431 Class 07

Thomas E. Love

2018-09-18

Today's Agenda

1. Working with a Categorical Outcome
2. Assessing Normality through Visualization
3. Kidney Cancer Maps

NHANES – As we've seen before...

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

set.seed(20180911) # note same seed as Classes 5 and 6

nh_2 <- sample_n(NHANES, size = 1000) %>%
    select(ID, Gender, Age, Height, Weight, BMI,
           Pulse, Race1, HealthGen, Diabetes)

nh_3 <- nh_2 %>%
    filter(Age > 20 & Age < 80) %>%
    select(ID, Gender, Age, Height, Weight, BMI,
           Pulse, Race1, HealthGen, Diabetes) %>%
    na.omit
```

General Health Status: A Categorical Outcome

Here's a Table of the General Health Status results. This is a self-reported rating of each subject's health on a five point scale (Excellent, Very Good, Good, Fair, Poor.)

```
nh_3 %>%
    select(HealthGen) %>%
    table() %>%
    addmargins()
```

```
.
Excellent     Vgood      Good      Fair      Poor
       69       206       223        76        14
      Sum
      588
```

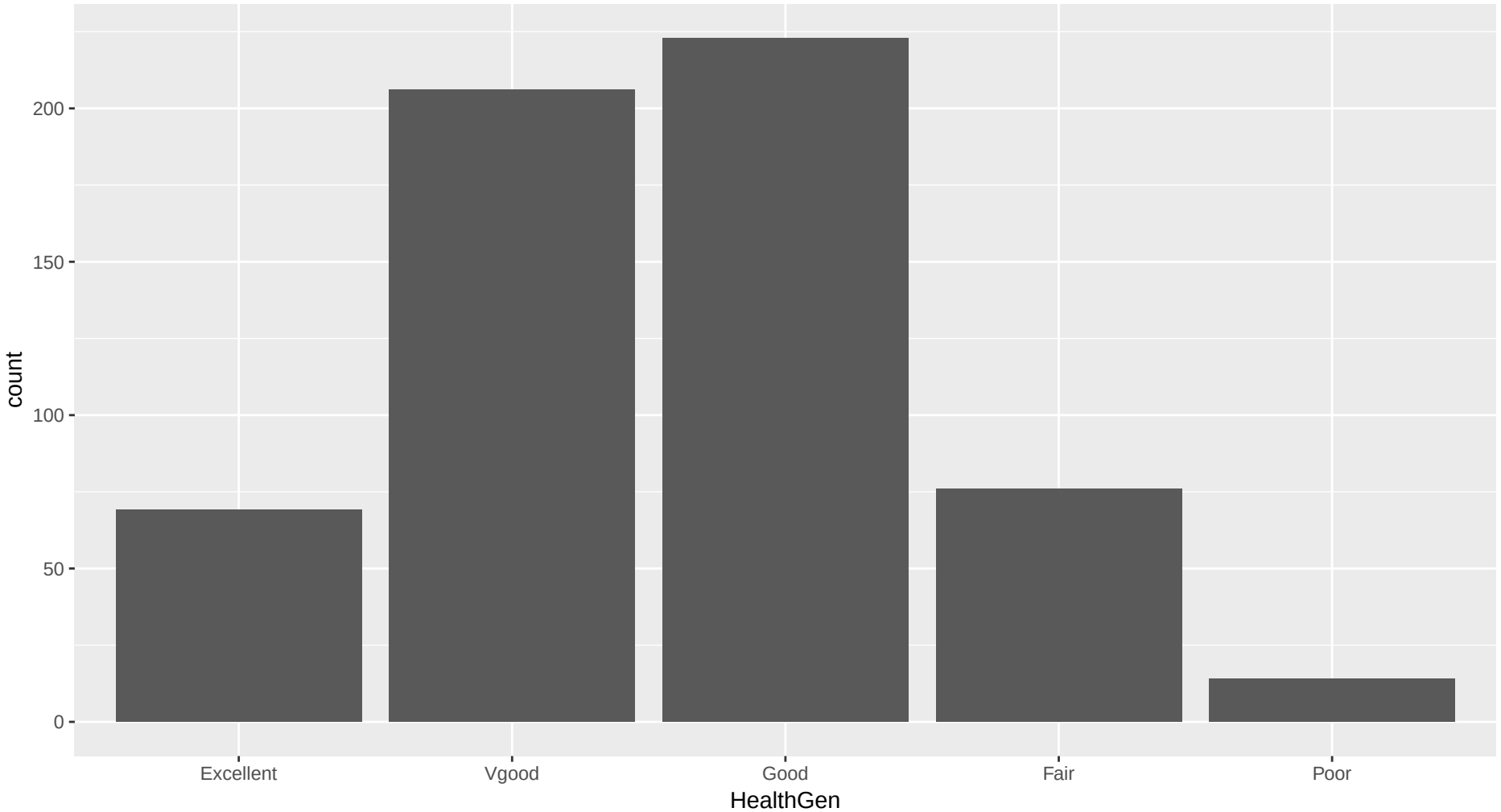
The `HealthGen` data are categorical, which means that summarizing them with averages isn't as appealing as looking at percentages, proportions and rates.

Bar Chart for Categorical Data

Usually, a **bar chart** is the best choice for a graphing a variable made up of categories.

