as possible as quickly as possible
2. Plot as much of the actual data as you can
3. For large data sets, subsample before plotting
4. Use log transforms for ratio measurements
5. Missing values can have a mighty impact on conclusions

Leek: Chapter 9 Written Analyses

Elements: title, introduction/motivation, description of statistical tools used, results with measures of uncertainty, conclusions indicating potential problems, references

1. What is the question you are answering?
2. Lead with a table summarizing your tidy data set (critical to identify data versioning issues)
3. For each parameter of interest report an estimate and measure of uncertainty on the scientific scale of interest
4. Summarize the importance of reported estimates
5. Do not report every analysis you performed

Leek: Chapter 10 Creating Figures

Communicating effectively with figures is non-trivial. The goal is clarity.

> *When viewed with an appropriately detailed caption, (a figure should) stand alone without any further explanation as a unit of information.*

1. Humans are best at perceiving position along a single axis with a common scale
2. Avoid chartjunk (gratuitous flourishes) in favor of high-density displays
3. Axis labels should be large, easy to read, in plain language
4. Figure titles should communicate the plot's message
5. Use a palette (like `viridis`) that color-blind people can see (and distinguish) well

Check out Karl Broman's excellent presentation on displaying data badly at https://github.com/kbroman/Talk_Graphs

What about R for Data Science?

I'd be trying to get through *Explore* (sections 2-8) before our first Quiz.

- Section 11 on Data import
- Section 18 on Pipes
- Section 27 on R Markdown and maybe 28 on Graphics for communication

Today's Packages

The R packages we're using are `NHANES`, `magrittr`, `janitor` and `tidyverse`.

```
library(NHANES); library(magrittr)
library(janitor); library(tidyverse)

theme_set(theme_bw())
```

I always load the `tidyverse` last, and `theme_bw()` is one of my two favorite ggplot themes.

- Also, I set the code chunk to `message = FALSE` when I want to hide several messages that come up when loading.

So my package loading code chunk header (inside the brackets) looks like:

```
{r load_packages, message = FALSE}
```

CWRU Colors

CWRU's color guide (see the README) specifies CWRU blue and CWRU gray

```
cwru.blue <- '#0a304e'
cwru.gray <- '#626262'
```

I'd like to use those later today.

Today's Example

We're going to work with subjects who participated in NHANES: National Health and Nutrition Examination Survey.

> *The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.*

Use `?NHANES` to learn more about the data. The NHANES package contains 5000 observations from each of the 2009-10 and 2011-12 administrations.

- See the Course Notes for a related series of examples.
- Baumer, Kaplan and Horton (2017) *Modern Data Science with R* have developed similar examples.

Creating an nh2 data set

```
set.seed(20200908) # so we can get the same sample again

nh2 <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    select(ID, SurveyYr, Age, Height, Weight, BMI, Pulse,
           SleepHrsNight, BPSysAve, BPDiaAve, Gender,
           PhysActive, SleepTrouble, Smoke100,
           Race1, HealthGen, Depressed) %>%
    rename(SleepHours = SleepHrsNight, Sex = Gender,
           SBP = BPSysAve, DBP = BPDiaAve) %>%
    filter(Age > 20 & Age < 80) %>% ## ages 21-79 only
    drop_na() %>% # removes all rows with NA
    sample_n(., size = 1000) %>% # sample 1000 rows
    clean_names() # from the janitor package (snake case)
```

Codebook for `nh2` (ID and Quantitative Variables)

Name	Description
`id`	Identifying code for each subject
`survey_yr`	2011_12 for all, indicates administration date
`age`	Age in years at screening of subject (must be 21-79)
`height`	Standing height in cm
`weight`	Weight in kg
`bmi`	Body mass index ($\frac{weight}{(height_{meters})^2}$ in $\frac{kg}{m^2}$)
`pulse`	60 second pulse rate
`sleep_hrs`	Self-reported hours (usually gets) per night
`sbp`	Systolic Blood Pressure (mm Hg)
`dbp`	Diastolic Blood Pressure (mm Hg)

Codebook for nh2 (Categorical Variables)

Binary Variables

Name	Levels	Description
sex	F, M	Sex of study subject
phys_active	No, Yes	Moderate or vigorous sports/recreation?
sleep_trouble	No, Yes	Has told a provider about trouble sleeping?
smoke100	No, Yes	Smoked at least 100 cigarettes in lifetime?

Multi-Categorical Variables

Name	Levels	Description
race1	5	Self-reported Race/Ethnicity
health_gen	5	Self-reported overall general health
depressed	3	How often subject felt depressed in last 30d

A Look at Body-Mass Index

Let's look at the *body-mass index*, or BMI. The definition of BMI for adult subjects (which is expressed in units of kg/m^2) is:

$$\text{BMI} = \frac{\text{weight in kg}}{(\text{height in meters})^2} = 703 \times \frac{\text{weight in pounds}}{(\text{height in inches})^2}$$

BMI is, essentially, a measure of a person's *thinnness* or *thickness*.

- BMI from 18.5 to 25 indicates optimal weight
- BMI below 18.5 suggests person is underweight
- BMI above 25 suggests overweight.
- BMI above 30 suggests obese.

