Bivariate Summaries

Ryan Miller

Outline

1. Relationships between variables
   - Explanatory and response variables, association vs. causation
2. Describing associations (two categorical variables)
   - Contingency tables, conditional proportions, stacked bar charts
3. Describing associations (one categorical and one quantitative variable)
   - Side-by-side graphs, differences in distribution

Relationships between variables

Two variables, X and Y, are **associated** if the values of X share a relationship with the values of Y

- Usually, we designate an **explanatory variable** (suspected cause) and a **response variable** (suspected outcome)
  - This is done using prior knowledge (ie: Exam #1 score could cause final grade, but not vice versa)

Relationships between variables

Two variables, X and Y, are **associated** if the values of X share a relationship with the values of Y

- Usually, we designate an **explanatory variable** (suspected cause) and a **response variable** (suspected outcome)
  - This is done using prior knowledge (ie: Exam #1 score could cause final grade, but not vice versa)

Note:

1. Association is general term, there are more specific types of association (ie: linear, non-linear, etc.)

Relationships between variables

Two variables, X and Y, are **associated** if the values of X share a relationship with the values of Y

- Usually, we designate an **explanatory variable** (suspected cause) and a **response variable** (suspected outcome)
  - This is done using prior knowledge (ie: Exam #1 score could cause final grade, but not vice versa)

Note:

1. Association is general term, there are more specific types of association (ie: linear, non-linear, etc.)
2. Observing an association between X and Y does not imply that X causes Y, or that Y causes X, *causation* is a complex topic that we'll discuss soon

Two categorical variables

A **contingency table** (two-way frequency table) is a straightforward way of expressing relationships between two categorical variables:

Table 1: Birth age as a risk factor for breast cancer in a CDC cohort study initiated in the 1980s

	Cancer	No
First birth before 25	65	4475
First birth at 25 or later	31	1597

Based upon the data in this table, is there an association between these variables?

Analyzing contingency tables

Conditional proportions (ie: row proportions or column proportions) are used to find associations in contingency tables:

Table 2: Row proportions for Table 1 (see prev slide)

	Cancer	No
First birth before 25	0.0143	0.9857
First birth at 25 or later	0.0190	0.9810

Analyzing contingency tables

- Table 2 shows a slight association
  - The **risk difference** (ie: difference in conditional proportions) is 0.0190 - 0.0143 = 0.0047, or about half of a percent

Analyzing contingency tables

- Table 2 shows a slight association
  - The **risk difference** (ie: difference in conditional proportions) is 0.0190 - 0.0143 = 0.0047, or about half of a percent
  - The **relative risk** (ie: ratio of conditional proportions) is 0.0190/0.0143 = 1.33, suggesting a 33% increase in the risk of breast cancer for women belonging to the birth after 25 group
- Whether the observed association is small enough to be explained by *random chance* is a topic we'll discuss later

Practice

The contingency table below describes the survival of crew members and first class passengers aboard the Titanic cruise ship:

	Survived	Died
Crew	212	673
1st Class	203	122

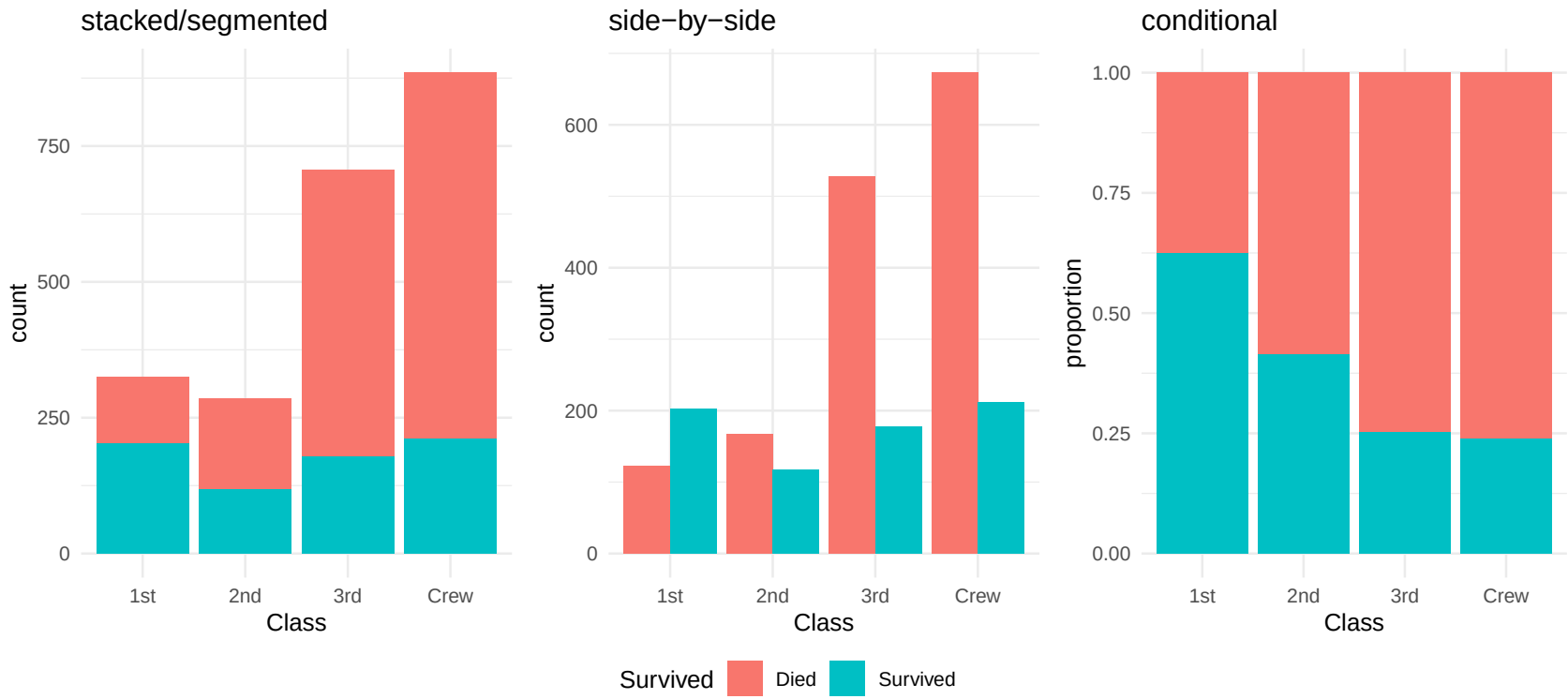
1) Which group was more likely to survive the shipwreck?
2) Did you use row or column proportions? Why is the other choice unable to answer this question?

Practice (solution)

1) Using *row proportions*, $\frac{212}{623+212} = 0.24$, or 24% of the crew survived; while $\frac{203}{122+203} = 0.62$, or 62% of first class passengers survived.
2) This question cannot be answered using column proportions. Notice the proportion of survivors who were crew is $\frac{212}{212+203} = 0.51$, while the proportion of survivors who were first class passengers is $\frac{203}{212+203} = 0.49$
   - Conditioning on the column variable is problematic here because the *marginal distribution* of 1st class/crew is *skewed towards crew*
   - In other words, most of the survivors were crew members because there were so many more crew members, not because the individual crew members were more likely to survive

Graphing two categorical variables

There are several ways to create bar charts that depict two categorical variables:



Which graph most clearly displays the association? Are there any limitations of this choice?