```
ggplot(data = nh_3, aes(x = HealthGen)) +
    geom_bar()
```

Original Bar Chart of General Health

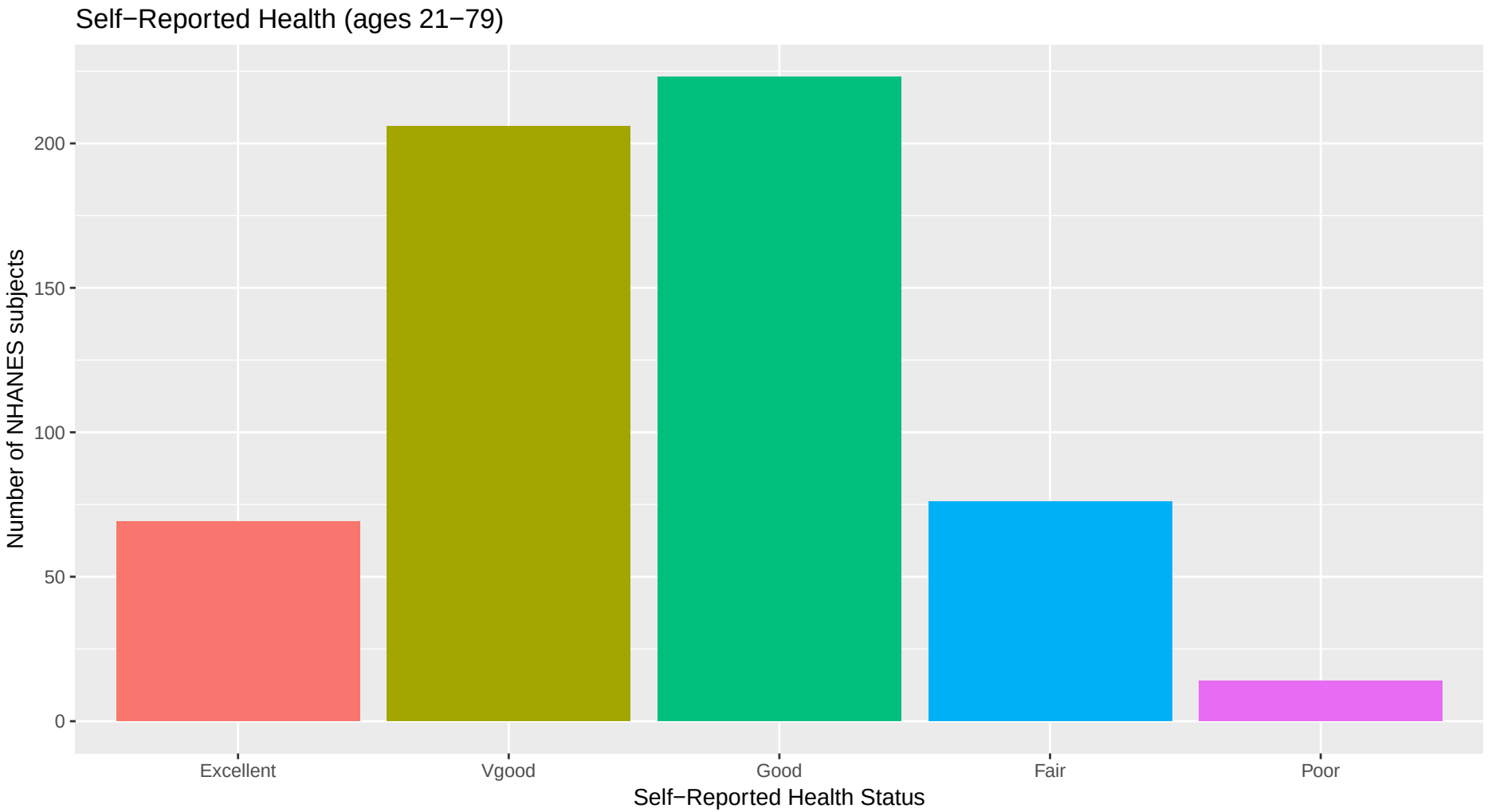


Improving the Bar Chart

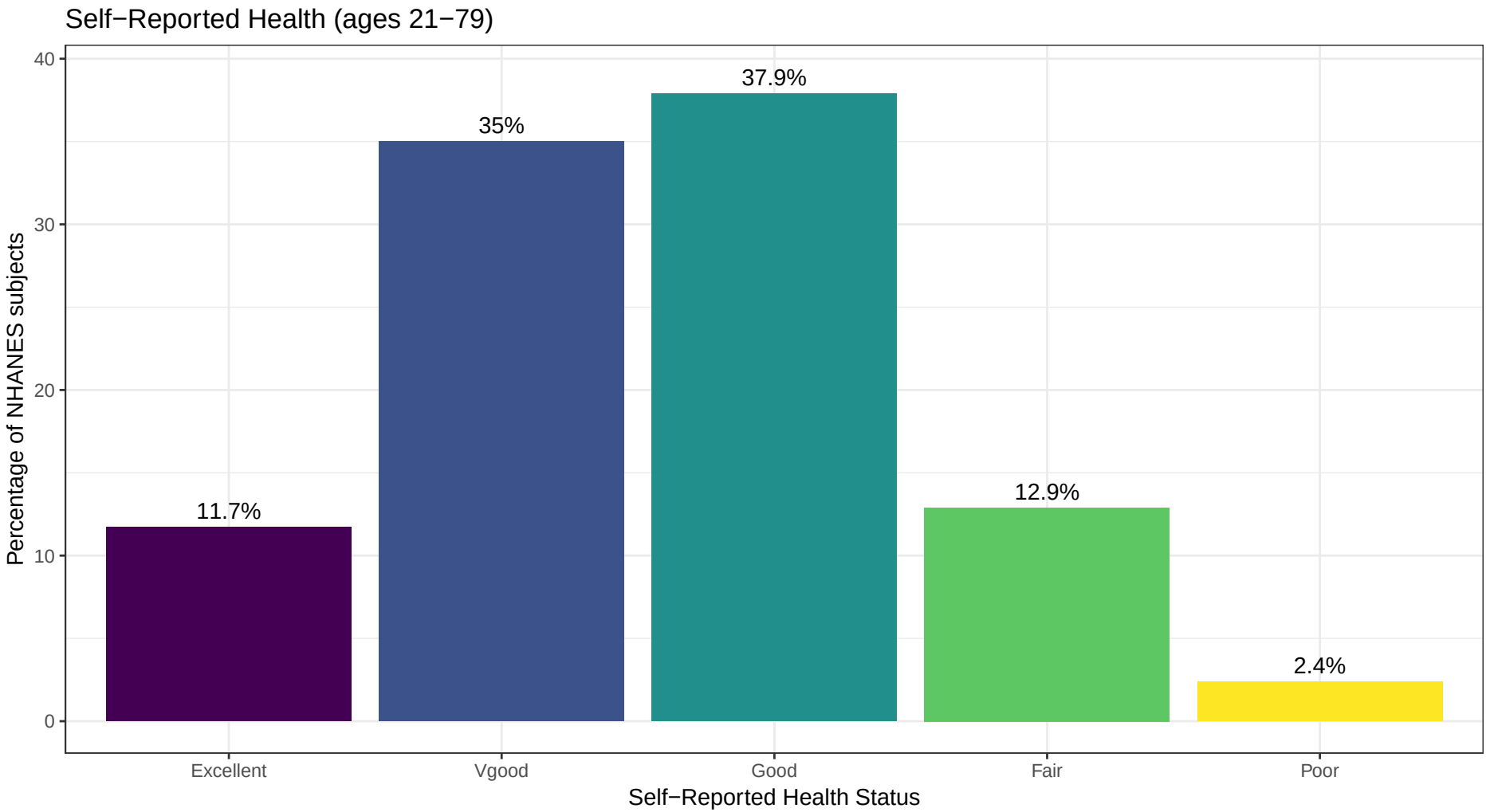
There are lots of things we can do to make this plot fancier.

```
ggplot(data = nh_3,
       aes(x = HealthGen, fill = HealthGen)) +
    geom_bar() +
    guides(fill = FALSE) +
    labs(x = "Self-Reported Health Status",
         y = "Number of NHANES subjects",
         title = "Self-Reported Health (ages 21-79)")
```

The Improved Bar Chart



Or, we can really go crazy. . . (code on next slide)



What crazy looks like. . .

```
nh_3 %>%
    count(HealthGen) %>%
    ungroup() %>%
    mutate(pct = round(prop.table(n) * 100, 1)) %>%
    ggplot(aes(x = HealthGen, y = pct, fill = HealthGen)) +
    geom_bar(stat = "identity", position = "dodge") +
    scale_fill_viridis_d() +
    guides(fill = FALSE, col = FALSE) +
    geom_text(aes(y = pct + 1,     # nudge above top of bar
                  label = paste0(pct, '%')),  # prettify
              position = position_dodge(width = .9),
              size = 4) +
    labs(x = "Self-Reported Health Status",
         y = "Percentage of NHANES subjects",
         title = "Self-Reported Health (ages 21-79)") +
    theme_bw()
```

Working with Cross-Tabulations

We can add a marginal total, and compare subjects by Gender, as follows. . .

```
nh_3 %>%
    select(Gender, HealthGen) %>%
    table() %>%
    addmargins() %>%
    knitr::kable()
```

	Excellent	Vgood	Good	Fair	Poor	Sum
female	39	116	100	33	8	296
male	30	90	123	43	6	292
Sum	69	206	223	76	14	588

Getting Row Proportions

We'll use `prop.table` and get the row proportions by feeding it a 1.

