Observational studies; descriptive statistics

Patrick Breheny

August 30

Observational studies

- We have said that randomized controlled experiments are the gold standard for determining cause-and-effect relationships in human health
- However, such experiments are not always possible, ethical, or affordable
- A much simpler, more passive approach is to simply observe people's decisions and the consequences that seem to result from them, then attempt to link the two
- Such studies are called *observational studies*

Smoking

- For example, smoking studies are observational – no one is going to take up smoking for ten years just to please a researcher
- However, the idea of treatment/exposure (smokers) and control (nonsmokers) groups is still used, just as it was in controlled experiments
- The essential difference, however, is that the subject assigns themselves to the exposure/control group – the investigators just watch
- Because of this, confounding is possible: hundreds of studies have shown that smoking is *associated* with various diseases, but none can definitively prove *causation*

Controlling for confounders

- However, just because confounding is possible in such studies does not mean that investigators are powerless to address it
- Instead, well-conducted observational studies make strong efforts to identify confounders and *control for* their effect
- There are many techniques for doing so; the most direct approach is to make comparisons separately for smaller and more homogeneous groups

Controlling for confounders (cont'd)

- For example, studying the association between heart disease and smoking could be misleading, because men are more likely to have heart disease and also more likely to smoke
- A solution is to compare heart disease rates separately: compare male smokers to male nonsmokers, and the same for females
- Age is another common confounding factor that epidemiologists are often concerned with controlling for

The value of observational studies

- Hundreds of very carefully controlled and well-conducted studies of smoking have been conducted in the past several decades
- Most people would agree that these studies make a very strong case that smoking is dangerous, and that alerting the public to this danger has saved thousands of lives
- Observational studies are clearly a very powerful and necessary tool
- Furthermore, observational studies have tremendous value as initial studies to build up support for larger, more resource-intensive controlled experiments
- However, they can be very misleading – identifying confounders is not always easy, and is sometimes more art than science

Racial bias in Florida

- A study of racial bias in the administration of the death penalty was published in the *Florida Law Review*
- The sample consists of 674 defendants convicted of multiple homicides in Florida between 1976 and 1987, classified by the defendant's and the victims' races:

	White defendants		Black defendants	
Victims' race	Total	Death penalty	Total	Death penalty
White	467	53	48	11
Black	16	0	143	4

Evidence for racial bias against whites

- From the table, the overall percentage of white defendants who received the death penalty is

$$\frac{53+0}{467+16} = 11.0\%$$

- And for black defendants,

$$\frac{11+4}{48+143} = 7.9\%$$

- This would seem to be evidence of racial bias against white defendants

Controlling for victim's race

- However, let's control for the potentially confounding effect of victim's race by calculating the percent who received the death penalty separately for white victims and black victims:

	% sentenced to death	
Victims' race	White	Black
White	11.3	22.9
Black	0.0	2.8

- This table indicates racial bias against blacks

What's going on?

- This may seem paradoxical: if blacks are more likely to receive the death penalty for white victims, and also for black victims, how can whites be more likely to receive the death penalty overall?
- The answer is that both races are much more likely to be involved in murders in which the victim is the same race as the defendant (97% of white defendants were on trial for the murder of white victims; 75% of black defendants were on trial for the murder of black victims)
- Furthermore, Florida juries were much more likely to award the death penalty in cases involving white victims (12.5%) than black victims (2.5%)
- Thus, the apparent racial bias against whites could be due to the confounding factor of the victims' race

Weighted averages

- Due to the threat of confounding in observational studies, it is often useful to obtain an overall average that has been adjusted for the confounding factor
- One such method is to calculate a *weighted average*
- In a regular average, every observation gets an equal weight of $1/n$ – an equivalent way of writing the average is

$$\bar{x} = \sum_{i=1}^{n} \frac{1}{n} x_i$$

- In a weighted average, every observation gets its own weight w_i:

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i$$

where the weights must add up to 1

Death penalty rates as weighted averages

- We can express death penalty rates as weighted averages; this allows us to separate the confounder from the outcome
- I'll use the following notation: For a given defendant race (i.e., white or black):
    - Let w_w denote the proportion on trial for the murder of a white victim
    - Let w_b denote the proportion on trial for the murder of a black victim
    - Let $\bar{x}_w$ denote the percent sentenced to death for the murder of a white victim
    - Let $\bar{x}_b$ denote the percent sentenced to death for the murder of a black victim

Death penalty rates as weighted averages (cont'd)

