Descriptive statistics

Patrick Breheny

February 5, 2026

Descriptive statistics

- Humans are not good at extracting patterns from large amounts of raw data; we understand data much better when it is summarized
- *Descriptive statistics* summarize data in a way that highlights important features and patterns
- These summaries are usually presented as tables or figures

Tables and figures

- Tables of descriptive statistics are extremely common; nearly all published studies include at least one table describing their sample
- Figures are often better than tables at revealing trends, patterns, and relationships in the data
- Graphics also play a central role in "exploratory" data analysis — as we'll see today, plots often reveal aspects of the data that summary statistics can hide
- With modern computing making it easy to produce high-quality graphics, figures now serve as the primary way scientific results are communicated

Today's focus

- Today we will cover
  - Descriptive statistics like the mean, median, and standard deviation
  - The difference between continuous and categorical data
  - Several important graphical summaries like the bar plot, box plot, and histogram
- In addition to going over the underlying concepts, I'll also discuss how to carry out these calculations and make these figures in R, although we'll go into more of the details in the next lab

Types of data

- The best way to summarize and present data depends on the type of data
- There are two main types of data:
  - *Categorical data*: Data that takes on distinct values (*i.e.*, it falls into categories), such as sex (male/female), alive/dead, blood type (A/B/AB/O), stages of cancer
  - *Continuous data*: Data that takes on a spectrum of fractional values, such as time, age, temperature, cholesterol levels
- The distinction between categorical (also called *discrete*) and continuous data is fundamental and we will return to it throughout the course

Summarizing categorical data

- Summarizing categorical data is pretty straightforward – you just *count* how many times each category occurs (also the *frequency*)
- Instead of counts, we are often interested in *proportions* or *percents* (which are proportions times 100)
- A percent is a special type of *rate*, a rate per hundred
- Counts (frequencies), percents, and rates are the basic summary statistics for categorical data, and are often displayed in tables or bar charts

Frequencies in R

- Suppose our admissions data from assignment 1 is called `ucb`
- In R, the function to obtain frequencies is called `xtabs()`, short for "cross-tabulation"

```
xtabs(~ Dept, ucb)
# Dept
#   A   B   C   D   E   F
# 933 585 918 792 584 714
xtabs(~ Gender + Admit, ucb)
#          Admit
# Gender    Admitted Rejected
#   Male        1198     1493
#   Female       557     1278
```

Proportions

To calculate proportions, the R function is `proportions()`:

```
xtabs(~ Dept, ucb) |>
  proportions()
# Dept
#      A      B      C      D      E      F
# 0.2061 0.1293 0.2028 0.1750 0.1290 0.1578
```

Note the use of the pipe operator |> here, which is a convenient way to chain together multiple operations

Margins

For multi-way tables, we often want to pass an argument to `proportions()`; here, specifying `"Gender"` means that we want separate proportions for each gender:

```
xtabs(~ Gender + Admit, ucb) |>
  proportions("Gender")
#          Admit
# Gender    Admitted Rejected
#   Male      0.4452   0.5548
#   Female    0.3035   0.6965
```

You can also type `proportions(1)` if you prefer less typing (since `Gender` is the first term in the formula)

Three-way tables

The same logic extends to three-way tables:

```
xtabs(~ Gender + Admit + Dept, ucb) |>
  proportions(c("Gender", "Dept"))
```

the function `c()` here stands for "combine"; we want to calculate the proportion over both gender and department here

Bar plots

- The most common graphical way of displaying summary statistics for categorical data is the *bar chart*, in which the height of a bar represents the frequency of the category
- A common variation of the bar plot is the *stacked bar chart*, which provide a visual way to break down a category into smaller groups as we do in a cross-tabulation
- Finally, *faceting* (or "small multiples") repeats the same plot across subsets of the data, enabling us to see how relationships change

ggplot

- R has a number of optional add-on packages called *libraries*
- We won't really use them in this class, with one exception: R has a very useful library for plotting called `ggplot2`
- To load the library, we use the `library()` function:

