Chapter 2. Summarizing data

2.1 Numerical data

- Scatterplots for paired data
- Dot plot: one-variable scatterplot
- Sample mean: $\bar{x}$ common way to measure the center of a distribution of data
- Population mean: μ Greek letter mu, average of all observations in the population
- Histograms for larger samples, show data density
- Long tails for skew: long left tail means left skewed
- Unimodal, bimodal, multimodal
- Deviation, variance, standard deviation
- Box plots, quartiles, median

Robust statistics

- Median and IQR: extreme observations have little effect on their values
- Transforming data: rescale data using a function
    - Goals: see data structure differently, reduce skew, assist in modeling, straighten a nonlinear relationship in a scatterplot
    - Types: Log, square root, inverse, etc.
- Map data: e.g., intensity map for geographical data (US, worldwide, etc.)

2.2 Categorical data

- Contingency tables, bar plots
- Stacked bar plot, side-by-side bar plot, standardized stacked bar plot
- Mosaic plot (1 or 2 variables versions)
- Pie chart: when differences are small between groups, easier to see them in bar plots
- Numerical comparison across groups
    - Side-by-side boxplot
    - Hollow histograms

2.3 Malaria vaccine

		outcome		
		infection	no infection	Total
treatment	vaccine	5	9	14
	placebo	6	0	6
	Total	11	9	20

Figure 2.29: Summary results for the malaria vaccine experiment.

- Random noise: observed outcome in the data sample may not reflect true relationships between variables because of random noise (differences observed due to chance alone)

- When sample size is small, it is unclear if the observed difference represents efficacy of the vaccine or whether it is simply due to chance.

H_0 Independence model:

- Variables *treatment* and *outcome* are independent; have no relationship; the observed difference was due to chance.

H_A Alternative model:

- Variables *treatment* and *outcome* are not independent; vaccine affected the rate of infection.

Checking for independence:

- Choose between these 2 competing models by assessing if the data conflict so much with H_0 that the notion of independence is rejected and conclude that H_A is supported and the vaccine was effective.

Simulations:

- Complete another randomization where we will pretend we know that vaccine does not work

		outcome		
		infection	no infection	Total
treatment (simulated)	vaccine	7	7	14
	placebo	4	2	6
	Total	11	9	20

Figure 2.30: Simulation results, where any difference in infection rates is purely due to chance.

- Repeat the simulation 100 times and we start to get a good idea of what represents the **distribution of differences from chance alone** (see below for infection rate differences from 100 simulations produced under the independence model, H_0)

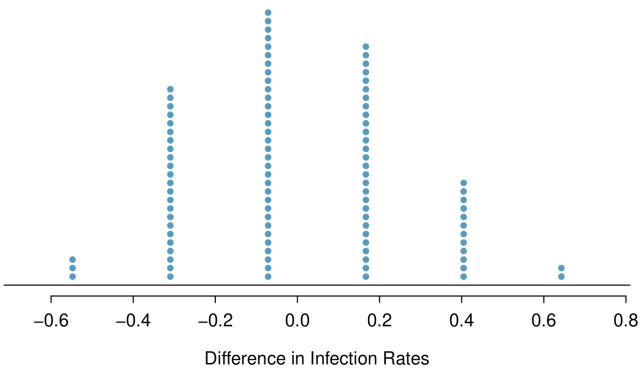

Figure 2.31: A stacked dot plot of differences from 100 simulations produced under the independence model, H_0, where in these simulations infections are unaffected by the vaccine. Two of the 100 simulations had a difference of at least 64.3%, the difference observed in the study.

- Distribution of these simulated differences is centered around 0 because it is assuming that the independence model was true
- How often do you observe a difference of at least 64.3% (0.643; what was observed in data) according to this distribution of simulated differences? Only about 2% of the time!
- 2 possible interpretations:
    - Independence model is still correct; we just happened to observe this difference on a very rare chance. -> conclude we *do not have sufficiently strong evidence* to conclude vaccine had an effect in this clinical setting.
    - Alternative model: the vaccine has an effect on the infection rate, and the difference observed was actually due to vaccine being effective at combatting malaria, which explains the large 64.3% difference. -> conclude *evidence is sufficiently strong to reject* H_0 and assert that vaccine was useful.
- Formal study: we would usually reject the independence model in this case.

Statistical inference:

- Built on evaluating whether differences are due to chance; which model is most reasonable given the data
- We may sometimes choose the wrong model but statistical inference gives us tools to control and evaluate how often these errors occur