Summarizing Data

Ryan Miller

Why summarize?

A restaurant server seeking better understand their tips collects data on each table they serve. From this data, 20 rows are displayed below. What do these data tell you?

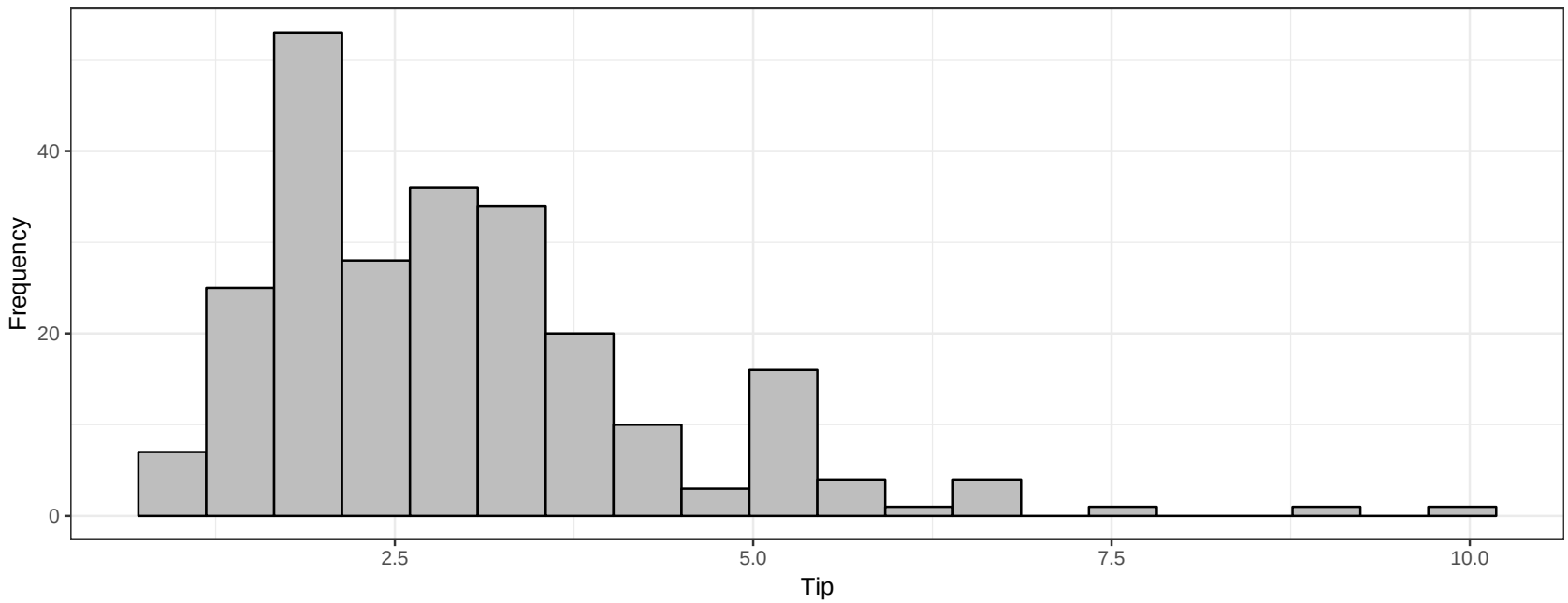
total_bill	tip	sex	smoker	day	time	size
12.69	2.00	Male	No	Sat	Dinner	2
13.13	2.00	Male	No	Sun	Dinner	2
11.87	1.63	Female	No	Thur	Lunch	2
14.07	2.50	Male	No	Sun	Dinner	2
26.59	3.41	Male	Yes	Sat	Dinner	3
24.55	2.00	Male	No	Sun	Dinner	4
21.01	3.50	Male	No	Sun	Dinner	3
19.49	3.51	Male	No	Sun	Dinner	2
25.00	3.75	Female	No	Sun	Dinner	4
11.69	2.31	Male	No	Thur	Lunch	2
16.21	2.00	Female	No	Sun	Dinner	3
8.52	1.48	Male	No	Thur	Lunch	2
20.08	3.15	Male	No	Sat	Dinner	3
13.27	2.50	Female	Yes	Sat	Dinner	2
3.07	1.00	Female	Yes	Sat	Dinner	1
19.81	4.19	Female	Yes	Thur	Lunch	2
15.69	3.00	Male	Yes	Sat	Dinner	3
20.29	3.21	Male	Yes	Sat	Dinner	2
13.94	3.06	Male	No	Sun	Dinner	2
34.81	5.20	Female	No	Sun	Dinner	4

Why summarize?