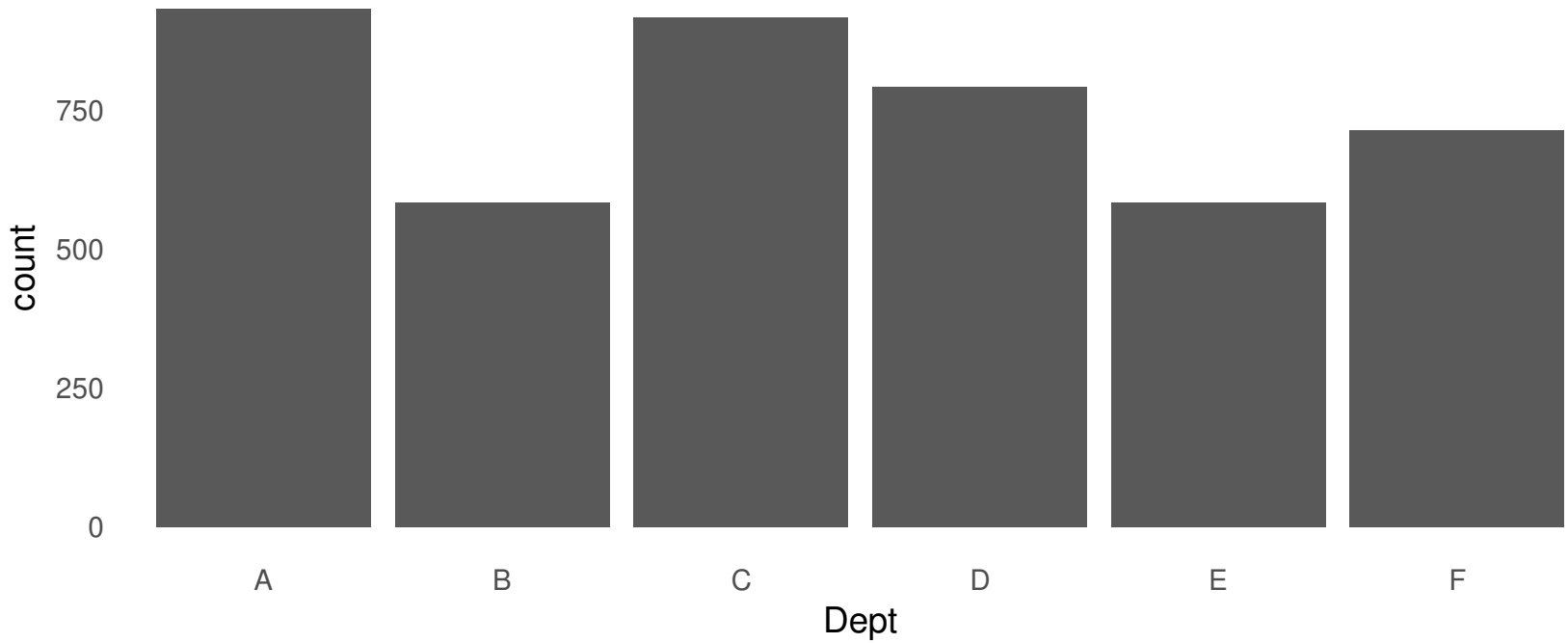
```
library(ggplot2)
```

ggplot syntax

- Once it's been loaded, using ggplot is essentially a three-step process
  - First, we specify the data set we're using
  - Second, we set up the aesthetics: what we want on the horizontal and vertical axes
  - Finally, we say what type of geometric object we're using to represent the data
- In our case, my infant mortality data is called `infmort`, I want the infant mortality rate on the horizontal axis, and I want to draw a histogram

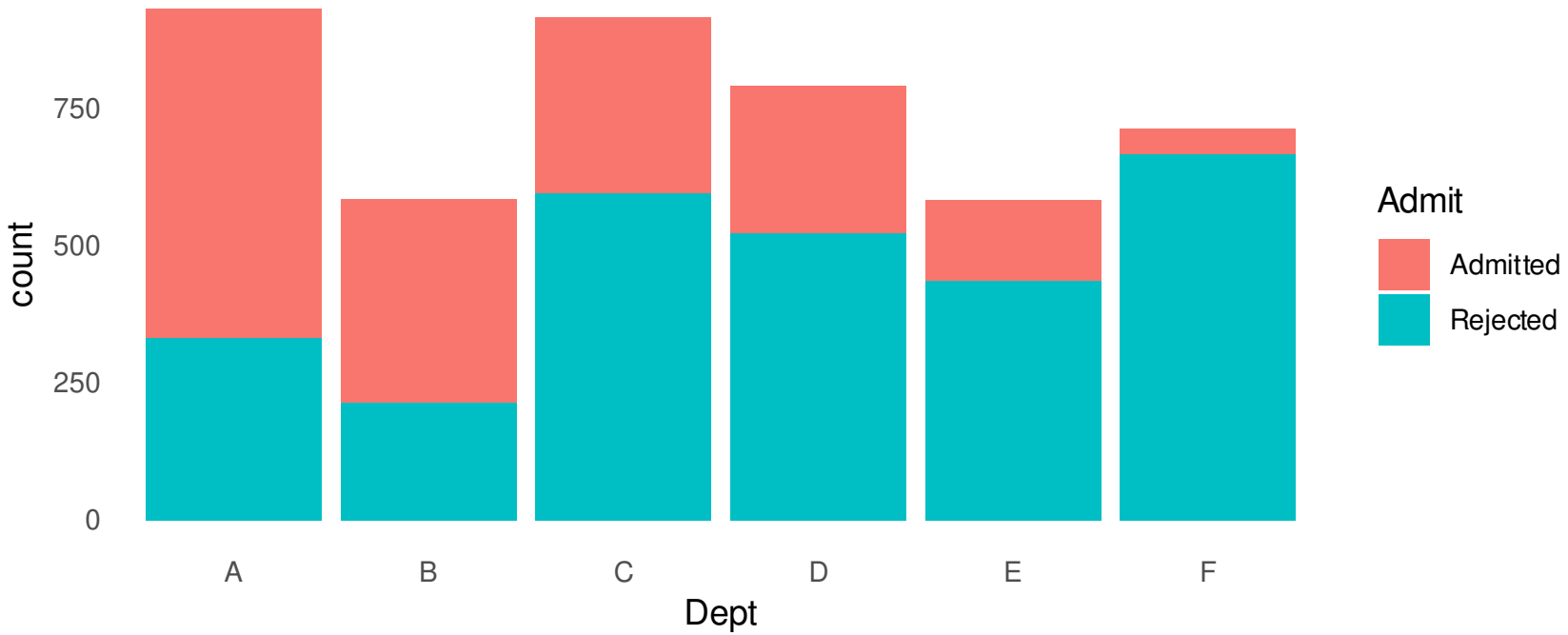
A basic bar chart

```
ggplot(ucb, aes(Dept)) + geom_bar()
```