One categorical and one quantitative variable

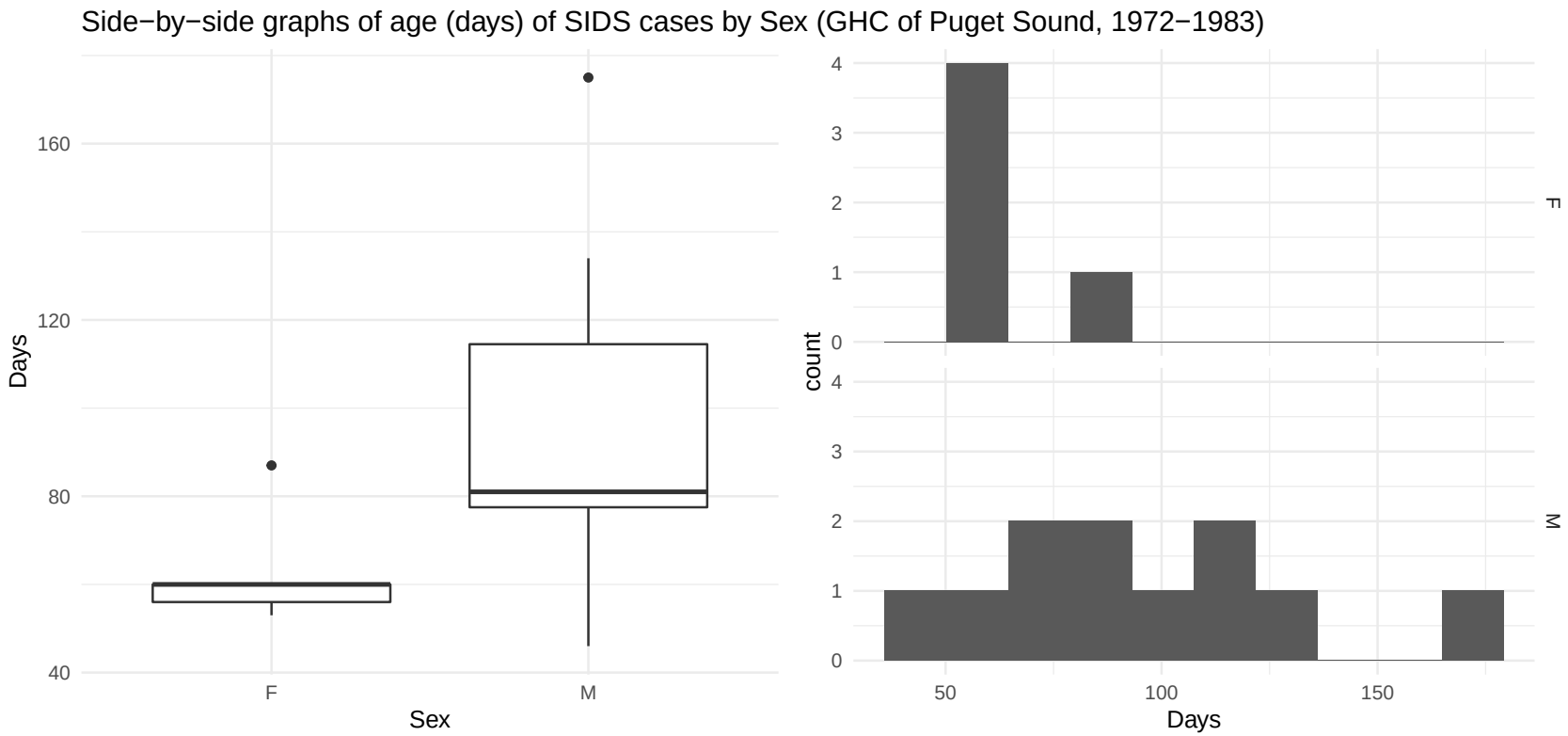
Graphing the association

- Side-by-side graphs, such as boxplots or histograms

Describing or quantifying the association

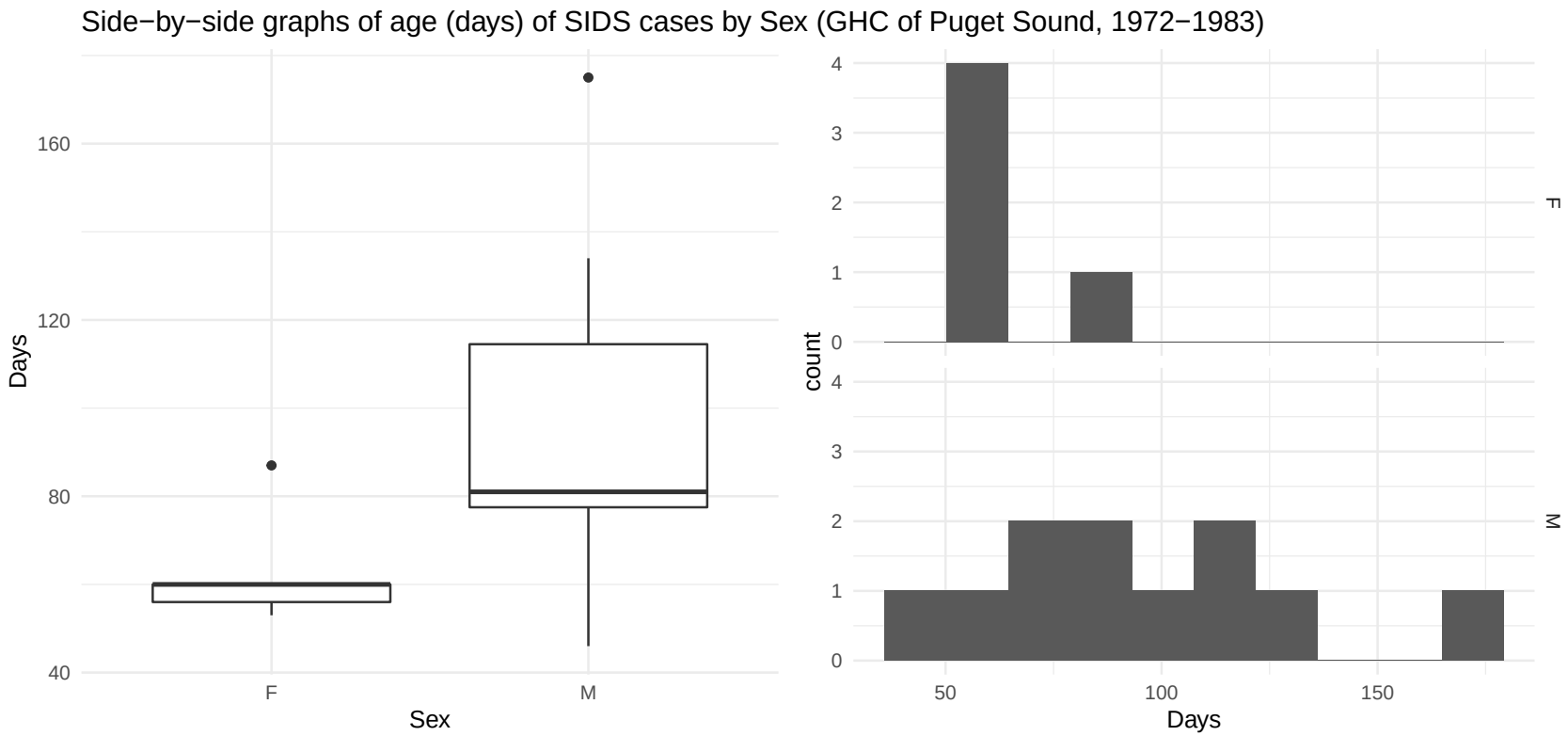
- Comparative summary statistics (ie: side-by-side comparisons or differences in means/medians/etc.)

