Descriptive statistics

Patrick Breheny

February 1

Tables and figures

- Human beings are not good at sifting through large streams of data; we understand data much better when it is summarized for us
- We often display summary statistics in one of two ways: *tables* and *figures*
- Tables of summary statistics are very common (we have already seen several in this course) – nearly all published studies in medicine and public health contain a table of basic summary statistics describing their sample
- However, figures are usually better than tables in terms of distilling clear trends from large amounts of information

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

Categorical data

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

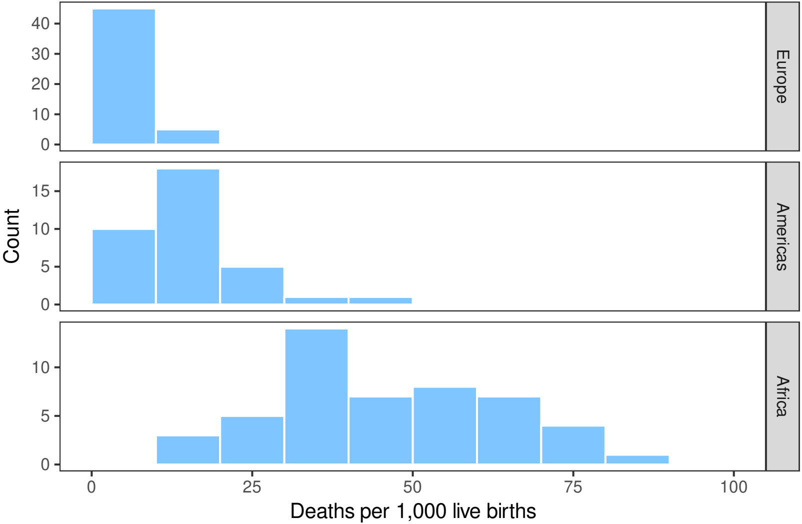
Continuous data

- 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

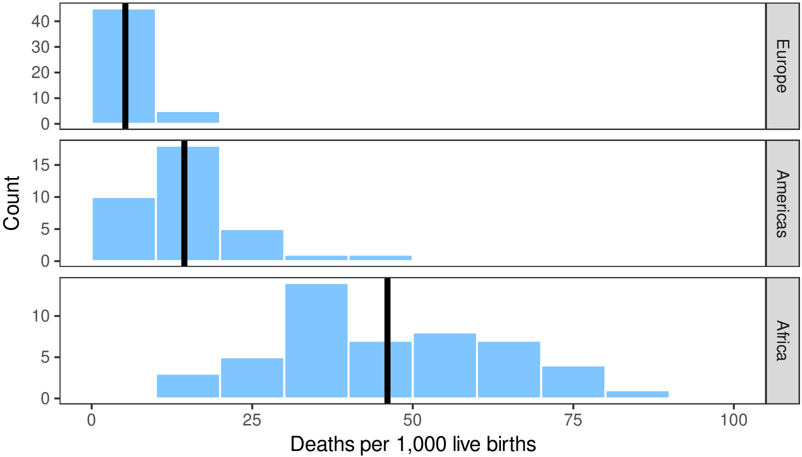


Summarizing continuous data

- 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

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 return to our small example with the numbers $\{4, 5, 1, 9\}$
- Each of these numbers deviates from the mean by some amount:

$$\begin{array}{ll} 4 - 4.75 = -0.75 & 5 - 4.75 = 0.25 \\ 1 - 4.75 = -3.75 & 9 - 4.75 = 4.25 \end{array}$$

- 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

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

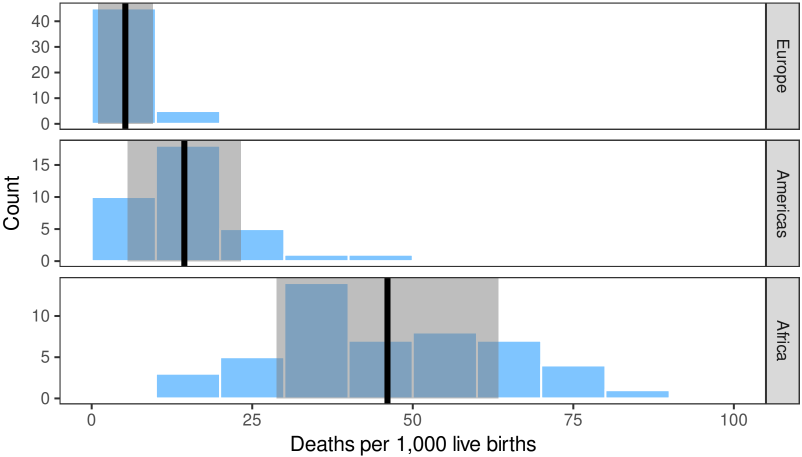
- Wait a minute; why $n-1$?
- The reason (which we will discuss further in a few weeks) is that dividing by n turns out to underestimate the true standard deviation
- Dividing by $n-1$ instead of n corrects some of that bias
- The standard deviation of $\{4, 5, 1, 9\}$ is 3.30 (recall that we got 2.86 if we divide by n)