A First Set of Exploratory Questions

Variables of Interest: `bmi`, `phys_active`, `health_gen`, `pulse`

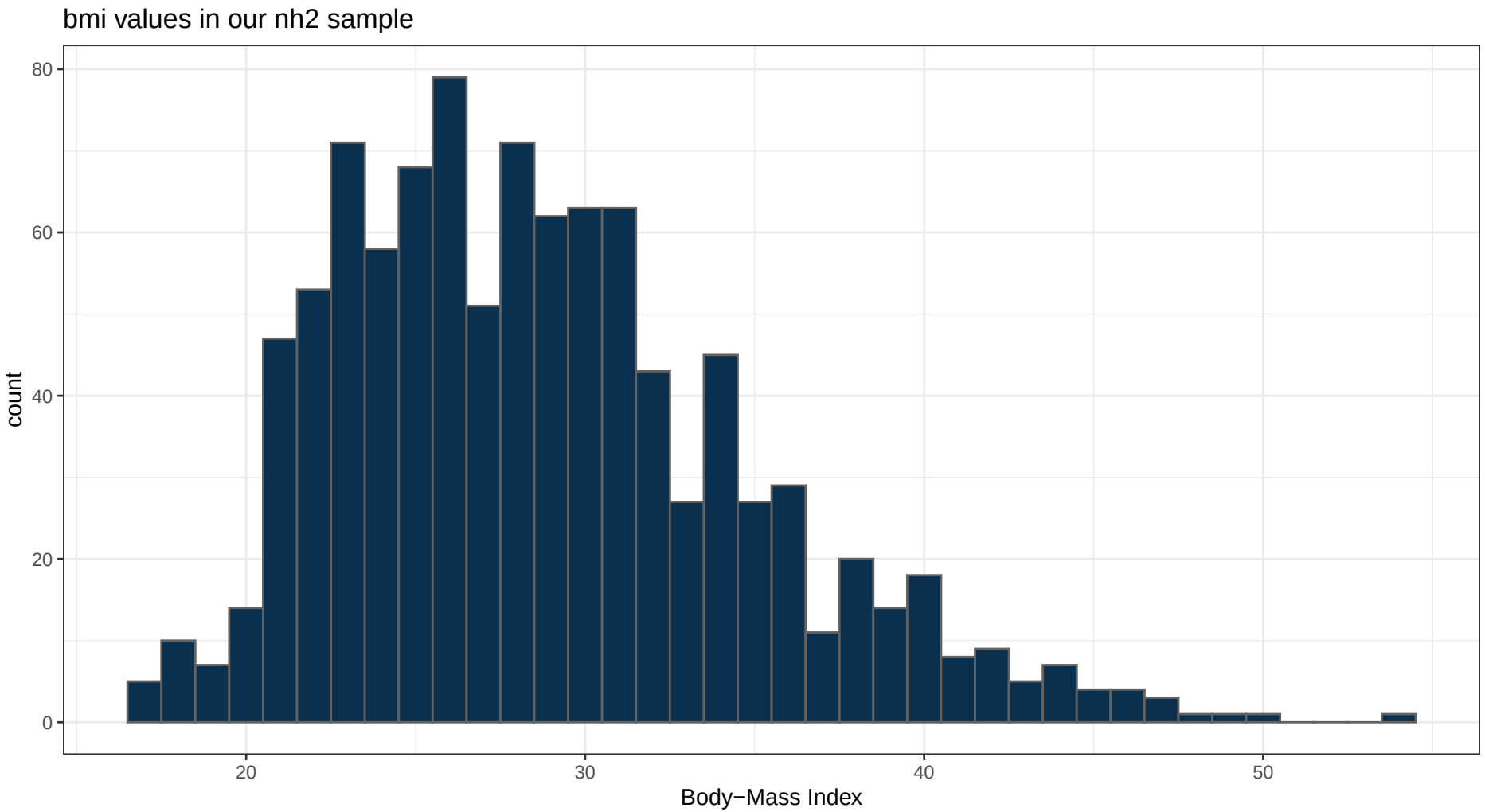
1. What is the distribution of `bmi` in our `nh2` sample of adults?
2. How does the distribution of `bmi` vary by whether the subject is physically active?
3. How does the distribution of `bmi` vary by the subject's self-reported general health?
4. What is the association between `bmi` and the subject's pulse rate?
5. Does that `bmi-pulse` association differ in subjects who are physically active, and those who are not?

Note: These are NOT what anyone would call research questions, which involve generating scientific hypotheses, among other things. These are merely triggers for visualizations and (small) analyses.