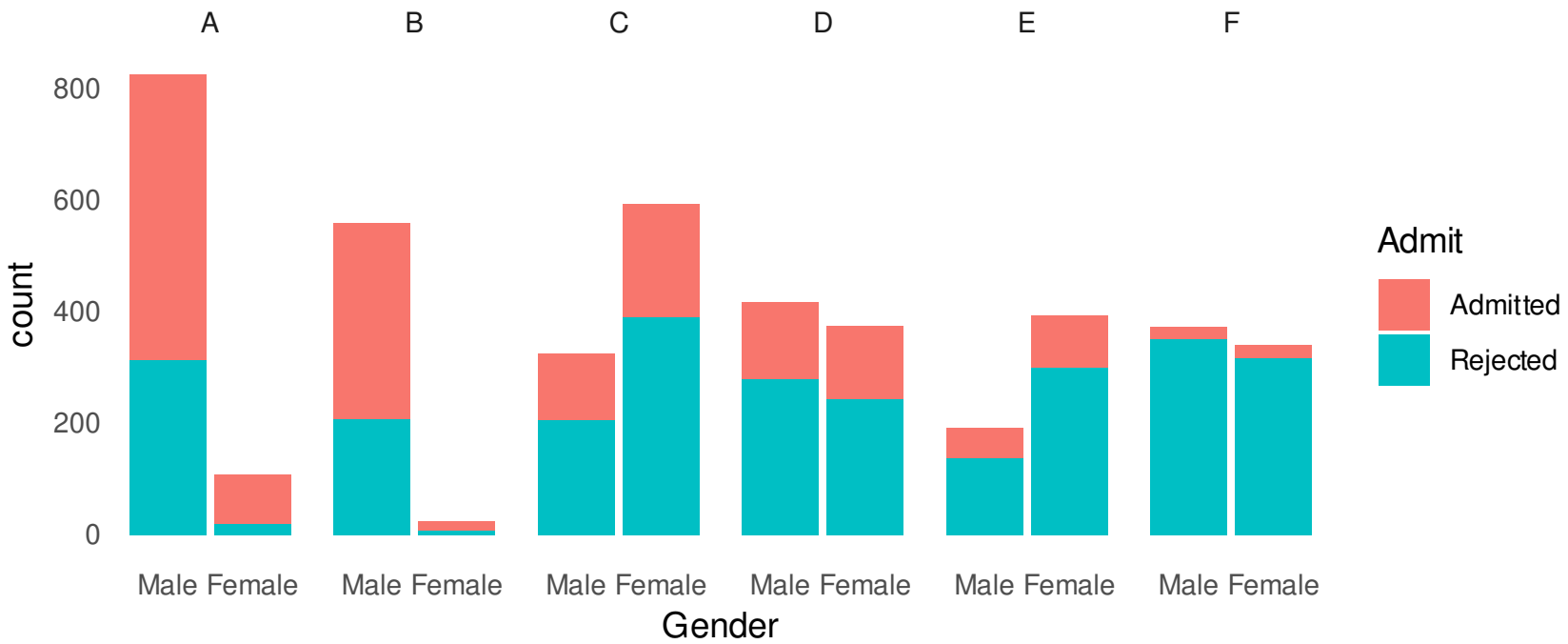
A stacked bar chart

```
ggplot(ucb, aes(Dept, fill = Admit)) +
  geom_bar()
```



A faceted stacked bar chart

```
ggplot(ucb, aes(Gender, fill = Admit)) +
  geom_bar() +
  facet_grid(~ Dept)
```