- Presenting data without any *summarization* is rarely useful
  - Human's simply aren't good at processing that much information
- Summarization reduces the data to a single number (or a small set of numbers)
  - For now, we will focus on **univariate** summaries (those involving a single variable) and **bivariate** summaries (those involving two variables)

Distributions

- For a single variable, we often want to describe how the variable is *distributed*
  - A variable's **distribution** describes values that are possible and how frequently they occur
- Below is a **histogram**, one way of showing a distribution (for a quantitative variable)
  - \$2-3 tips are most common, but larger tips of \$5+ do occasionally occur



Distributions

- Distributions aren't really a summary, but they help us understand summarization
  - The "most common" tips could be more precisely defined by using the **mean** or **median**
  - The less common larger tips could be more precisely defined by using the **maximum** or **90% percentile**
- Each of the four bolded terms is a different *univariate* summary measure
  - Lab #1 will go into further detail on these summary measures

Variability

- Distributions also display **variation** in the data, a fundamental concept in statistics
  - Variation is commonly measured by the **standard deviation**, which roughly corresponds to the average distance of each data-point from the mean
  - Lab #1 will provide a more precise, mathematical definition of standard deviation

Variability - The 68-95-99 Rule

For symmetric, bell-shaped distributions, the standard deviation is related to the percentage of cases within a certain distance of the mean