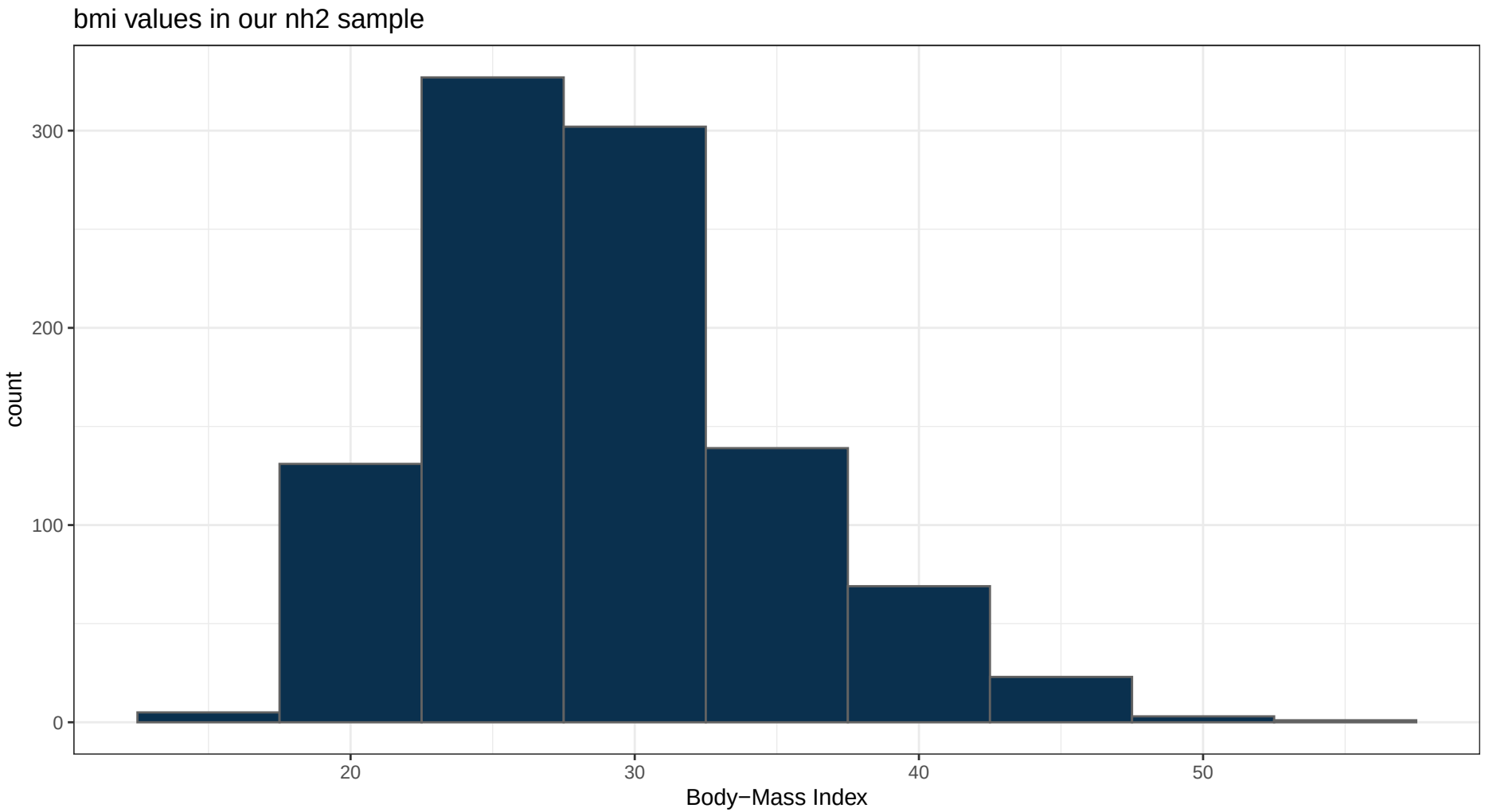
Histogram of BMI in nh2 with binwidth = 1

```
ggplot(nh2, aes(x = bmi)) +
    geom_histogram(binwidth = 1, fill = cwru.blue,
                   col = cwru.gray) +
    labs(title = "bmi values in our nh2 sample",
         x = "Body-Mass Index")
```

Histogram of BMI in `nh2` with binwidth = 1



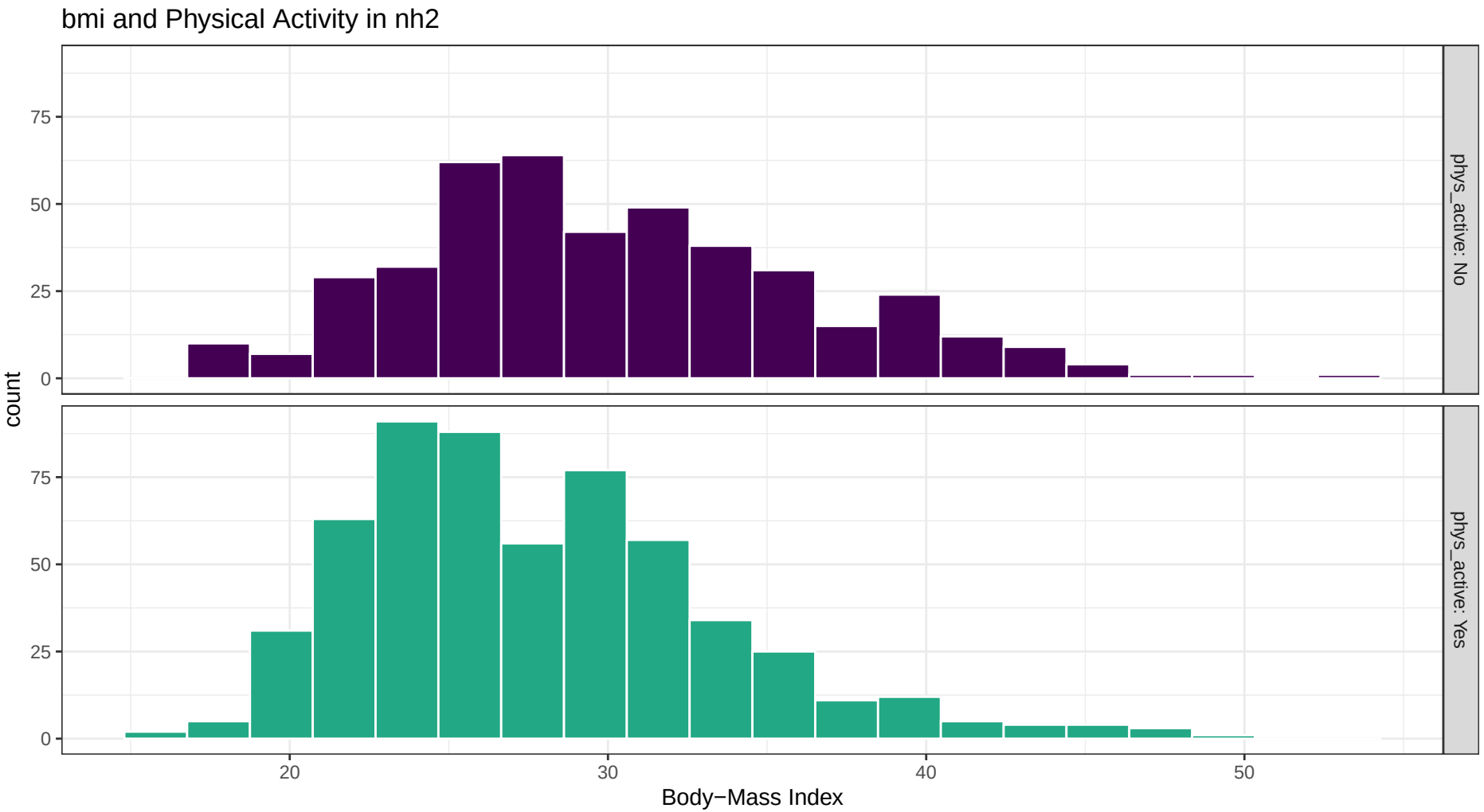
Histogram of BMI in `nh2` with binwidth = 5