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:

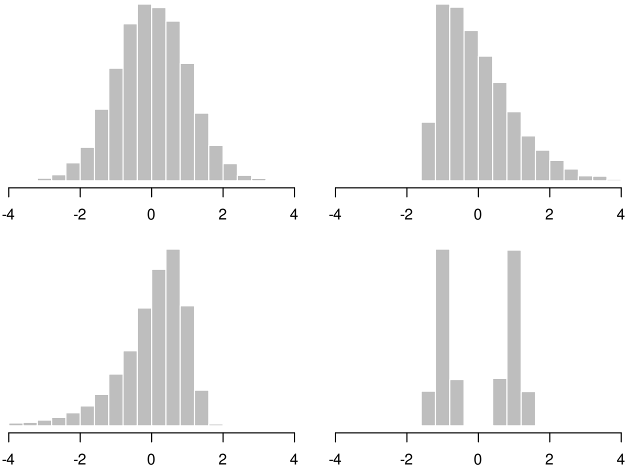


The 68%/95% rule in action

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:

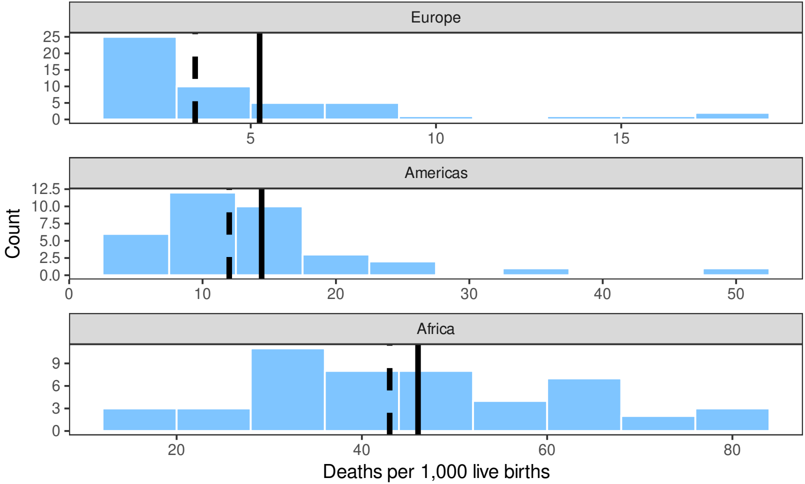


Percentiles

- 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 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 of a distribution of continuous values
- 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

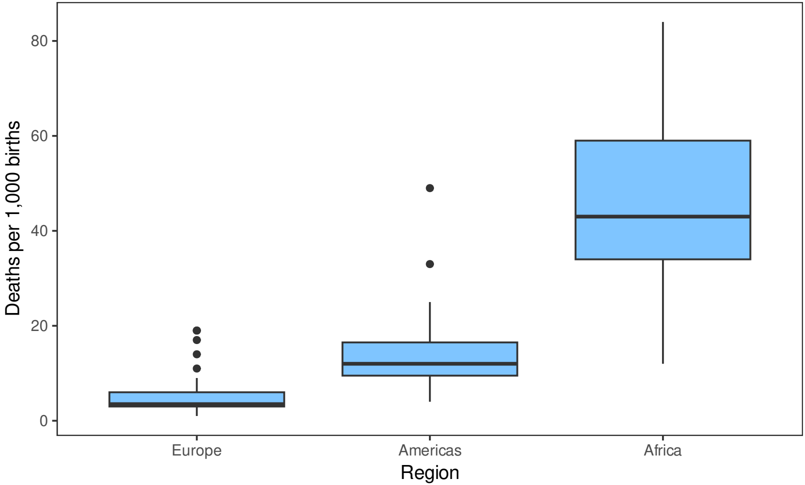
- We won't spend much time on this, but the distance from the first quartile (i.e., the 25th percentile) to the third quartile (75th percentile) is called the *interquartile range*, or IQR
- This range always contains the middle 50% of the data, and the IQR provides an alternative measure of how spread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

- In lab, we saw that bar charts provide an effective way of comparing two (or more) categorical variables (e.g., survival 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