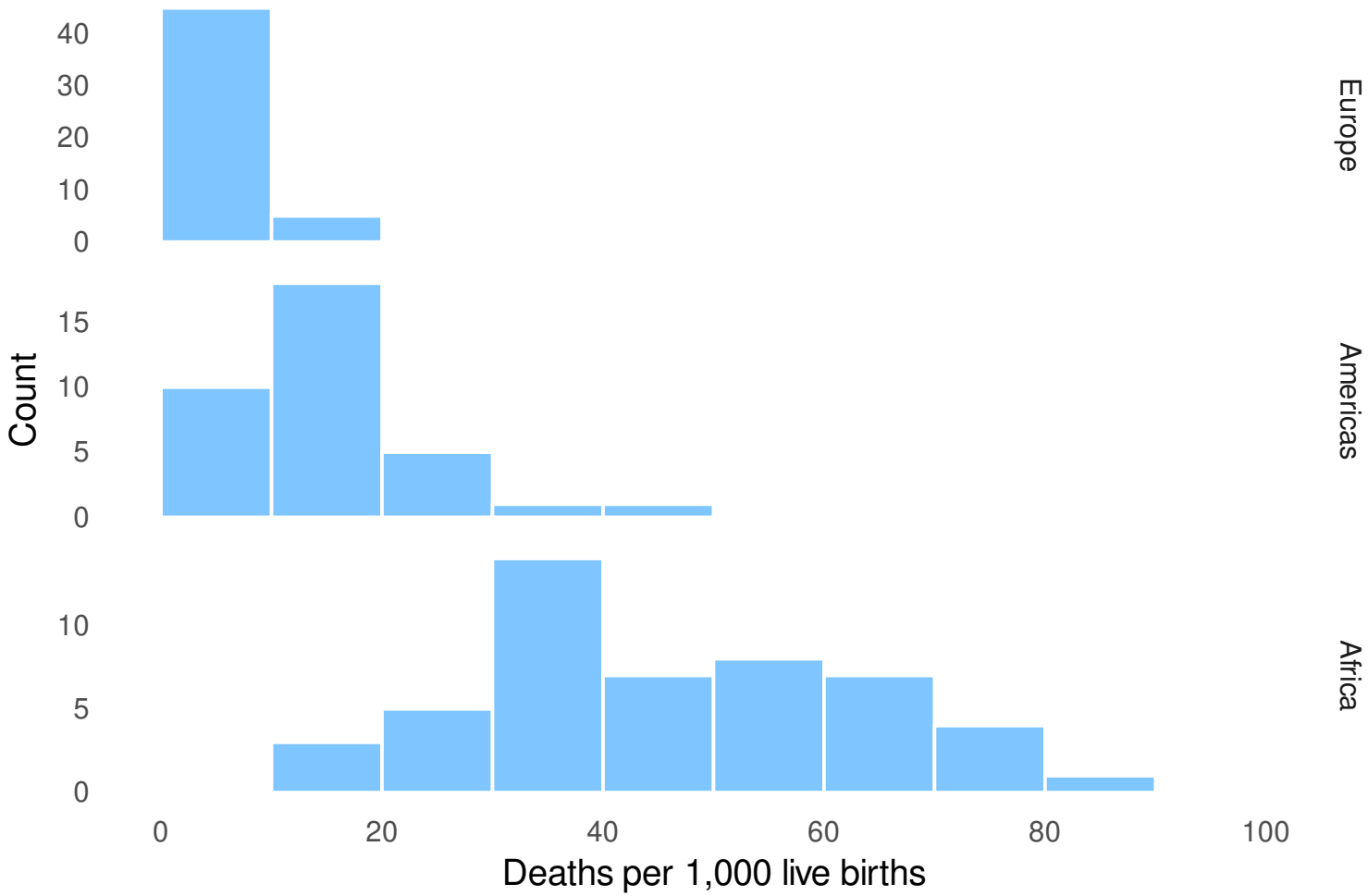
Continuous data

- For continuous data, instead of a finite number of categories, observations can take on a potentially infinite number of values
- Summarizing continuous data is therefore much less straightforward
- To introduce concepts for describing and summarizing continuous data, we will look at data on infant mortality rates from 2019 for 134 nations in three geographical regions: Africa, Europe, and the Americas

Histograms

- One very useful way of looking at continuous data is with *histograms*
- To make a histogram, we divide a continuous axis into equally spaced intervals, then count and plot the number of observations that fall into each interval
- This allows us to see how our data points are distributed

Infant mortality rate histograms



Summarizing continuous data

- As we can see, continuous data comes in a variety of shapes
- Nothing can replace seeing the picture, but if we had to summarize our data using just one or two numbers, how should we go about doing it?
- The aspect of the histogram we are usually most interested in is, "Where is its center?"
- This is typically represented by the average

```
mean(infmort$Rate)
# [1] 22.57
```

Calculating averages by subgroup

If you want to calculate the mean separately for different subgroups, you can use the `by()` function:

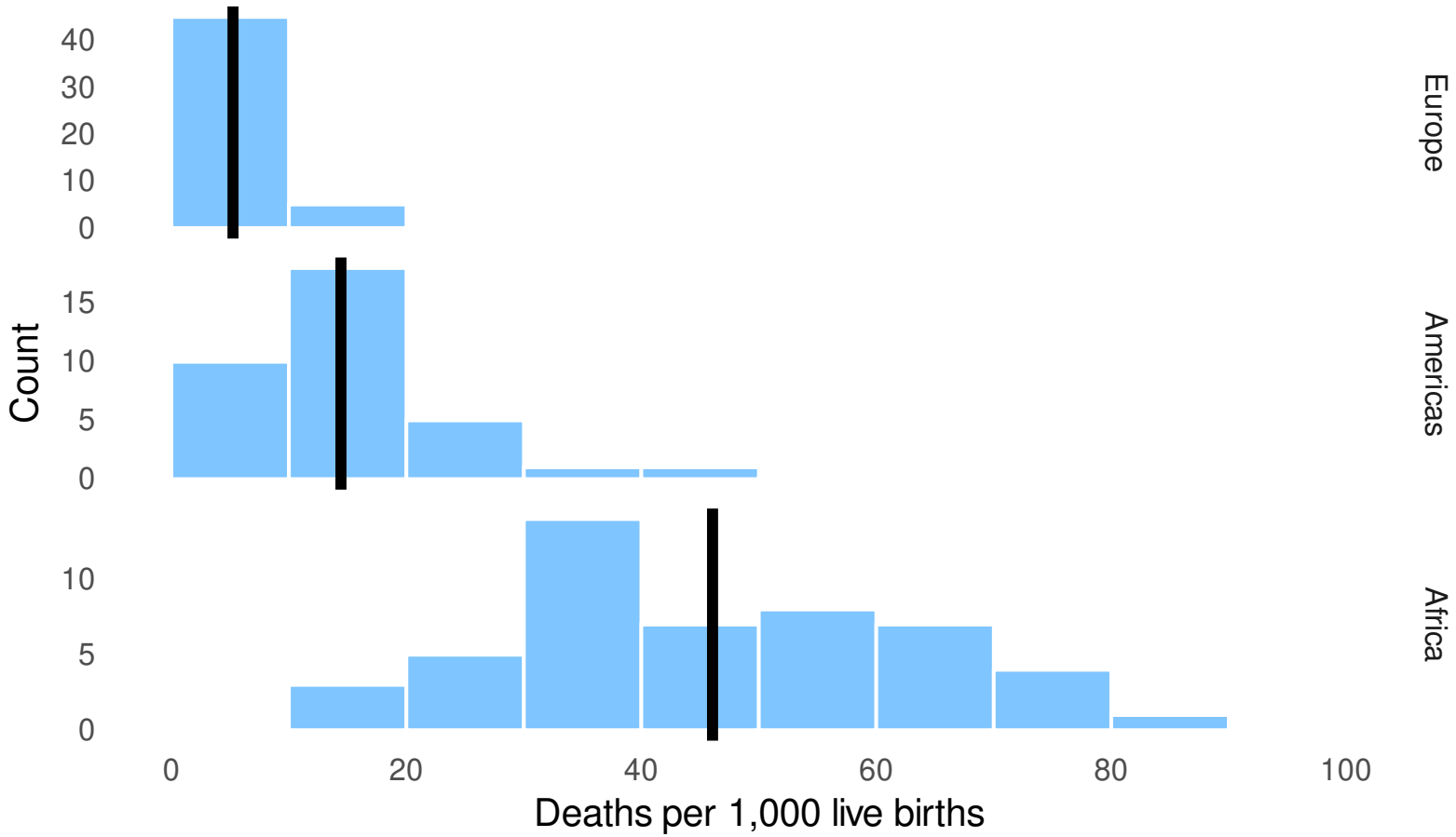
```
by(infmort$Rate, infmort$Region, mean)
# infmort$Region: Europe
# [1] 5.24
# ------------------------------------------------------------
# infmort$Region: Americas
# [1] 14.43
# ------------------------------------------------------------
# infmort$Region: Africa
# [1] 46.06
```