BMI Histograms faceted by Physical Activity Status

```
ggplot(nh2, aes(x = bmi, fill = phys_active)) +
    geom_histogram(bins = 20, col = "white") +
    labs(title = "bmi and Physical Activity in nh2",
         x = "Body-Mass Index") +
    scale_fill_viridis_d(end = 0.6) +
    guides(fill = FALSE) +
    theme_bw() +
    facet_grid(phys_active ~ ., labeller = "label_both")
```

BMI Histograms faceted by Physical Activity Status



Average BMI by Physical Activity Status, I

Create a tibble that helps us answer:

- What is the "average" BMI in each activity group?
- How many people fall into each activity group?

```
nh2 %>%
    group_by(phys_active) %>%
    summarize(count = n(), mean(bmi), median(bmi))
```

```
`summarise()` ungrouping output (override with `.groups` argum
```

```
# A tibble: 2 x 4
  phys_active count `mean(bmi)` `median(bmi)`
  <fct>       <int>       <dbl>          <dbl>
1 No            431        30.1           29.4
2 Yes           569        27.8           26.9
```

Average BMI by Physical Activity Status, II

Making this look a bit more presentable as a table... (and using `message = FALSE`)

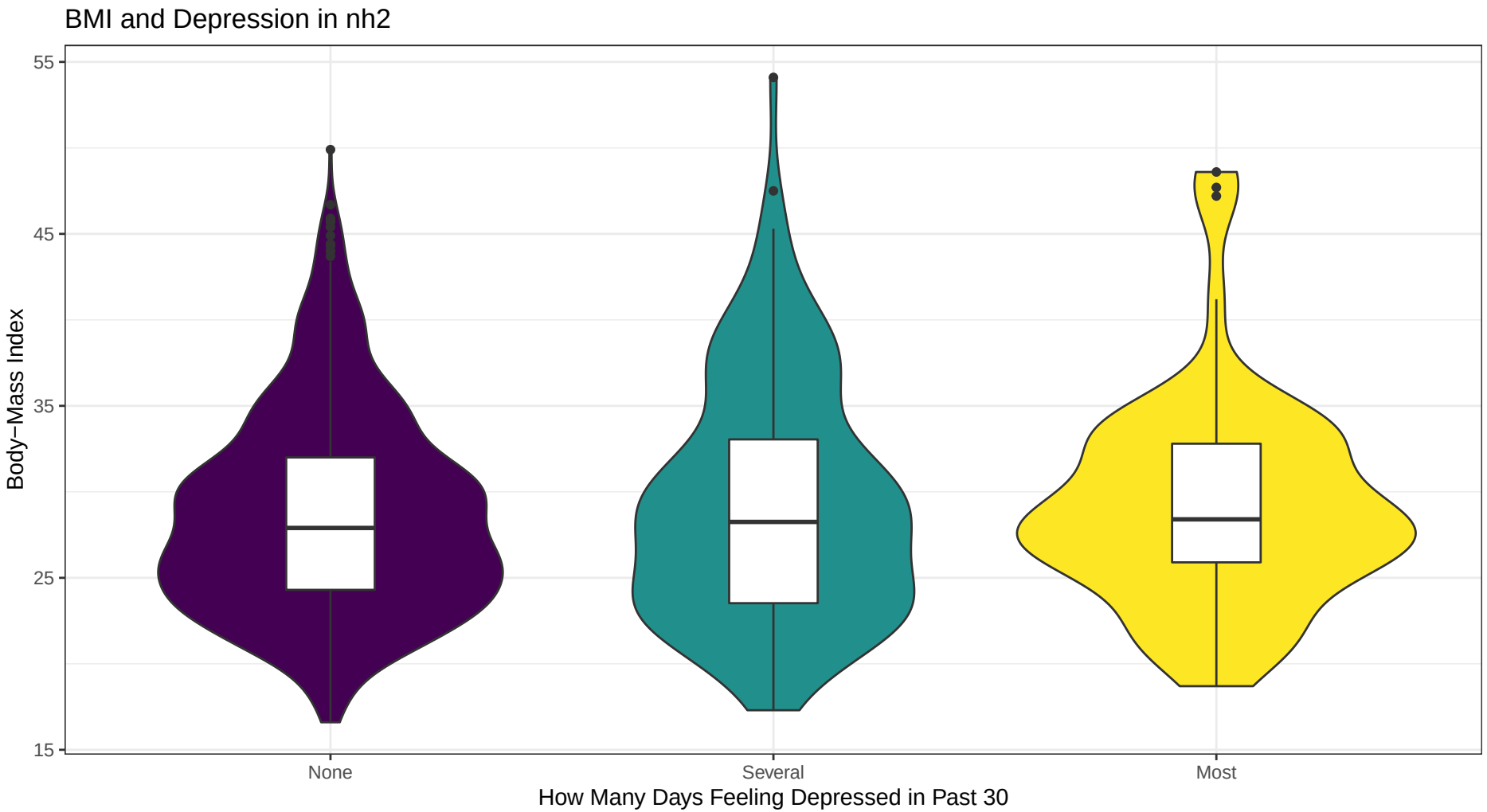
```
nh2 %>%
    group_by(phys_active) %>%
    summarize("Count" = n(),
              "Mean(BMI)" = round(mean(bmi),2),
              "Median(BMI)" = median(bmi)) %>%
    knitr::kable()
```