Side-by-side graphs



Does sex appear to be associated with the age of SIDS cases following diphtheria-tetanus-pertussis (DTP) immunization?

Side-by-side graphs



Does sex appear to be associated with the age of SIDS cases following diphtheria-tetanus-pertussis (DTP) immunization?

- Yes, the medians appear to differ by sex

Comparative summary statistics

Table 3: Comparative summary statistics for the age (days) of SIDS cases by Sex (GHC of Puget Sound, 1972-1983)

Sex	F	M
Min	53	46
Q1	60.0	114.5
Mean	63.20000	96.45455
Median	60	81
Max	87	175
StDev	13.62718	36.77870

Practice

The Breast Cancer Deaths Dataset is available at this link, or under "Data" on our course website. The variables we will focus on are:

- "Time" - the number days the patient survived after beginning treatment
- "Grade" - tumor classification type (higher is worse)
- "Cycles" - the number of chemotherapy cycles the patient had undergone

Note: these data are subset of a larger study and were filtered to exclude survivors

1) Use StatKey to determine whether the "Grade" and "Cycles" are associated (Hint: treat "Cycles" as categorical)
2) Use StatKey to determine whether the "Time" and "Grade" are associated

Practice (solution)

1) Yes - these variables appear to be associated. Using the "Two Categorical Variables" menu on StatKey, we see that 41.4% of patients with Grade III tumors undergo 6 cycles, 47.7% of Grade II undergo 6 cycles, and 76.9% of Grade I undergo 6 cycles. Because these conditional proportions are so different, it seems that higher grades are associated with decreased changes of completing 6 chemotherapy cycles.
2) Yes - these variable appear to be associated. Using the "one Quantitative and one Categorical Variable" menu on StatKey, we see that the average time for Grade I patients is 983 days, but it is only 632 days for Grade III patients.

Common misconceptions

1) When a nominal categorical variable is involved in a comparison, a difference between at least two groups is sufficient to establish an association (even if the other groups have identical distributions)
2) While we typically focus on the mean or median, an association can be evidenced by meaningful differences in any segment of a distribution (ie: differences in Q3 despite there being equal means/medians)

Conclusion

1. Relationships between variables
   - An *explanatory variable* is suspected to cause changes in a *response variable*
   - Associations between explanatory and response variables do not prove *causation* (no matter how strong)
2. Describing associations (two categorical variables)
   - Associations should be expressed using contingency tables and conditional proportions (ie: differences in proportions)
   - Bar charts can be used for graphical displays
3. Describing associations (one categorical and one quantitative variable)
   - Associations should be expressed using comparative summary statistics (ie: differences in means)
   - Side-by-side boxplots or histograms provide graphical displays