Alternatively, if you want to avoid typing the name of the data set each time, you can use the `with()` function:

```
with(infmort, by(Rate, Region, mean))
```

The average and the histogram

The average represents the center of mass of the histogram:



Spread

- The second most important bit of information from the histogram to summarize is, "How spread out are the observations around the center"?
- This is most typically represented by the *standard deviation*
- To understand how standard deviation works, let's consider the numbers $\{4, 5, 1, 9\}$
- Each of these numbers deviates from the mean by some amount:

$$4 - 4.75 = -0.75 \quad 5 - 4.75 = 0.25$$
$$1 - 4.75 = -3.75 \quad 9 - 4.75 = 4.25$$

- How should we measure the overall size of these deviations?

Root-mean-square

- Taking their mean isn't going to tell us anything (why not?)
- We could take the average of their absolute values:

$$\frac{|-0.75| + |0.25| + |-3.75| + |4.25|}{4} = 2.25$$

- But it turns out that for a variety of reasons, the *root-mean-square* works better as a measure of overall size:

$$\sqrt{\frac{(-0.75)^2 + (0.25)^2 + (-3.75)^2 + (4.25)^2}{4}} \approx 2.86$$

The standard deviation

- The formula for the standard deviation is

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- Wait a minute; why $n-1$?
- The reason (which we will discuss further in a few weeks) is that dividing by n turns out to underestimate the true standard deviation
- Dividing by $n-1$ instead of n corrects some of that bias

Standard deviation in R

- The standard deviation of $\{4, 5, 1, 9\}$ is 3.30:

```
sd(c(4, 5, 1, 9))
# [1] 3.304
```

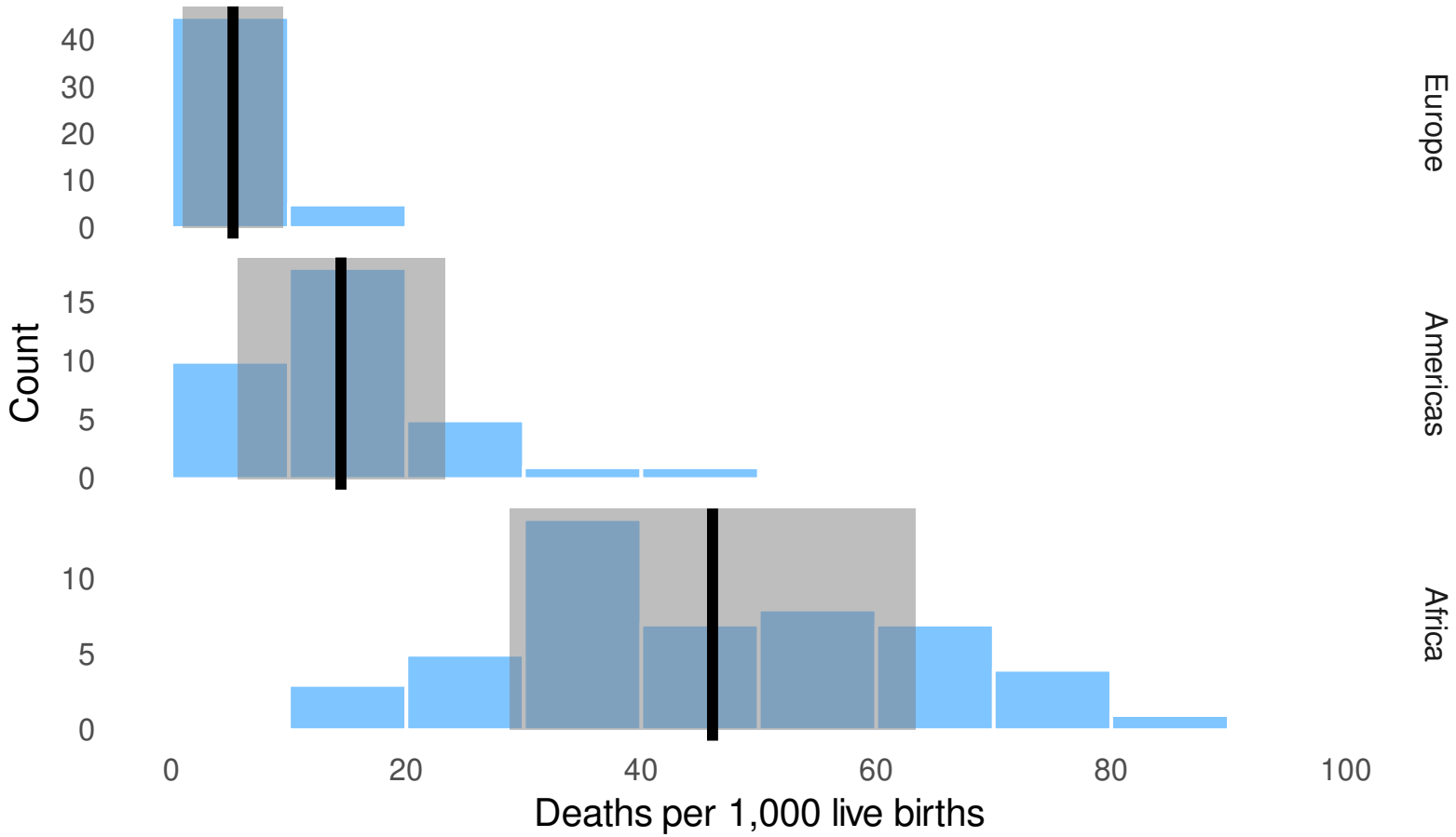
- Recall that we got 2.86 if we divide by n
- As n gets larger, the difference between the two is minimal