```
nh_3 %>%
    select(Gender, HealthGen) %>%
    table() %>%
    prop.table(.,1) %>%
    round(.,2) %>%
    knitr::kable()
```

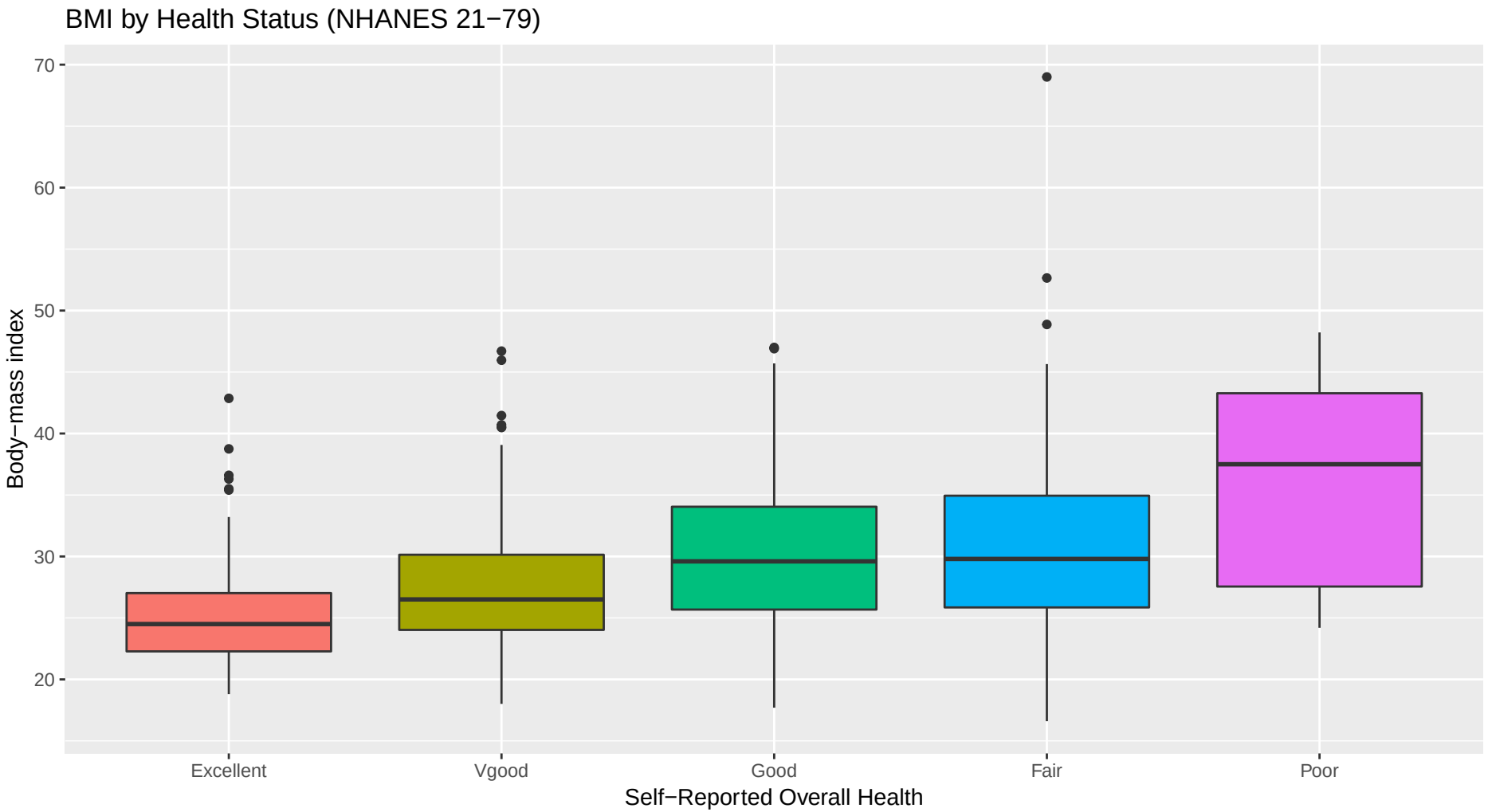
	Excellent	Vgood	Good	Fair	Poor
female	0.13	0.39	0.34	0.11	0.03
male	0.10	0.31	0.42	0.15	0.02

BMI by General Health Status

Let's consider now the relationship between self-reported overall health and body-mass index.

```
ggplot(data = nh_3,
       aes(x = HealthGen, y = BMI, fill = HealthGen)) +
    geom_boxplot() +
    labs(title = "BMI by Health Status (NHANES 21-79)",
         y = "Body-mass index",
         x = "Self-Reported Overall Health") +
    guides(fill = FALSE)
```

What happens with the `Poor` category?



Summary Table of BMI distribution by HealthGen

```
nh_3 %>%
    group_by(HealthGen) %>%
    summarize("BMI n" = n(),
              "Mean" = round(mean(BMI),1),
              "SD" = round(sd(BMI),1),
              "min" = round(min(BMI),1),
              "Q25" = round(quantile(BMI, 0.25),1),
              "median" = round(median(BMI),1),
              "Q75" = round(quantile(BMI, 0.75),1),
              "max" = round(max(BMI),1)) %>%
    knitr::kable()
```

- Resulting table is shown in the next slide.

Not many self-identify in the `Poor` category

HealthGen	BMI n	Mean	SD	min	Q25	median	Q75	max
Excellent	69	25.5	4.9	18.8	22.3	24.5	27.0	42.9
Vgood	206	27.7	5.2	18.0	24.0	26.5	30.1	46.7
Good	223	30.1	5.9	17.7	25.7	29.6	34.0	47.0
Fair	76	31.2	8.7	16.6	25.9	29.8	34.9	69.0
Poor	14	36.7	8.5	24.2	27.6	37.5	43.3	48.2

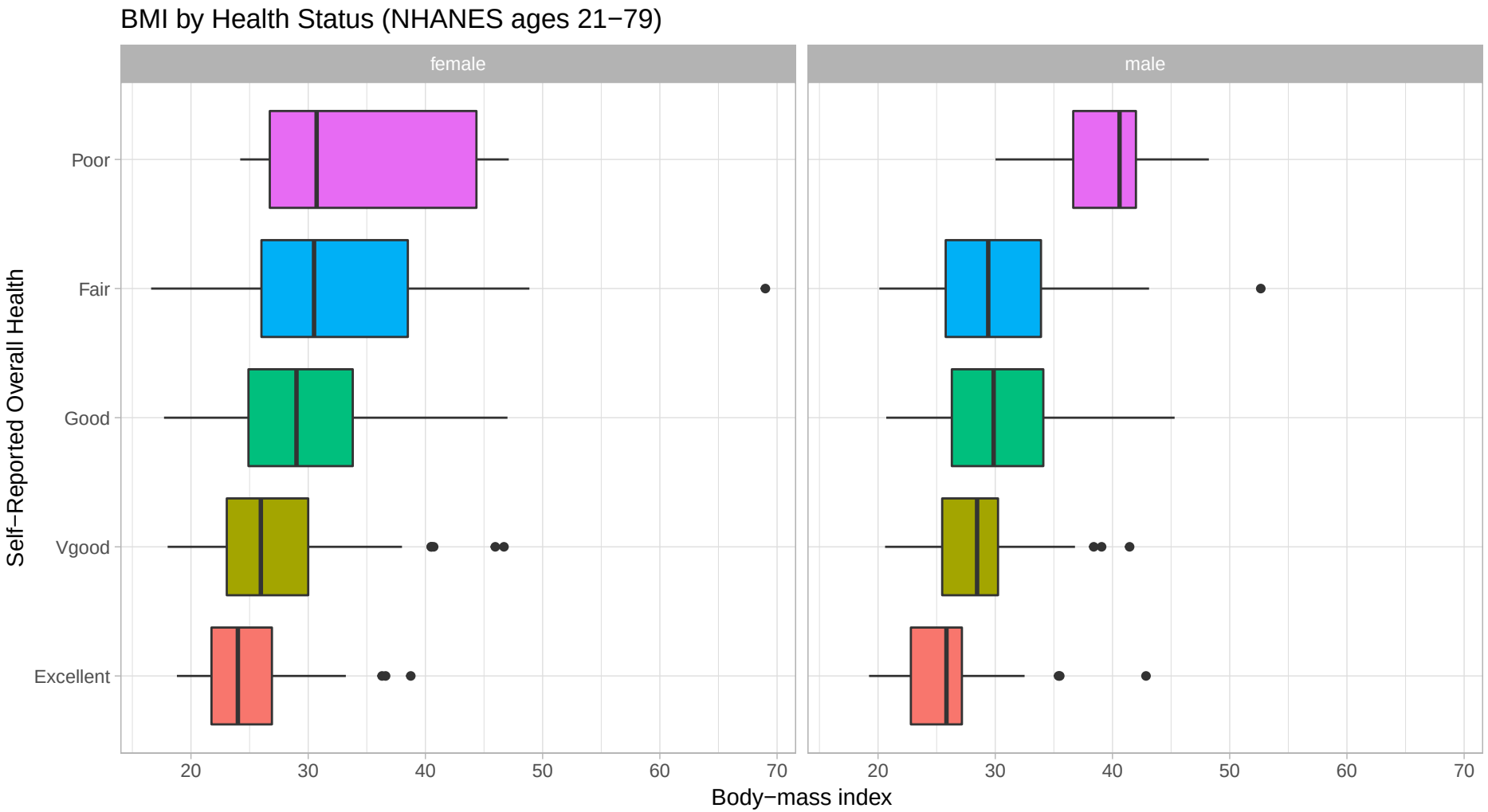
BMI by Gender and General Health Status

We'll start with two panels of boxplots to try to understand the relationships between BMI, General Health Status and Gender.

