Chapter 2 reading notes

5 september 2023

Vocabulary

- scatterplot : compares two numeric variables to each other.
- linear/nonlinear
- dotplot -> like scatterplot, but only one variable, in a straight line.
- Variance: Square the deviations from the mean then take the average
- Standard Deviation, the square root of variance.
- 70 / 95% within 2 standard deviations
- boxplot
- boxplot IQR - between q1 and q3, 50%.
- boxpolot whiskers -> never past 1.5X IQR
- intensity map (AKA heat map)
- contingency plots, summarize two variables, with totals on the right column and bottom row.
- Bar Charts - Stacked, side by side, proportion
- Mosaic charts, almost liek a box plot, but uses a more fixed space. Can have horiz or veritical columns.
- H0: Independence Model -> No effect on rate
- HA: Alternative Method. The vaccine has an effect.
- Reject H0 and accept the alnternative hypothesis.

Formulas

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Mean	Variance	Standard Deviation

Questions:

On page 58,

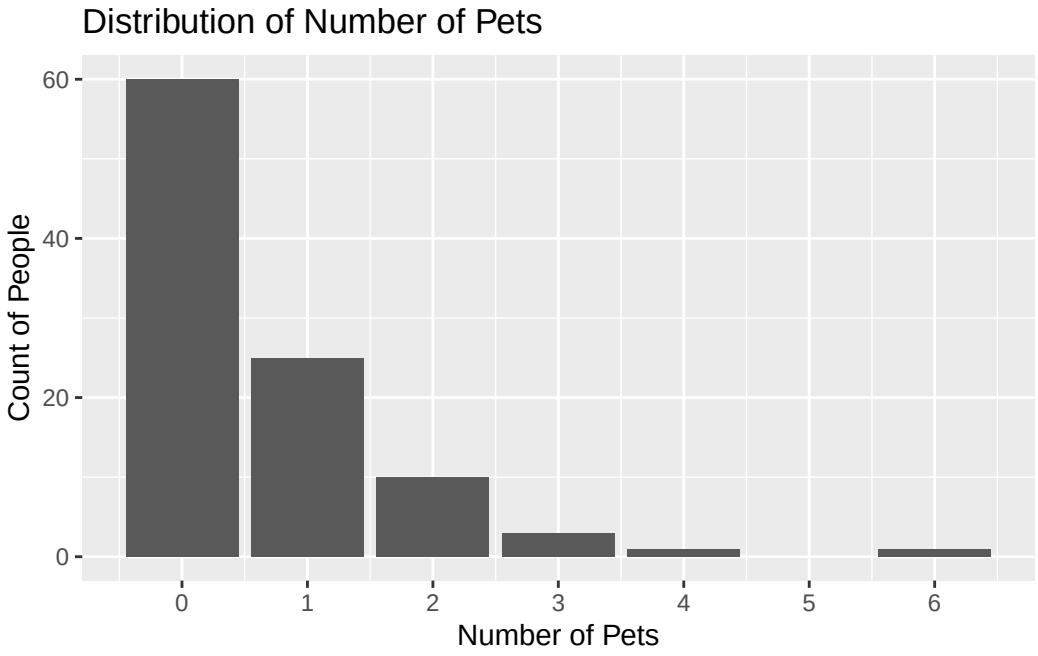
Number of pets per household (right left or symmetric)

```
library(ggplot2)

# Sample data
data <- data.frame(
  pets = c(rep(0, 60),  # 60 people have 0 pets
           rep(1, 25),  # 25 people have 1 pet
           rep(2, 10),  # and so on...
           rep(3, 3),
           rep(4, 1),
           rep(5, 0),
           rep(6, 1))
)

# Plot
ggplot(data, aes(x=pets)) +
  geom_bar(aes(y=..count..)) +
  labs(title="Distribution of Number of Pets",
       x="Number of Pets",
       y="Count of People") +
  scale_x_continuous(breaks = 0:6)  #
```

```
Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(count)` instead.
```



From ChatGPT:

tidiverse

The tidyverse is a collection of R packages designed to simplify data science tasks using a consistent and coherent philosophy. The term "tidyverse" was coined by Hadley Wickham, and it's more than just a collection of packages—it represents a philosophy of data handling, analysis, and visualization in R.

Here's a breakdown:

Philosophy: Central to the tidyverse is the concept of "tidy data," a standardized way of storing data sets. When data is tidy:

- Each variable forms a column.
- Each observation forms a row.
- Each cell contains a single value.
- This structure makes it easier to manipulate, visualize, and model data.

Core Packages: While the tidyverse encompasses many packages, several are central to its ecosystem:

- ggplot2 for data visualization.
- dplyr for data manipulation.

- tidyr for converting data to a tidy format.
- readr for importing datasets into R.
- purrr for functional programming.
- tibble for tamed data frames.
- stringr for string manipulation.
- forcats for factor (categorical variable) manipulation.

Design Consistency: The packages within the tidyverse share a common design philosophy. They have consistent syntax, which makes them easy to learn and use together seamlessly.

Interactivity: Packages like shiny allow users to create interactive web apps, while ggplot2 extensions (like plotly) can make interactive plots.

Extension Packages: Beyond the core set, many other packages embrace tidyverse principles and integrate nicely with the core tools, even if they aren't officially part of the tidyverse.

Active Development & Community: The tidyverse has an active community of developers, educators, and users. This means you can often find solutions to problems, educational resources, and updates improving the suite's capabilities.

In essence, the tidyverse is both a collection of packages and a philosophy about how to analyze and visualize data. It's an approach to R programming designed to make data science tasks more intuitive and productive.