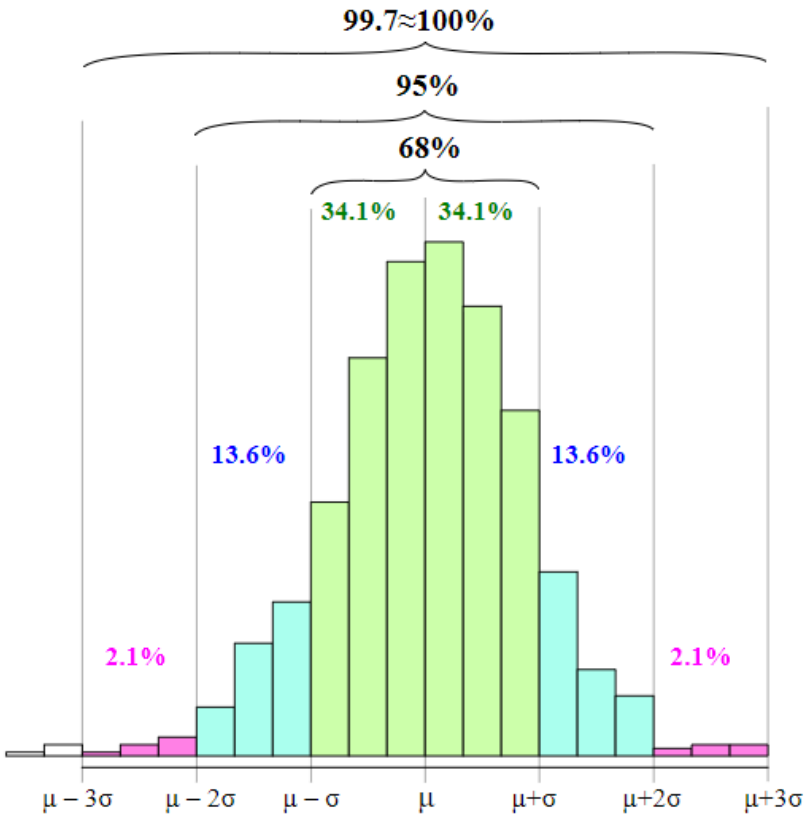


Image Source: https://en.wikipedia.org/wiki/68-95-99.7_rule

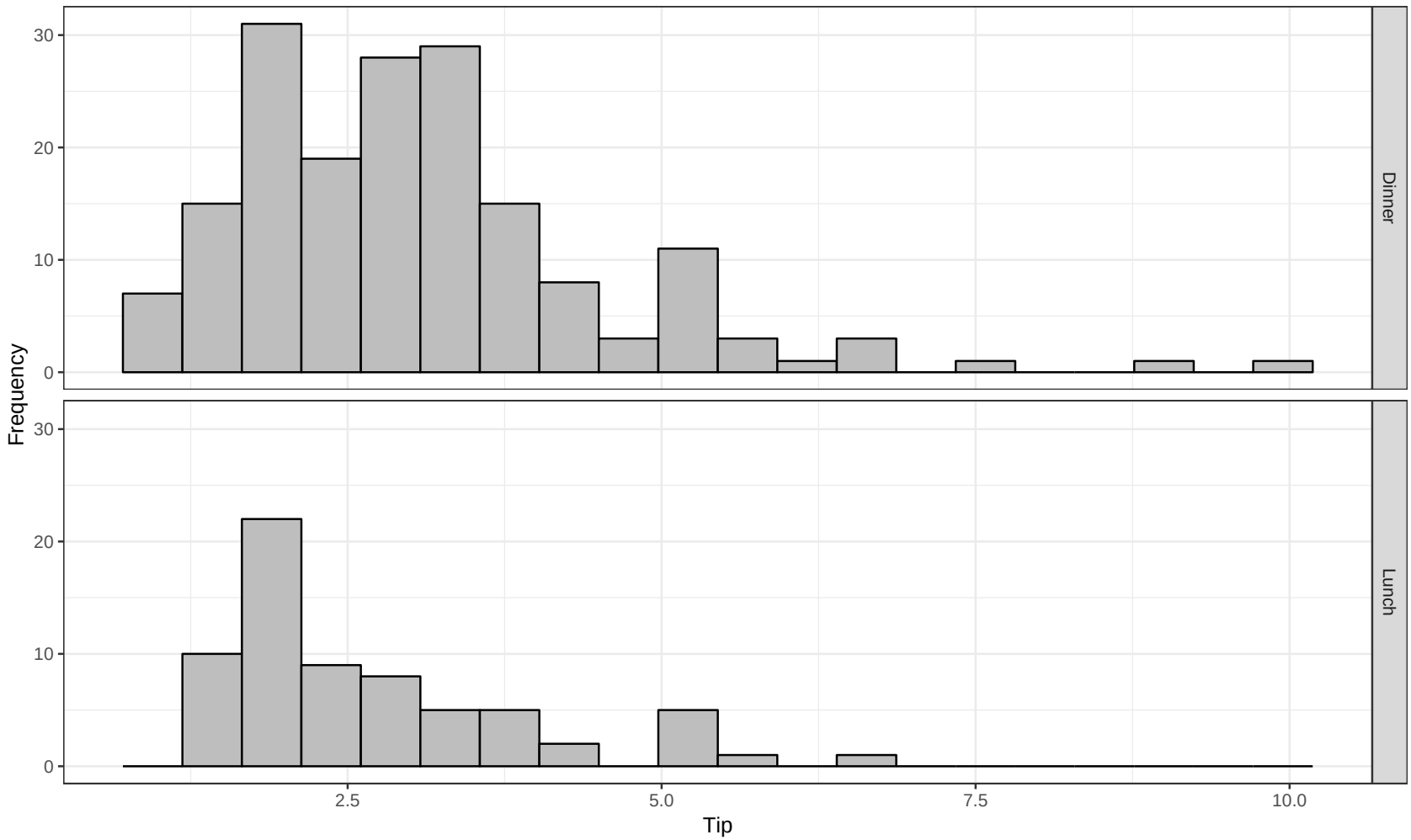
Association

- Most things we'd like to learn from our data involve two (or more) variables
- Two variables are **associated** if certain values of one variable tend to correspond with certain values of the other variable
- For example, the **two-way frequency table** below suggests "table size" and "time" *are associated*
  - 76.5% of lunches have size $= 2$, while only 59.1% of dinners have size $= 2$

Size	Dinner	Lunch
1	2	2
2	104	52
3	33	5
4	32	5
5	4	1
6	1	3

Practice

Using the graph below, are the variables “time” and “tip” associated? Be prepared to explain why or why not.