phys_active	Count	Mean(BMI)	Median(BMI)
No	431	30.12	29.4
Yes	569	27.84	26.9

BMI by Depression Status: Violin Plot

```
ggplot(nh2, aes(x = depressed, y = bmi, fill = depressed)) +
    geom_violin() +
    geom_boxplot(width = 0.2, fill = "white") +
    labs(title = "BMI and Depression in nh2",
         y = "Body-Mass Index",
         x = "How Many Days Feeling Depressed in Past 30") +
    scale_fill_viridis_d() +
    guides(fill = FALSE) +
    theme_bw()
```

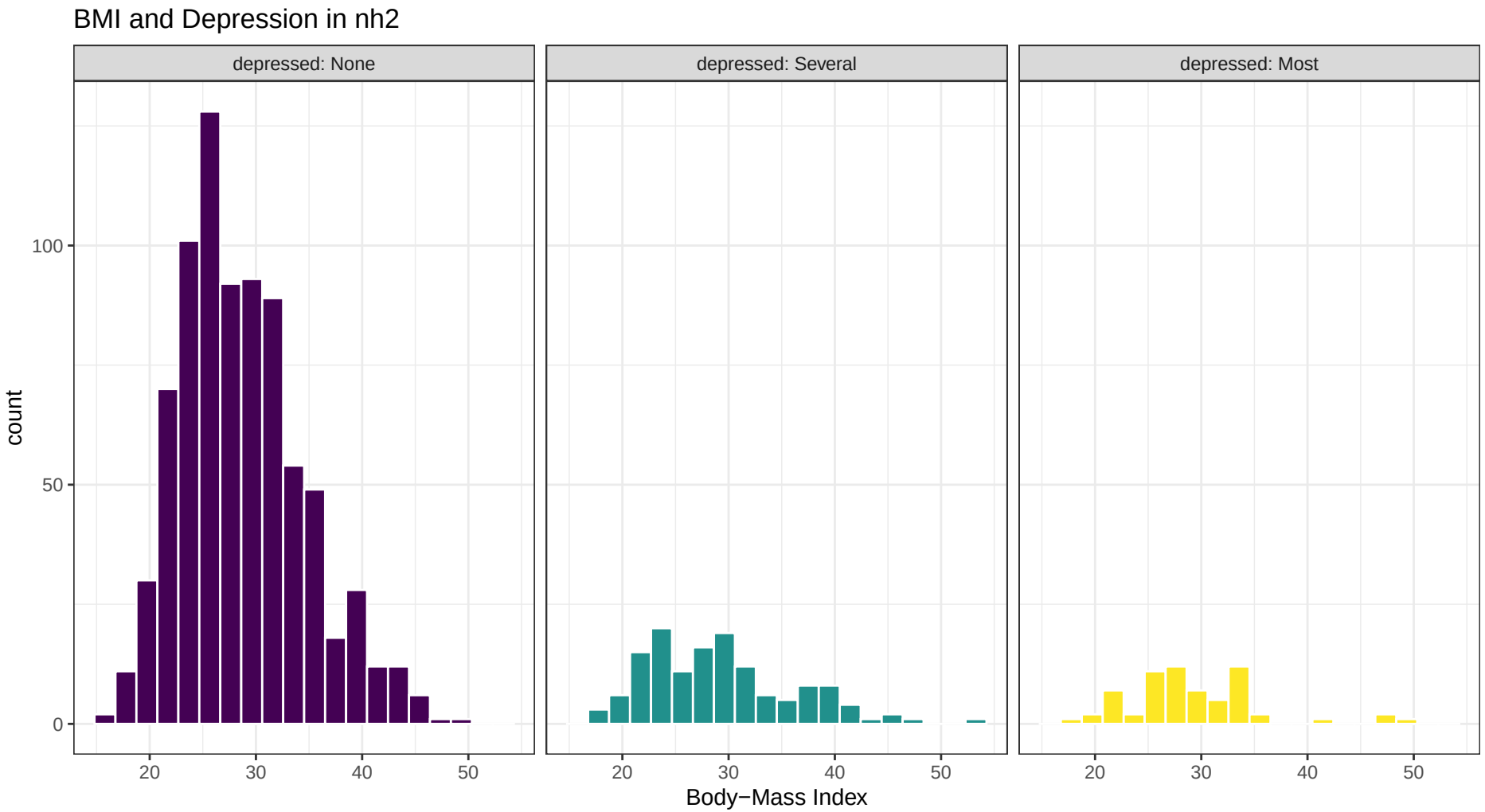
BMI by Depression Status: Violin Plot



BMI by Depression Status, Faceted Histograms

```
ggplot(nh2, aes(x = bmi, fill = depressed)) +