```
ggplot(data = nh_3,
       aes(x = HealthGen, y = BMI, fill = HealthGen)) +
    geom_boxplot() +
    guides(fill = FALSE) +
    facet_wrap(~ Gender) +
    coord_flip() +
    theme_light() +
    labs(title = "BMI by Health Status (NHANES ages 21-79)",
         y = "Body-mass index",
         x = "Self-Reported Overall Health")
```

- Note the use of `coord_flip` to rotate the graph 90 degrees.
- Note the use of `theme_light()`.

BMI by Gender and General Health Status Boxplots



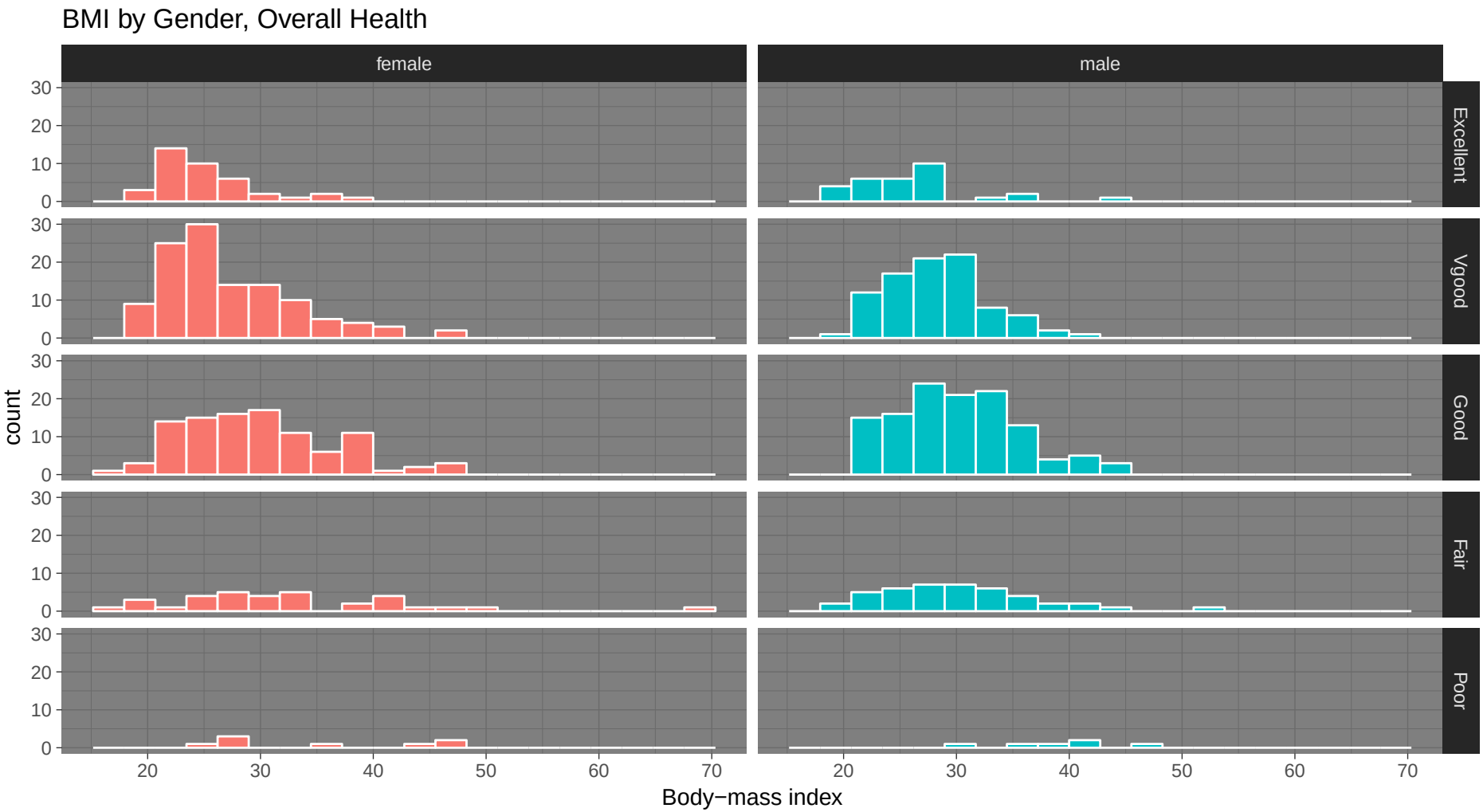
Histograms of BMI by Health and Gender

Here are doubly faceted histograms, which can help address similar questions.

```
ggplot(data = nh_3,
       aes(x = BMI, fill = Gender)) +
    geom_histogram(color = "white", bins = 20) +
    labs(title = "BMI by Gender, Overall Health",
         x = "Body-mass index") +
    guides(fill = FALSE) +
    facet_grid(HealthGen ~ Gender) +
    theme_dark()
```

- Note the use of `facet_grid` to specify rows and columns.
- Note the use of a new theme, called `theme_dark()`.

Histograms of BMI by Health and Gender

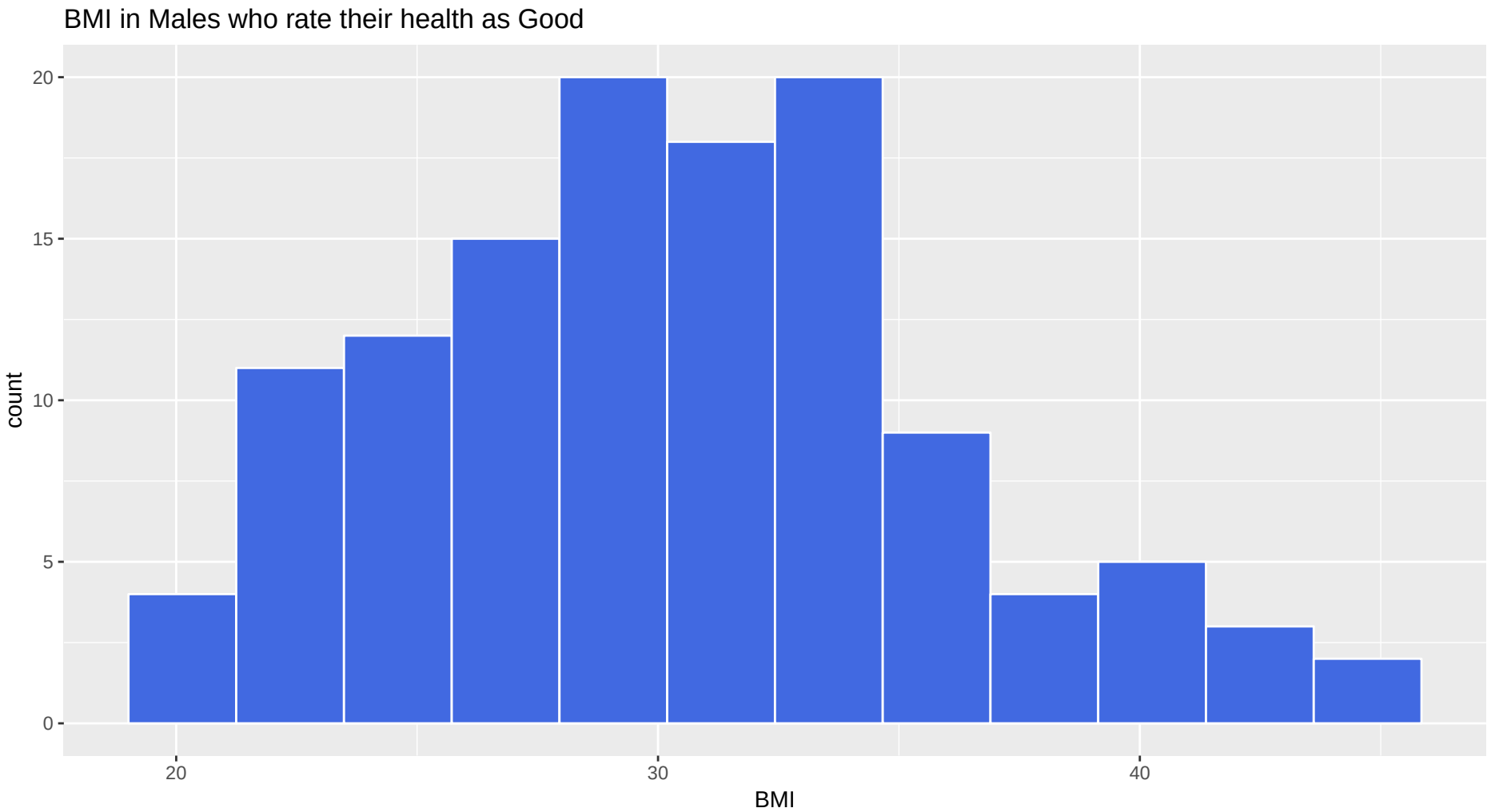


Assessing whether a Normal Model could be useful

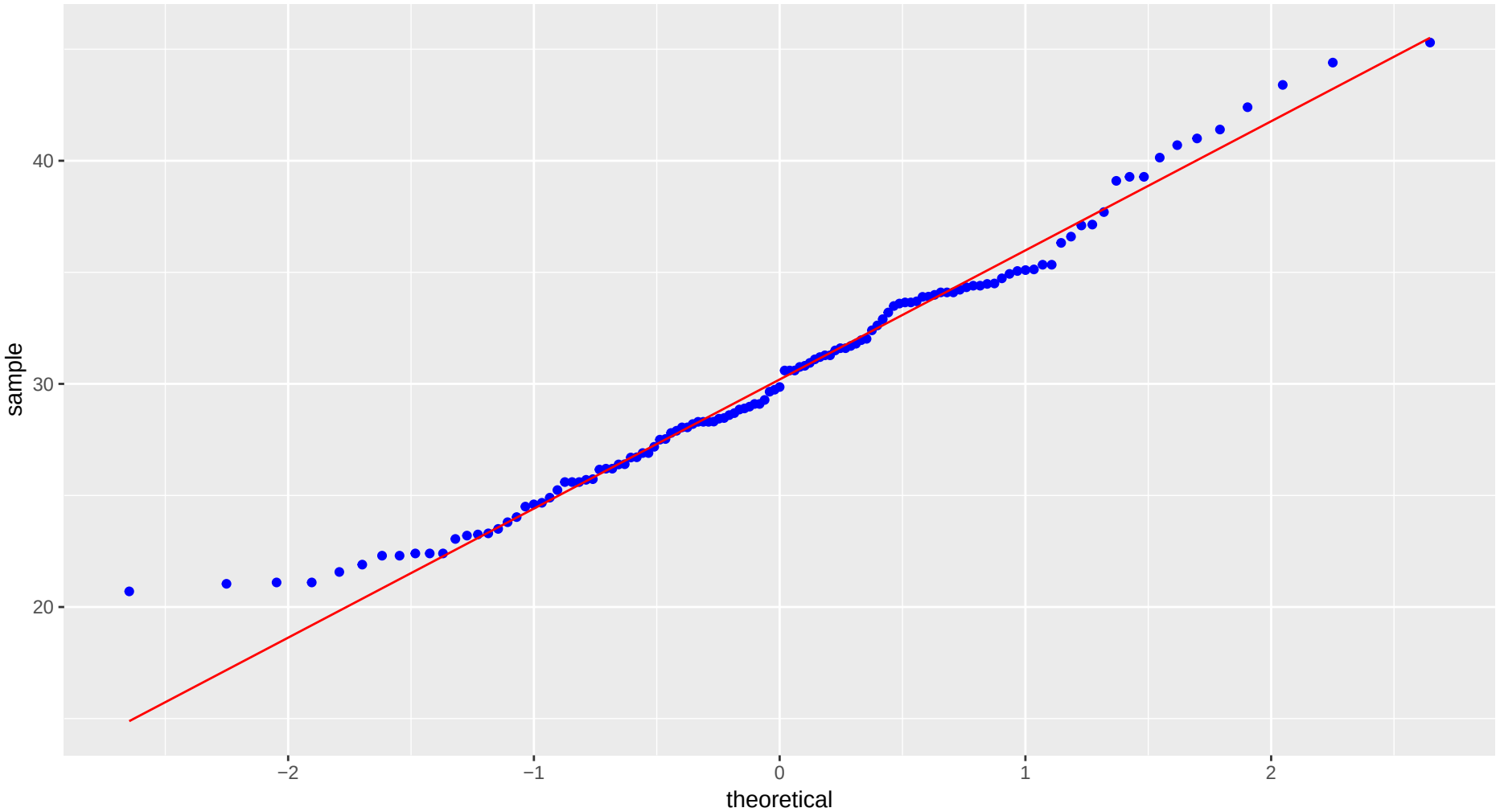
Let's look at the distribution of Males who rate their health as Good.