Meaning of the standard deviation

- The standard deviation (SD) describes how far away numbers in a list are from their average
- The SD is often used as a "plus or minus" number, as in "adult women tend to be about 5'4, plus or minus 3 inches"
- Most numbers (roughly 68%) will be within 1 SD away from the average
- Very few entries (roughly 5%) will be more than 2 SD away from the average
- This rule of thumb works very well for a wide variety of data; we'll discuss where these numbers come from in a few weeks

Standard deviation and the histogram

Background areas within 1 SD of the mean are shaded:

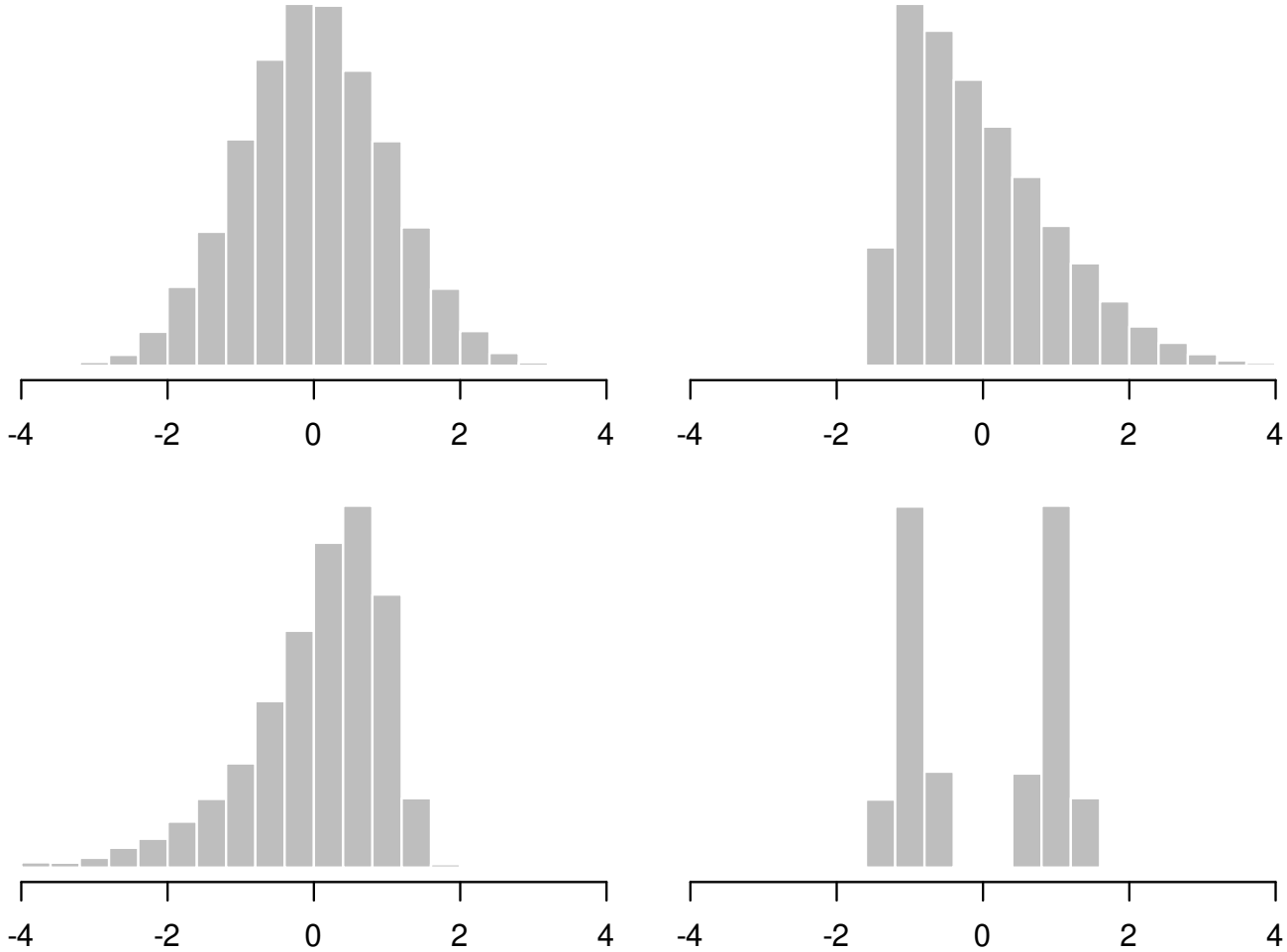


The 68%/95% rule in action

Region	SD1	SD2
Europe	0.90	0.92
Africa	0.69	0.98
Americas	0.80	0.94

Summaries can be misleading!