    geom_histogram(bins = 20, col = "white") +
    labs(title = "BMI and Depression in nh2",
         x = "Body-Mass Index") +
    scale_fill_viridis_d() +
    guides(fill = FALSE) +
    theme_bw() +
    facet_wrap(~ depressed, labeller = "label_both")
```

BMI by Depression Status, Faceted Histograms



BMI by Depression Status, Numerically

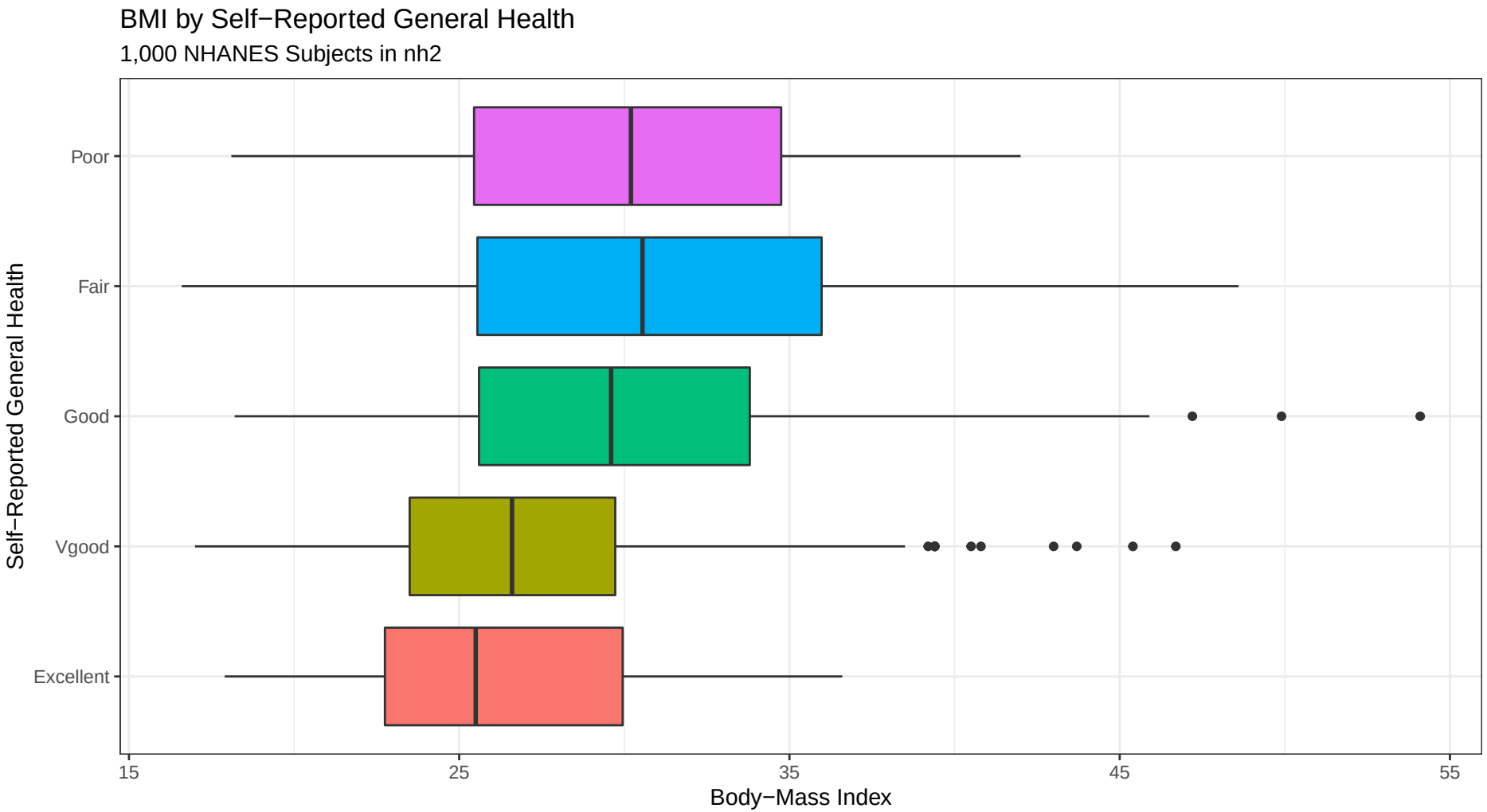
```
nh2 %>%
    group_by(depressed) %>%
    summarize("Count" = n(),
              "Mean(BMI)" = round(mean(bmi),2),
              "Median(BMI)" = median(bmi)) %>%
    knitr::kable()
```

depressed	Count	Mean(BMI)	Median(BMI)
None	797	28.71	27.90
Several	138	29.30	28.25
Most	65	29.18	28.40