```
nh_3 %>% filter(Gender == "male" & HealthGen == "Good") %>%
  ggplot(., aes(x = BMI)) +
  geom_histogram(bins = 12,
                 fill = "royalblue", col = "white") +
  labs(title = "BMI in Males who rate their health as Good")
```

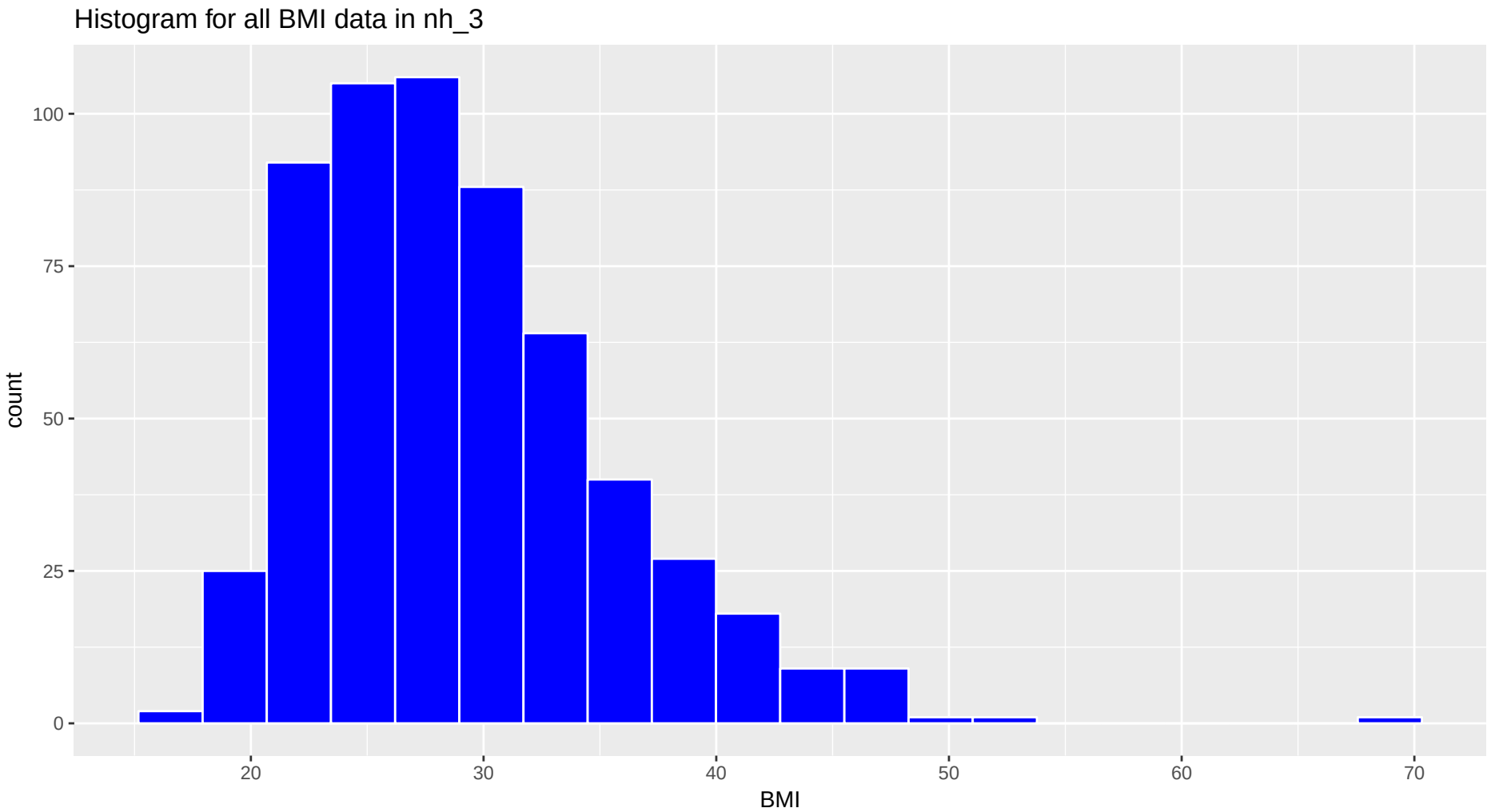
Could a Normal Model be useful here?