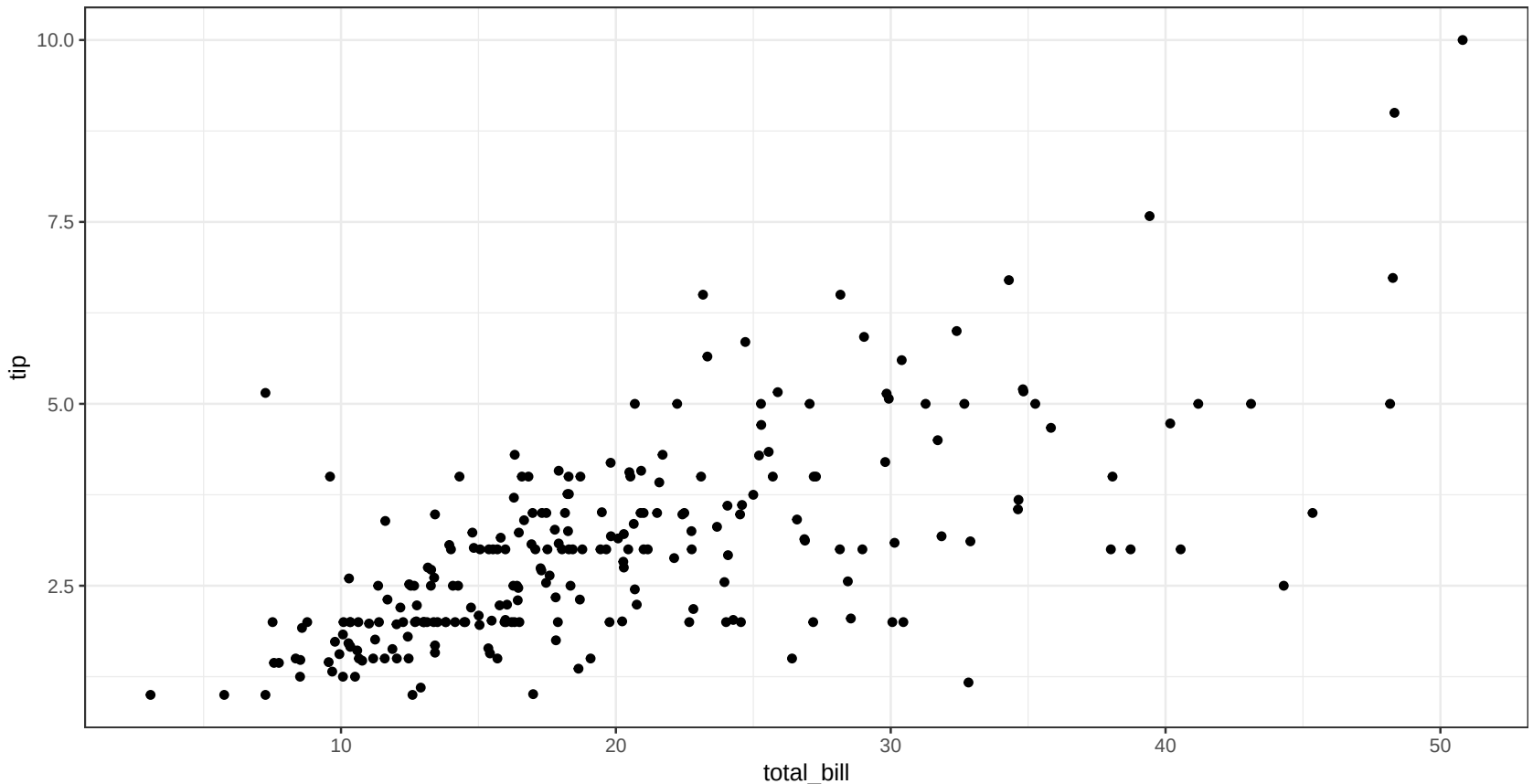


Explanatory and Response Variables

- When discussing association, we tend to think about *cause and effect*
  - "time" could influence "tip", but "tip" couldn't possibly influence "time"
- In this regard, an **explanatory variable** is one that is used to understand or predict a **response variable**
  - Not every two-variable relationship requires the designation of explanatory and response variables
  - Systolic blood pressure is strongly associated with diastolic blood pressure, but neither "explains" the other
- We will revisit *cause and effect* soon, for now we'll use the general term "association" when discussing relationships between variables, and *we'll avoid reading too much into why associations exist*