- White defendants:

$$\begin{aligned}\bar{x} &= w_w\bar{x}_w + w_b\bar{x}_b \\ &= (.967)11.3 + (.033)0 \\ &= 11.0\end{aligned}$$

- Black defendants:

$$\begin{aligned}\bar{x} &= w_w\bar{x}_w + w_b\bar{x}_b \\ &= (.251)22.9 + (.749)2.8 \\ &= 7.9\end{aligned}$$

- This allows us to see directly the effect of confounding: the white-victim death penalty percentage gets 97% of the weight for white defendants, but only 25% of the weight for black defendants

Average controlled for victims' race

- What would happen if these weights were the same (*i.e.* if victims' race was not a confounding factor and both races were equally likely to be on trial for the murder of a white victim)?
- Overall, 76.4% (515/674) of the victims were white and 23.6% were black; using these as weights,

$$\text{Whites:} \quad (.764)11.3 + (.236)0 = 8.6$$
$$\text{Blacks:} \quad (.764)22.9 + (.236)2.8 = 18.2$$

- By artificially forcing the distribution of victims' race to be the same for both groups, we obtain an average that is adjusted for the confounding factor of victim's race
- This allows us to isolate the effect of defendant's race upon his/her likelihood of receiving the death penalty, in the absence of the confounding effect of victim's race

Summary: Observational Studies

- Randomized controlled trials are not always possible or practical; for these reasons observational studies also play an important role in science
- Observational studies are always limited by confounding, although known confounders can be accounted for, either through design or statistical calculations
- We did a simple example with a weighted average; more sophisticated approaches to adjusting for confounders are discussed in Biostatistical Methods II (BIOS 5720)

Descriptive statistics

- Switching gears now, the rest of the lecture will deal with descriptive statistics
- 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 occurs throughout all of statistics

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 three basic summary statistics for categorical data, and are often displayed in tables or bar charts

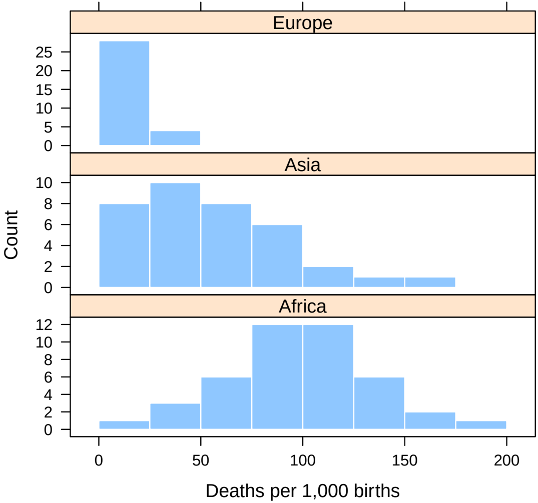
Continuous data

- 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 for 111 nations on three continents: Africa, Asia, and Europe

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

Histogram of infant mortality rates

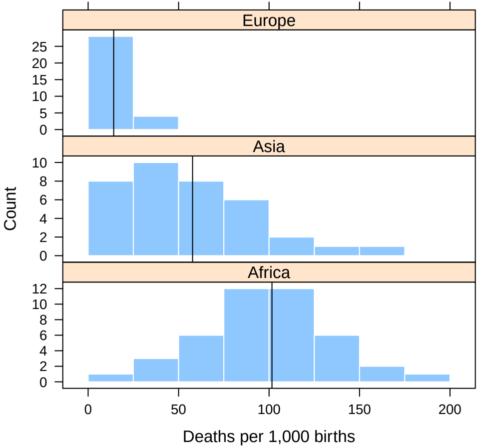


Summarizing continuous data

- 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 summarized 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 typically summarized by the *standard deviation*:

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

- The root-mean-square (RMS) is the most natural way of measuring the average size of an n-dimensional object
- The standard deviation is essentially the RMS of the deviations, except it has an $n-1$ in the denominator instead of n

Why $n-1$

- Why $n-1$ instead of n?
- The reason has to do with the *variance*, which is simply s^2
- We will return to this point in a few weeks, but it turns out that the "natural" estimator