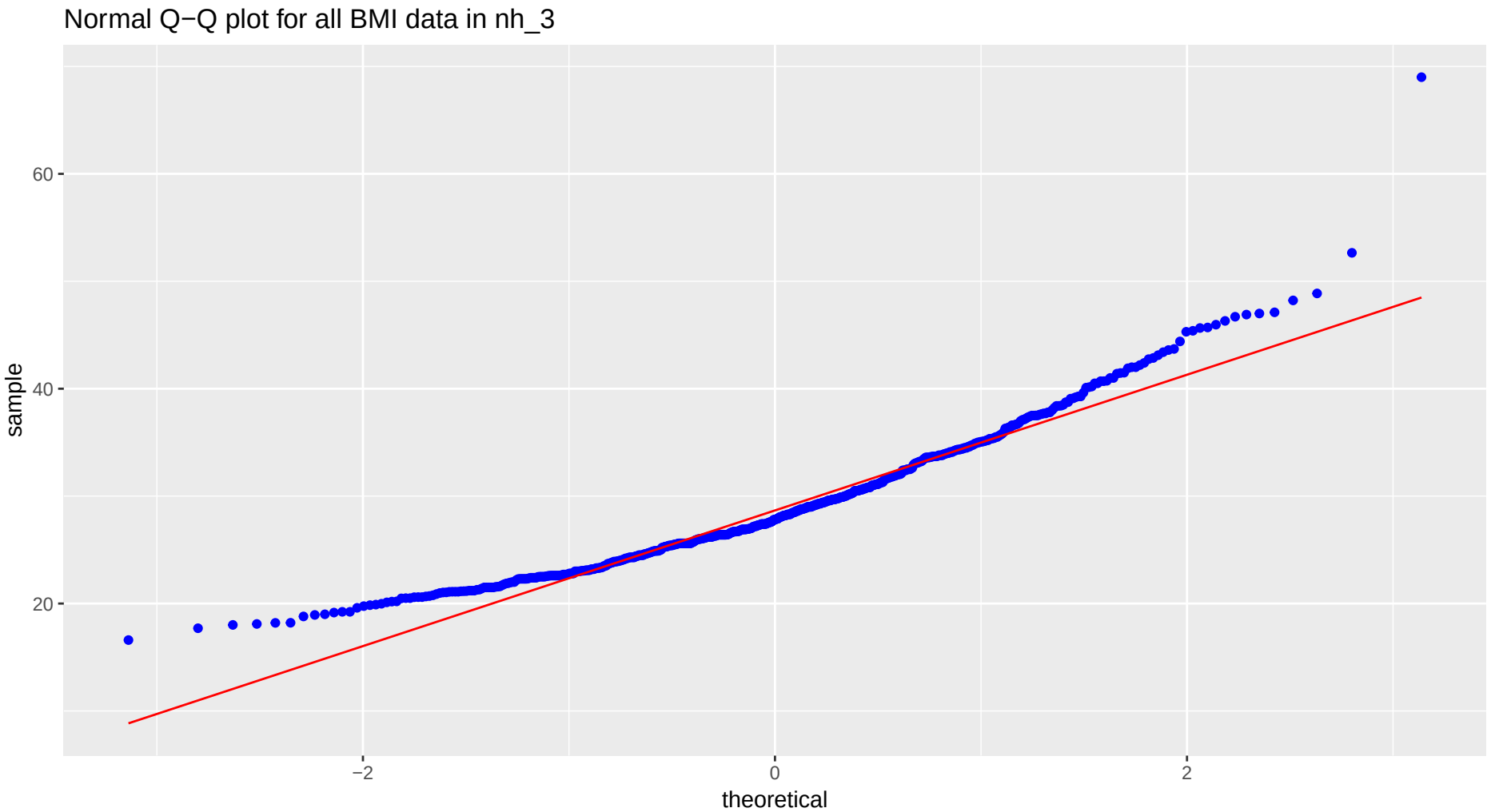
Normal Q-Q plot within `ggplot2`



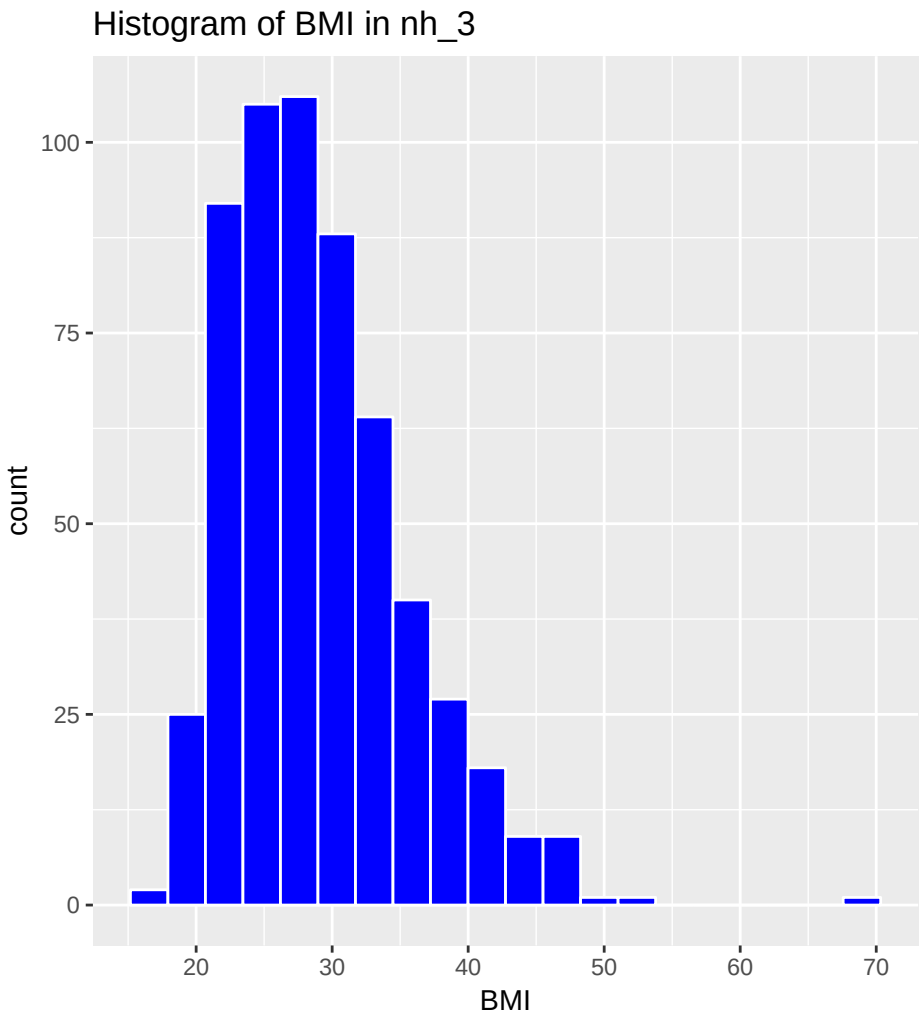
Histogram of all BMI values in `nh_3`

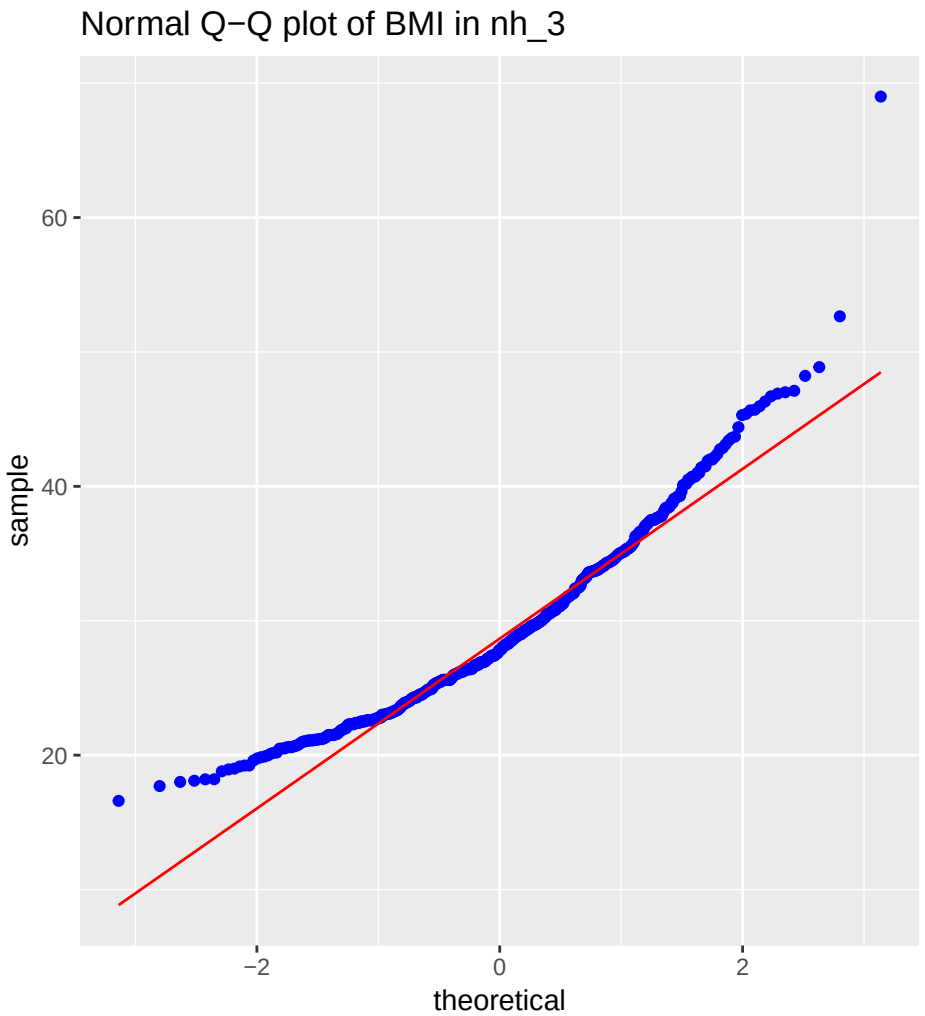


Normal Q-Q plot of all BMI values in `nh_3`



Two plots, side by side





Two plots, side by side (code)