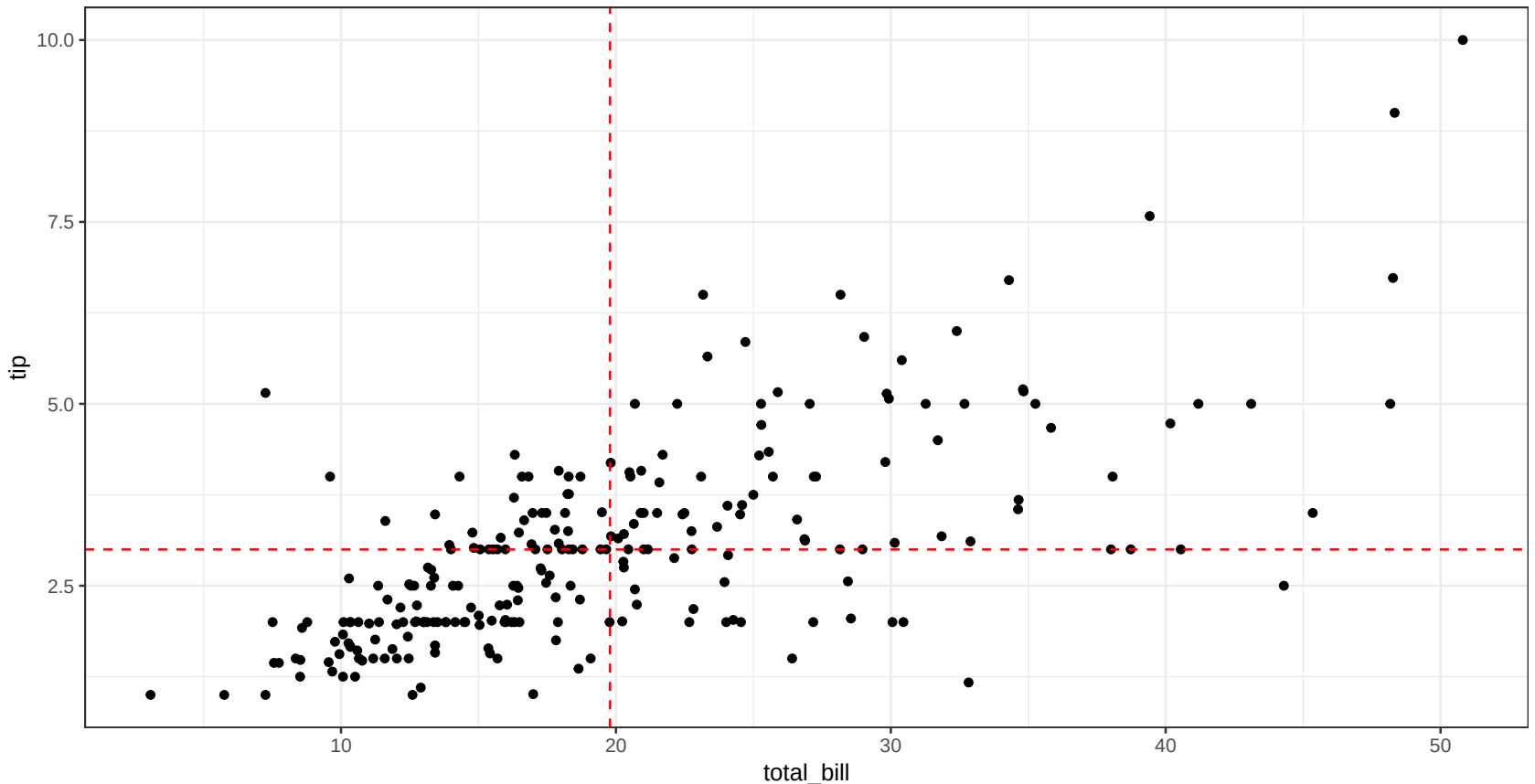
Practice

Using the scatterplot below, are the variables "total_bill" and "tip" associated? Why or why not? Which variable makes more sense to consider as an explanatory variable?



Practice - Solution

Dividing the scatterplot into quadrants (using each variable's mean), an association is evidenced by the abundance of data in the upper-right and lower-left quadrants.



Measuring Association

- Association can be quantified numerically depending upon the types of the variables in question
- For two categorical variables, association can be measured using **differences in proportions**
  - The proportion of tables with exactly 2 patrons is 0.174 higher for lunches than for dinners
- For one quantitative and one categorical variable, it can be measured using **differences in means**
  - The mean tip is \$1.6 higher for dinners than it is for lunches
- For two quantitative variables, it can be measured using the **correlation coefficient**
  - The correlation between tip and total bill is 0.676, suggesting higher bills are associated with higher tips
  - More info on the correlation coefficient is coming in Lab #1

Foreshadowing

- For the time being, we're going to focus on measuring and describing associations *in the data we are analyzing*
- For much of the remainder of this course, we'll learn about how to properly generalize associations that we find, using statistical methods to help us make broader conclusions

Conclusion

Right now, you should:

1. Understand the usefulness in summarizing data
2. Know the definition of association, how to identify when variables are associated, and how to quantify an association

If you want more information:

- Read Ch 2.1-2.4
- Read the Bradford Hill criteria (link) for causation