BMI by Self-Reported Health Status

```
ggplot(nh2, aes(x = health_gen, y = bmi,
                fill = health_gen)) +
    geom_boxplot() +
    theme_bw() +
    coord_flip() +
    guides(fill = FALSE) +
    labs(title = "BMI by Self-Reported General Health",
         subtitle = "1,000 NHANES Subjects in nh2",
         x = "Self-Reported General Health",
         y = "Body-Mass Index")
```

BMI by Self-Reported Health Status



BMI by Self-Reported Health Status

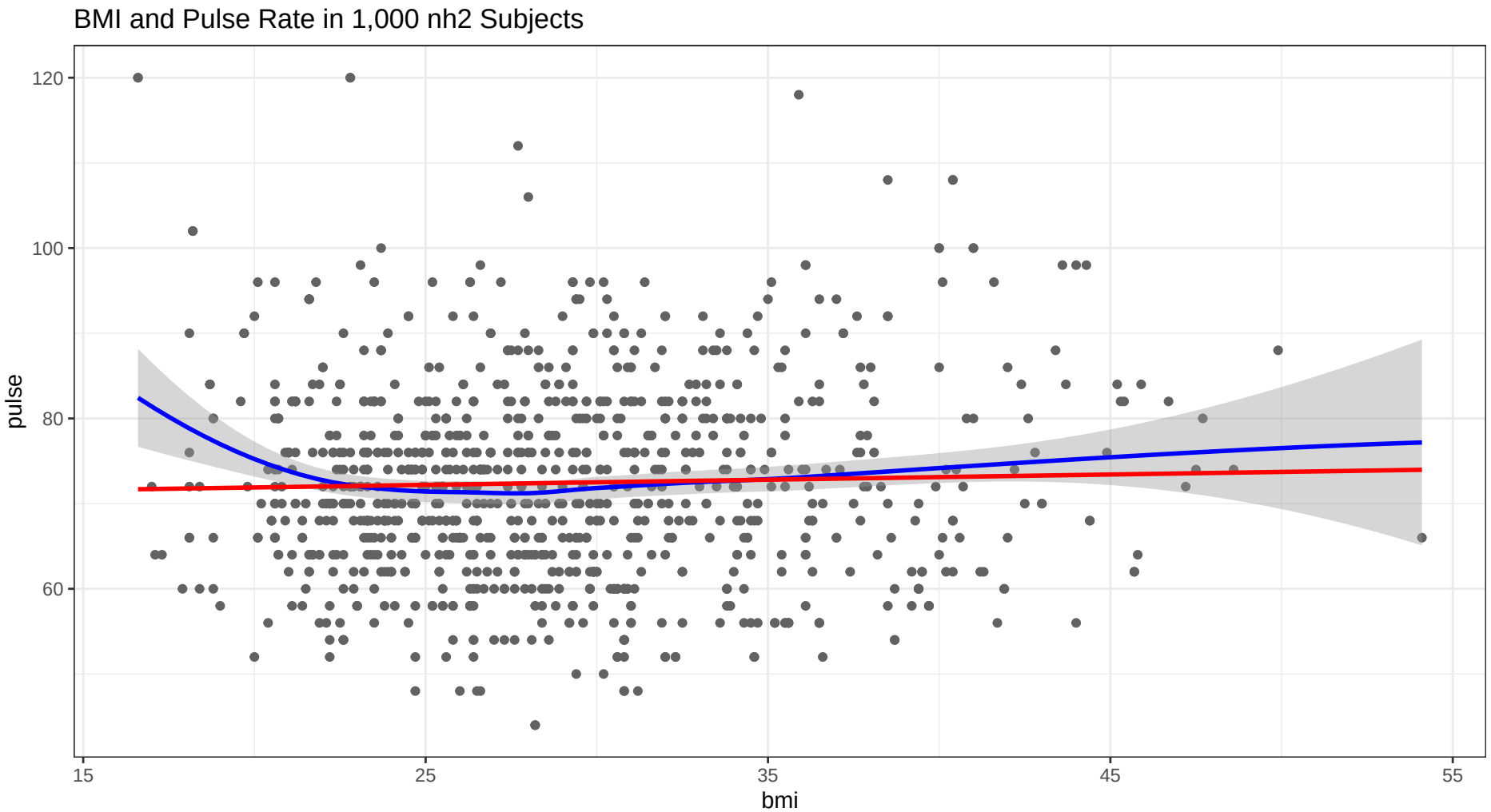
```
nh2 %>%
    group_by(health_gen) %>%
    summarize(count = n(), mean(bmi),
              median(bmi), sd(bmi)) %>%
    knitr::kable(digits = 2)
```

health_gen	count	mean(bmi)	median(bmi)	sd(bmi)
Excellent	128	26.35	25.50	4.53
Vgood	344	27.40	26.60	5.10
Good	365	30.08	29.60	6.18
Fair	140	31.07	30.55	7.17
Poor	23	30.10	30.20	6.56

Association of BMI and Pulse Rate

```
ggplot(nh2, aes(x = bmi, y = pulse)) +
    geom_point(col = cwru.gray) +
    geom_smooth(method = "loess", se = TRUE, col = "blue") +
    geom_smooth(method = "lm", se = FALSE, col = "red") +
    theme_bw() +
    labs(title = "BMI and Pulse Rate in 1,000 nh2 Subjects")
```

Association of BMI and Pulse Rate



Correlation Coefficient to Summarize Association?