```
plot_a <- ggplot(nh_3, aes(x = BMI)) +
  geom_histogram(bins = 20, fill = "blue", col = "white") +
  labs(title = "Histogram of BMI in nh_3")

plot_b <- ggplot(nh_3, aes(sample = BMI)) +
  geom_qq(col = "blue") + geom_qq_line(col = "red") +
  labs(title = "Normal Q-Q plot of BMI in nh_3")

gridExtra::grid.arrange(plot_a, plot_b, ncol = 2)
```

Superimpose a Normal model on histogram?

A Normal distribution is completely specified by the mean and standard deviation. For our men who rate their health as Good, we have:

```
nh_3 %>% filter(Gender == "male", HealthGen == "Good") %>%
  summarize(count = n(), mean(BMI), sd(BMI))
```

```
# A tibble: 1 x 3
  count `mean(BMI)` `sd(BMI)`
  <int>       <dbl>     <dbl>
1   123        30.4      5.51
```

So, we'd want a Normal model with that mean and standard deviation.

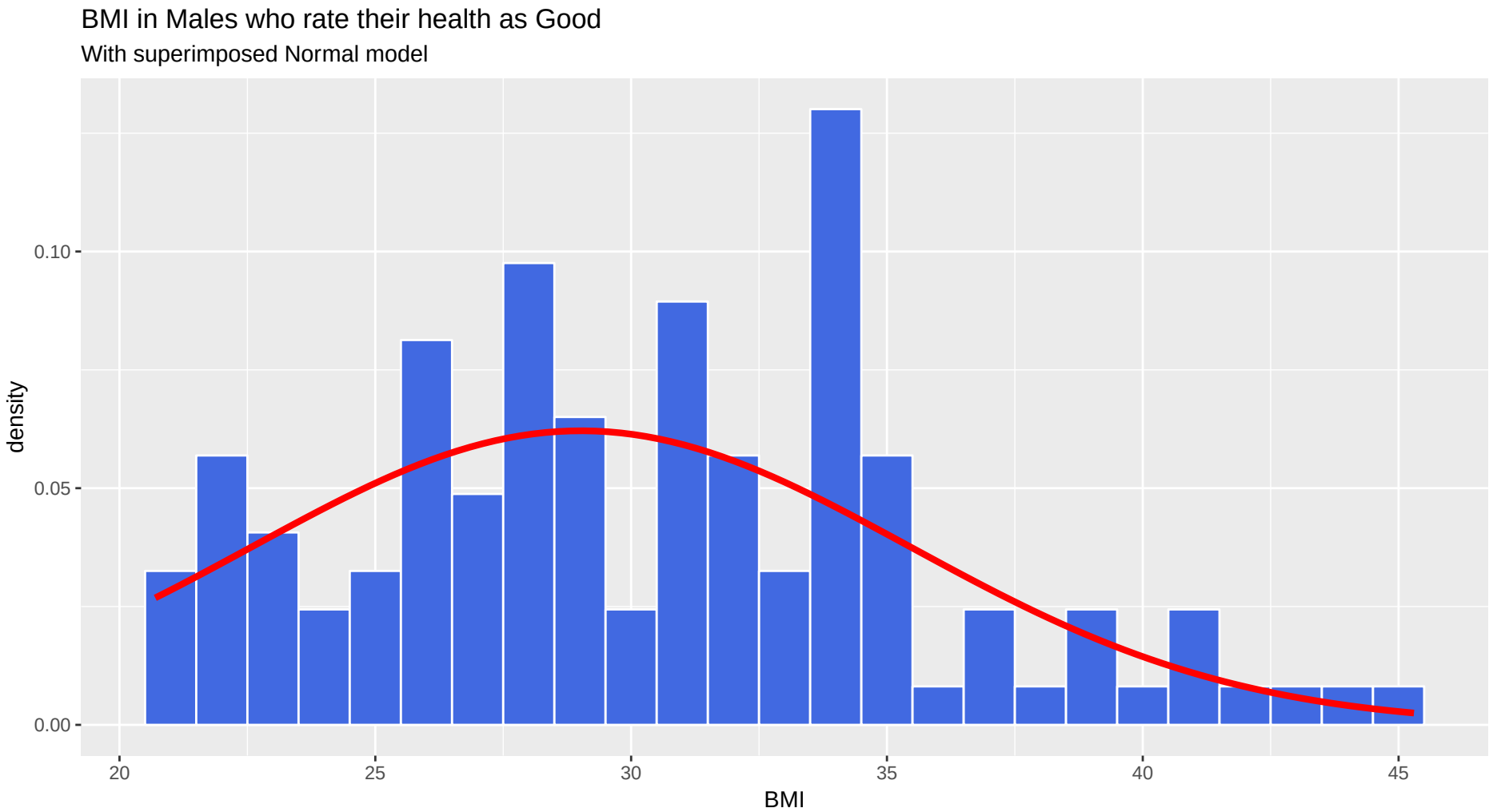
Superimposing a Normal model on a histogram

Code to superimpose a Normal model ("counts")

```
nh_3_maleGood <- nh_3 %>%
  filter(Gender == "male", HealthGen == "Good")

ggplot(nh_3_maleGood, aes(x = BMI)) +
  geom_histogram(binwidth = 1, fill = "royalblue",
                 col = "white") +
  stat_function(fun = function(x, mean, sd, n)
    n * dnorm(x = x, mean = mean, sd = sd),
    args = with(nh_3_maleGood,
                c(mean = mean(BMI),
                  sd = sd(BMI),
                  n = length(BMI))),
    col = "red", lwd = 1.5) +
  labs(title = "BMI in Males who rate their health as Good",
       subtitle = "With superimposed Normal model")
```

Could plot density function, add a Normal curve



Code: Density version of Normal model superimposition