All of the following have the same mean and standard deviation:



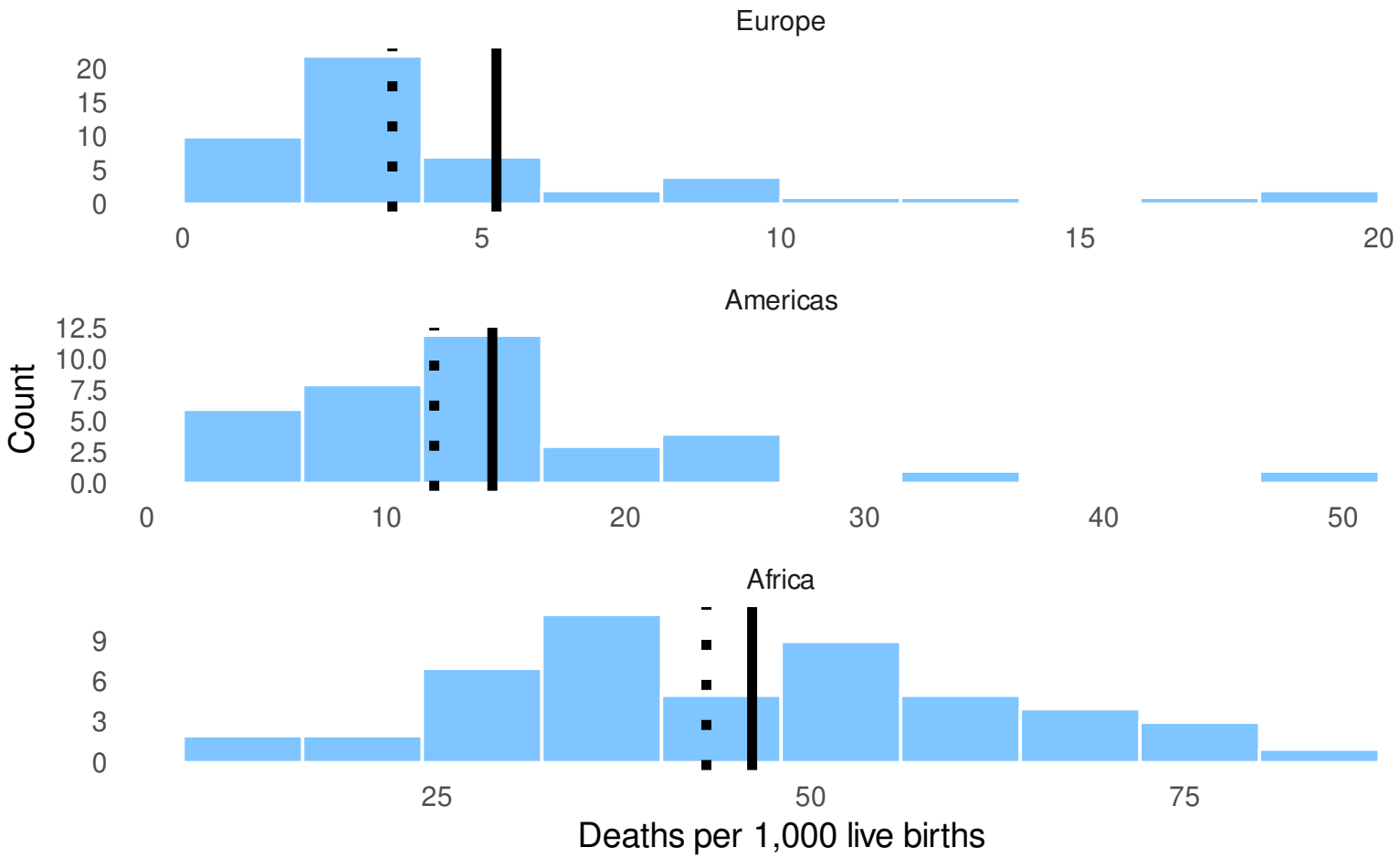
Percentiles

- The average and standard deviation are not the only ways to summarize continuous data
- Another type of summary is the *percentile*
- A number is the 25th percentile of a list of numbers if it is bigger than 25% of the numbers in the list
- The 50th percentile is given a special name: the *median*
- The median, like the mean, can be used to answer the question, "Where is the center of the histogram?"

```
median(infmort$Rate)
# [1] 13.5
quantile(infmort$Rate, 0.25)
#  25%
# 4.25
```

Median vs. mean

The dotted line is the median, the solid line is the mean:



Skew

- Focusing on Europe in particular, note that the histogram is not symmetric: the *tail* of the distribution extends further to the right than it does to the left
- Such distributions are called *skewed*
- The distribution of infant mortality rates in Europe is said to be *right skewed* or *skewed to the right*
- For asymmetric/skewed data, the mean and the median will be different

Hypothetical example

- Haiti had the highest infant mortality rate in the Americas at 49
- What if, instead of 49, it was 200?