The Pearson correlation coefficient is a very limited measure. It only describes the degree to which a **linear** relationship is present in the data. But we can look at it.

```
nh2 %$% cor(bmi, pulse)
```

```
[1] 0.03170466
```

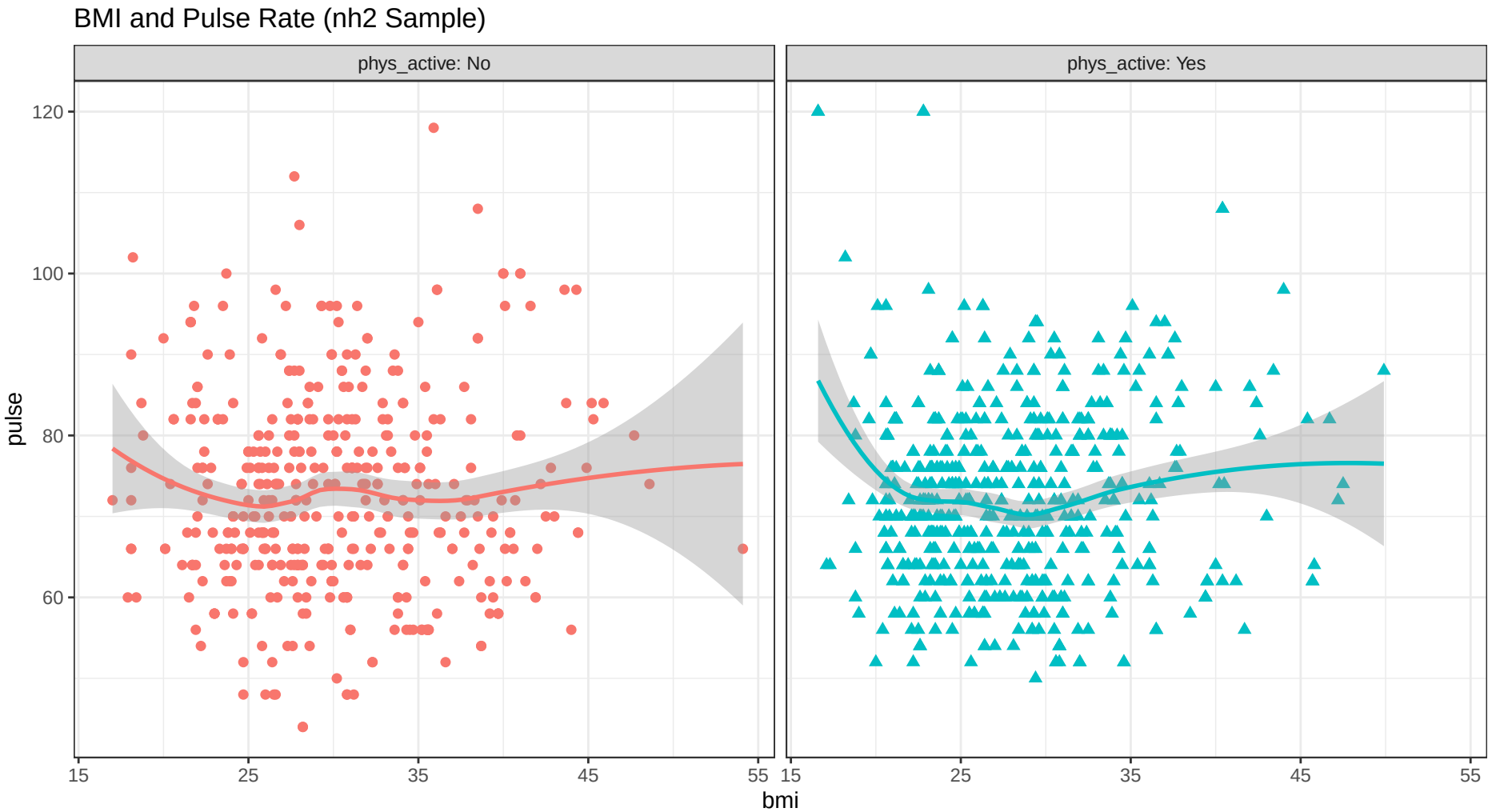
- The Pearson correlation ranges from -1 (perfect negative [as x rises, y falls] linear relationship) to +1 (perfect positive [as x rises, y rises] linear relationship.)
- Our correlation is pretty close to zero. This implies we have a very weak linear association in this case, across the entire sample.

Does Physical Activity affect the Pulse-BMI Association?

Let's change the shape and color of the points based on physical activity status.

```
ggplot(data = nh2, aes(x = bmi, y = pulse,
                       color = phys_active,
                       shape = phys_active)) +
    geom_point(size = 2) +
    geom_smooth(method = "loess", formula = y ~ x) +
    guides(color = FALSE, shape = FALSE) +
    labs(title = "BMI and Pulse Rate (nh2 Sample)") +
    facet_wrap(~ phys_active, labeller = "label_both") +
    theme_bw()
```

Does Physical Activity affect the Pulse-BMI Association?



Correlation(BMI, pulse) by Physical Activity?

- The Pearson correlation coefficient for the relationship between `bmi` and `pulse` in the full sample was quite weak, specifically, it was 0.032.
- Grouped by physical activity status, do we get a different story?

```
nh2 %>%
    group_by(phys_active) %>%
    summarize(cor(bmi, pulse)) %>%
    knitr::kable(digits = 3)
```

phys_active	cor(bmi, pulse)
No	0.022
Yes	0.036

Next Time

Scatterplots, Smoothing and Regression Models