```
nh_3 %>% filter(Gender == "male", HealthGen == "Good") %>%
  ggplot(., aes(x = BMI)) +
  geom_histogram(aes(y = ..density..), binwidth = 1,
                 fill = "royalblue", col = "white") +
  stat_function(fun = dnorm, col = "red", lwd = 1.5,
                args = list(mean = mean(nh_3$BMI, na.rm = T),
                            sd = sd(nh_3$BMI, na.rm = T))) +
  labs(title = "BMI in Males who rate their health as Good",
       subtitle = "With superimposed Normal model")
```

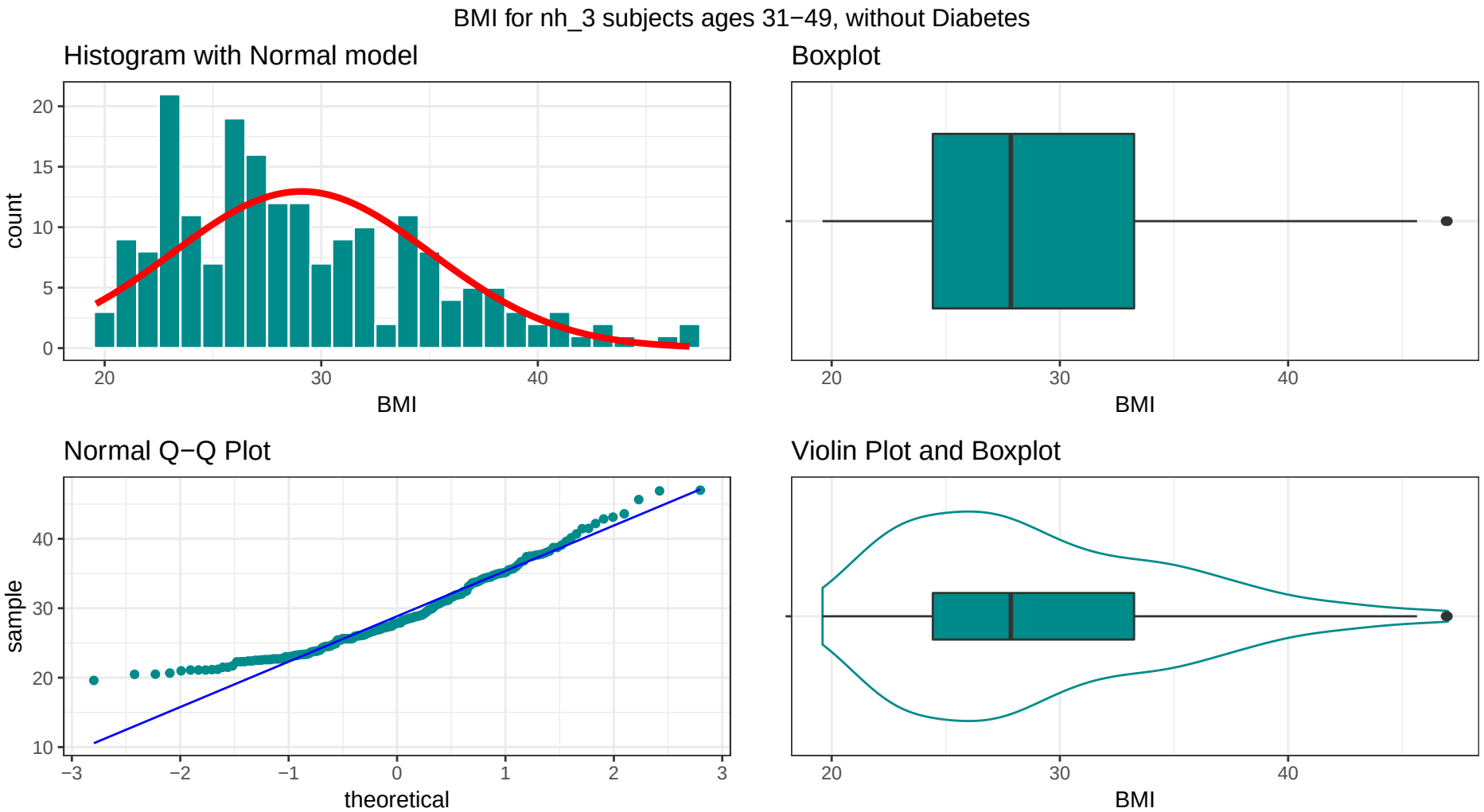
and see Section 8.3 of the Course Notes.

Does a Normal model fit well for my data?

1. Is a Normal Q-Q plot showing something close to a straight line, without clear signs of skew or indications of lots of outliers (heavy-tailedness)?
2. Does a boxplot, violin plot and/or histogram also show a symmetric distribution, where both the number of outliers is modest, and the distance of those outliers from the mean is modest?
3. Do numerical measures match up with the expectations of a normal model?

Let's start by looking at 1 and 2.

BMI among people without diabetes ages 31-49



Does a Normal model fit well for my data?

3. Do numerical measures match up with the expectations of a normal model?

- Is the mean close to the median (perhaps so that the skew_1 is less than 0.2 in absolute value)?
- In a Normal model, mean $\pm$ 1 standard deviation covers 68% of the data.
- In a Normal model, mean $\pm$ 2 standard deviations covers 95% of the data.
- In a Normal model, mean $\pm$ 3 standard deviations covers 99.7% of the data.

Normal model for nh_3 subjects ages 31-49, without Diabetes?

```
nh_3149nodm <- nh_3 %>%
  filter(Age > 30, Age < 50, Diabetes == "No")

mosaic::favstats(~ BMI, data = nh_3149nodm)
```

```
  min       Q1 median    Q3 max     mean       sd   n
 19.6 24.4325  27.85 33.25  47 29.09289 5.971646 194
 missing
       0
```

What is $skew_1$ here?

```
nh_3149nodm %>%
  summarize(skew1 = (mean(BMI) - median(BMI))/sd(BMI))
```

```
# A tibble: 1 x 1
  skew1
  <dbl>
1 0.208
```

How many of the observations are within 1 SD of the mean?

```
nh_3149nodm %>%
  count(BMI > mean(BMI) - sd(BMI),
        BMI < mean(BMI) + sd(BMI))
```

```
# A tibble: 3 x 3
  `BMI > mean(BMI) - sd(~ `BMI < mean(BMI) + sd~     n
  <lgl>                   <lgl>                  <int>
1 FALSE                   TRUE                      33
2 TRUE                    FALSE                     32
3 TRUE                    TRUE                     129
```

So 129 of the 194 (66.5%) observations are within 1 SD of the mean. How does this compare to the expectation under a Normal model?

How about the mean ± 2 standard deviations rule?

Remember the total sample size here is 194.

```
nh_3149nodm %>%
  count(BMI > mean(BMI) - 2*sd(BMI),
        BMI < mean(BMI) + 2*sd(BMI))
```

```
# A tibble: 2 x 3
  `BMI > mean(BMI) - 2 *~ `BMI < mean(BMI) + 2 ~     n
  <lgl>                   <lgl>                  <int>
1 TRUE                    FALSE                      9
2 TRUE                    TRUE                     185
```

So 185 of the 194 (95.4%) observations are within 2 SD of the mean. How does this compare to the expectation under a Normal model?

Conclusions

This is just a small piece of the toolbox for visualizations that we'll create in this class. Many additional tools are on the way, but the main idea won't change. Using the `ggplot2` package, we can accomplish several critical tasks in creating a visualization, including:

- Identifying (and labeling) the axes and titles
- Identifying a type of `geom` to use, like a point, bar or histogram
- Changing fill, color, shape, size to facilitate comparisons
- Building "small multiples" of plots with faceting

Good data visualizations make it easy to see the data, and `ggplot2`'s tools make it relatively difficult to make a really bad graph.

Group Task: Kidney Cancer Death Rates

Your map shows U.S. counties.

- The shaded counties are in the top 10% of age-standardized rates for death due to cancer of the kidney/ureter for white males, in 1980-1989.

Your Tasks

1. Describe the patterns you see in the map.
2. Speculate as to the cause of these patterns.

Highest kidney cancer death rates

Lowest kidney cancer death rates

Reminders

- Homework 3 is due Friday at Noon.
- Minute Paper after Class 7.
- So far, we've covered most of the material in Chapters 1-8 and 10 of our Course Notes.
  - Part A covers Chapters 1-14.
  - Next Time: More on Studying Association with Scatterplots and Correlations (Chapters 11-12)
  - Coming Soon: Using Transformations to "Normalize" data (Chapter 9)

Notes on the Kidney Cancer example, 1

I first asked you what you noticed about the map, in the hope that someone would point out the obvious pattern, which is that many of the countries in the Great Plains but relatively few near the coasts are shaded.

- Why might that be? Could these be the counties with more old people? Ah, but these rates are age-adjusted.
- They're mostly in rural areas: could the health care there be worse than in major cities? Or perhaps people living in rural areas have less healthy diets, or are exposed to more harmful chemicals? Maybe, but the confusing fact is that the highest 10% and the lowest 10% each show disproportionately higher rates in those Great Plains counties.

Notes on the Kidney Cancer example, 2

- Consider a county with 100 white males. If it has even one kidney death in the 1980s, its rate is 1 per thousand per year, which is among the highest in the nation. If it has no such deaths, its rate will be 0, which is the lowest in the nation.
- The observed rates for smaller counties are *much* more variable, and hence they are more likely to be shaded, even if nothing special is truly going on.
- If a small county has an observed rate of 1 per thousand per year, it's probably random fluctuation. But if a large county (like Cuyahoga) has a very high rate, it is probably a real phenomenon.

Source

My source for this example was Andrew Gelman and Deborah Nolan's book *Teaching Statistics: a bag of tricks* which is the source of a number of things we'll see in the course, including some of the "age guessing" example we've previously done.