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

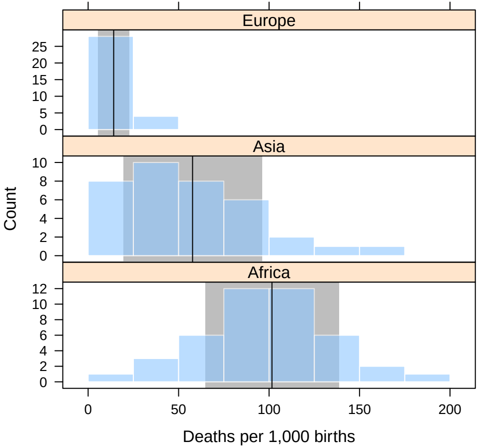
  systematically underestimates the true variance (i.e., it is biased); dividing by $n-1$ corrects this bias
- It is worth noting, however, that s is still biased for the true standard deviation

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:

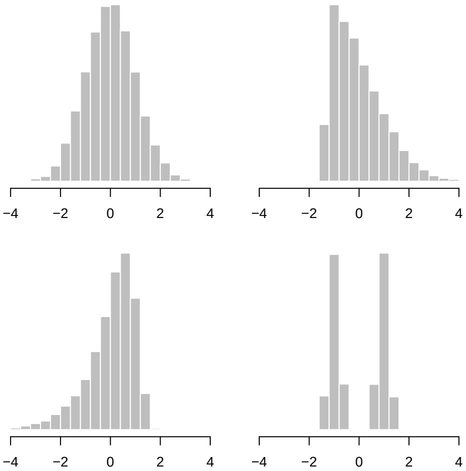


The 68%/95% rule in action

	% of observations within	
Continent	One SD	Two SDs
Europe	78	97
Asia	67	97
Africa	63	95

Summaries can be misleading!

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

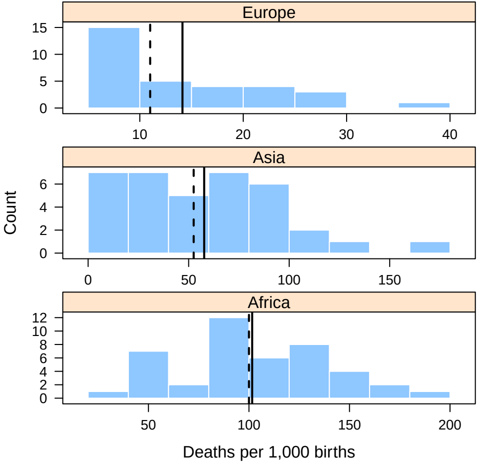


Percentiles

- 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

- Note that the histogram for Europe 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

Interquartile range

- Percentiles can also be used to summarize spread
- A common percentile-based measure is the *interquartile range* (IQR), defined as the difference between the 75th percentile (3rd quartile) and the 25th percentile (1st quartile)
- By construction, the IQR contains the middle 50% of the data

Robustness

- Azerbaijan had the highest infant mortality rate in Europe at 37
- What if, instead of 37, it was 200?

	Mean	Median	SD	IQR
Real	14.1	11	8.7	13.2
Hypothetical	19.2	11	33.8	13.2

- Note that the mean is sensitive to extreme values and the standard deviation is even more sensitive
- In comparison, the median and IQR are not; these statistics are *robust* to the presence of outlying observations

Five number summary

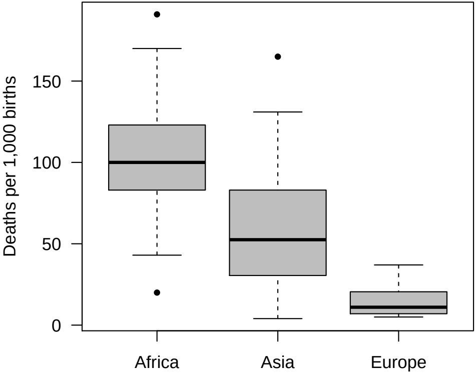
- The mean and standard deviation are a common way of providing a two-number summary of a distribution of continuous values
- Another approach, based on quantiles, is to provide a "five-number summary" consisting of: (1) the minimum, (2) the first quartile, (3) the median, (4) the third quartile, and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

	Europe	Asia	Africa
Min	5	4	20
First quartile	7	32	83
Median	11	52.5	100
Third quartile	20	83	123
Max	37	165	191

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 more than 1.5$\times$IQR away from the box an "outlier" and plot the observations using a special symbol (the 1.5 is customary but arbitrary and can be modified)
  - 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



One big advantage of box plots (compared to histograms) is the each with which they can be placed next to each other

Summary: Descriptive statistics

- 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 are superior for illustrating trends in the data
- Summary measures for categorical variables: counts, percents, rates
- Summary measures for continuous variables: mean, standard deviation, quantiles
- Ways to display continuous data: histogram, box plot