	Mean	Median
Real	14.4	12
Hypothetical	18.7	12

- The mean is now higher than 80% of the countries in the Americas
- Note that the average is sensitive to extreme values, while the median is not; statisticians say that the median is *robust* to the presence of outlying observations

Five number summary

- The mean and standard deviation are a common way of providing a "two-number summary" for continuous variables
- Another approach, based on quantiles, is to provide a "five-number summary" consisting of: (1) the minimum, (2) the first quartile, (3) the median, (4) the third quartile, and (5) the maximum

	Europe	Americas	Africa
0%	1.0	4.0	12
25%	3.0	9.5	34
50%	3.5	12.0	43
75%	6.0	16.5	59
100%	19.0	49.0	84

The interquartile range

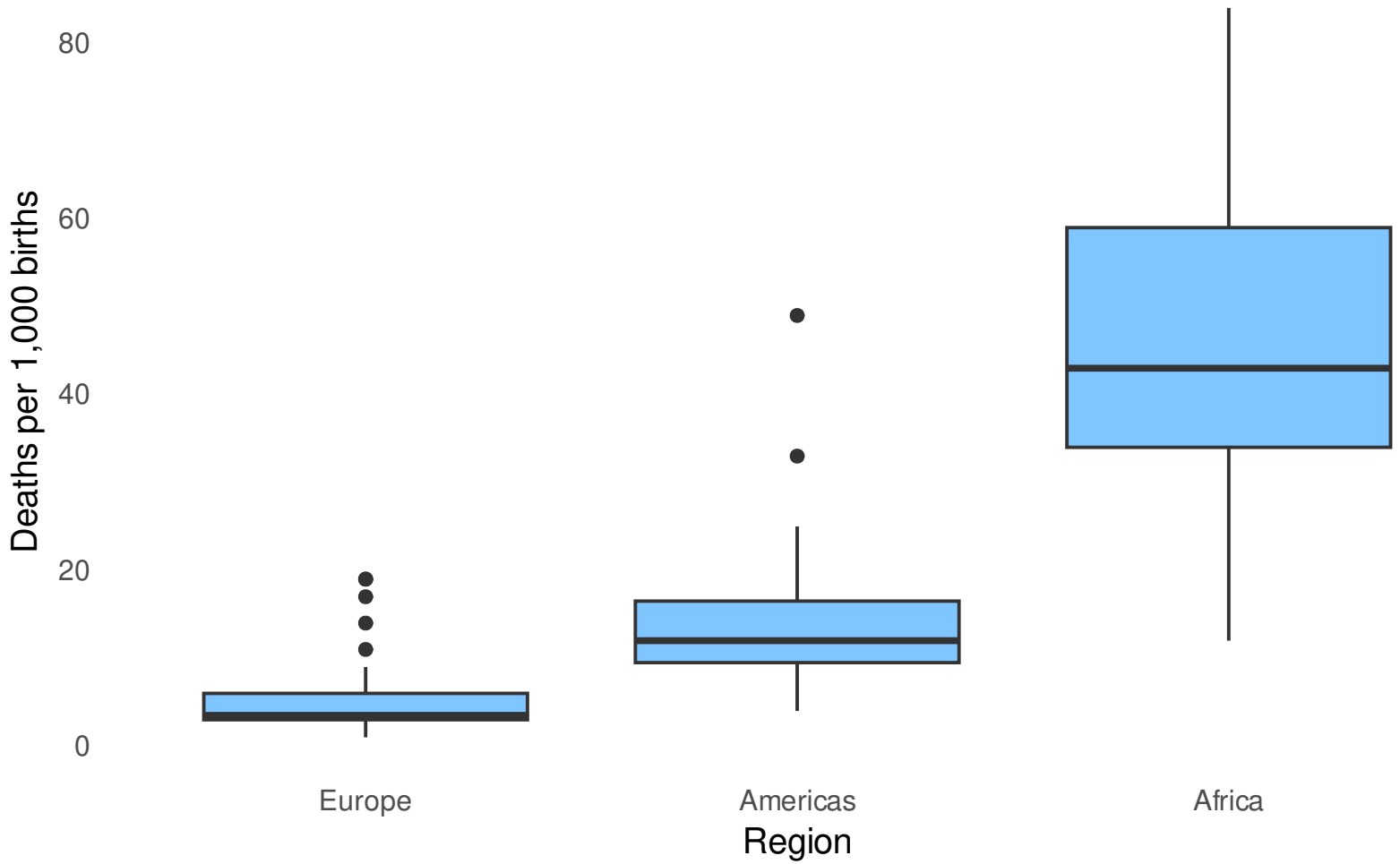
- We won't spend much time on this, but the distance from the first quartile (i.e., the 25th percentile) to the third quartile (75th percentile) is called the *interquartile range*, or IQR
- This range always contains the middle 50% of the data, and the IQR provides an alternative measure of how spread out the data is
- Returning to the hypothetical example in which Haiti's infant mortality rate is 200,

	SD	IQR
Real	8.8	7
Hypothetical	32.2	7

Box plots

- Quantiles are used in a type of graphical summary called a *box plot*
- Box plots are constructed as follows:
  - Calculate the three quartiles (the 25th, 50th, and 75th)
  - Draw a box bounded by the first and third quartiles and with a line in the middle for the median
  - Call any observation that is extremely far from the box an "outlier" and plot the observations using a special symbol (this is somewhat arbitrary and different rules exist for defining outliers)
  - Draw a line from the top of the box to the highest observation that is not an outlier; likewise for the lowest non-outlier

Box plots of the infant mortality rate data



Box plots and bar charts

- Bar charts provide an effective way of comparing categorical variables (e.g., admission and sex)
- Box plots provide a way to examine the relationship between a continuous variable and a categorical variable (e.g., infant mortality and continent)
- Next week, we will discuss how to summarize and illustrate the relationship between two continuous variables

Summary

- Raw data is complex and needs to be summarized; typically, these summaries are displayed in tables and figures
- Tables are useful for looking up information, but figures tend to be superior for illustrating trends in the data
- Summary measures for categorical variables: counts, percents, rates
- Plotting methods for categorical variables: bar charts
- Summary measures for continuous variables: mean, standard deviation, quantiles
- Plotting methods for continuous data: histogram, box plot