Stat 22000 Lecture Slides
Exploring Numerical Data

Yibi Huang
Department of Statistics
University of Chicago

Outline

In this slide, we cover mostly Section 1.2 & 1.6 in the text.

- Data and Types of Variables (1.2)
- Histograms (1.6.3)
- Mean and Median (1.6.2 & 1.6.6)
- Five-Number-Summary and Box Plots (1.6.5)
- Standard Deviation (1.6.4 & 3.1.5)
- Scatterplots (1.6.1)
- Transforming Data (1.6.7)

Please skip 1.6.8 (Mapping data).

Data basics

Data: Cases & Variables

- In a study, we collect information — data — from **cases**. **Cases** can be individuals, corporations, animals, or any objects of interest.

- A **variable** is a characteristic of a case. A variable varies among cases.
    - E.g., age, blood pressure, leaf length, first language

Data matrix

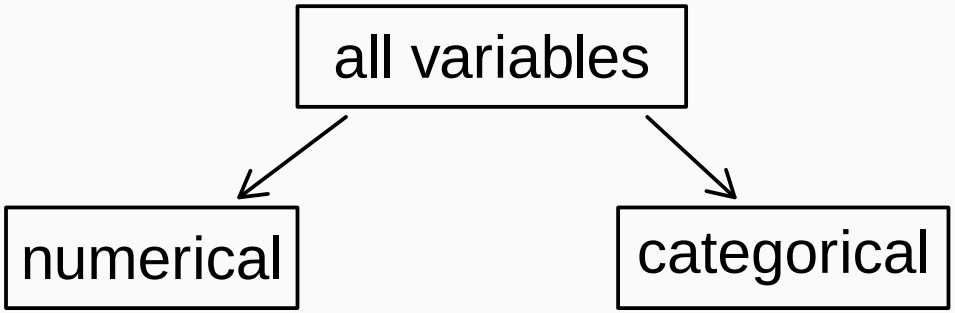
Example: Data Set `email50`

variable ↓

	spam	num_char	line_breaks	format	number	
1	no	21,705	551	html	small	
2	no	7,011	183	html	big	
3	yes	631	28	text	none	← *case*
⋮	⋮	⋮	⋮	⋮	⋮	
50	no	15,829	242	html	small	

Each row of data corresponds to a **case** (an email), and each column contains the values of one **variable** of all observations.

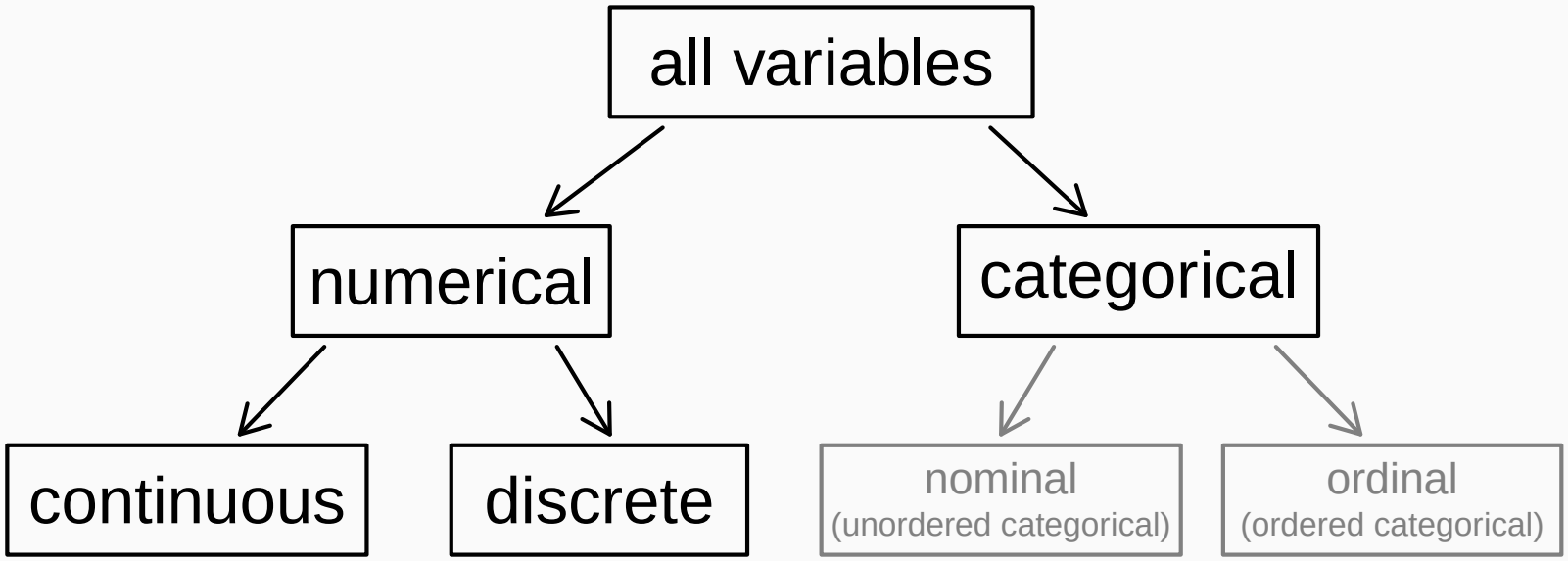
Types of variables



A variable is *numerical* when it can take a wide range of numerical values, and it is sensible to take arithmetic operations (addition, subtraction, average) with those values. Otherwise, it is *categorical*.

- e.g., Zip codes, area codes are NOT numerical variables

Types of variables



A numerical variable is

- *discrete* if its possible values form a set of separate numbers, such as 0, 1, 2, 3, ...
- *continuous* if its possible values form an interval.

A categorical variable with ordered categories is *ordinal*.

Types of variables (cont.)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

- spam
- num_char
- line_breaks
- format
- number

Types of variables (cont.)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

- `spam` ..... *categorical*
- `num_char`
- `line_breaks`
- `format`
- `number`

Types of variables (cont.)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

- spam ..... *categorical*
- num_char ..... *numerical*
- line_breaks
- format
- number

Types of variables (cont.)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

- `spam` ..... *categorical*
- `num_char` ..... *numerical*
- `line_breaks` ..... *numerical*
- `format`
- `number`

Types of variables (cont.)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

- `spam` ..... *categorical*
- `num_char` ..... *numerical*
- `line_breaks` ..... *numerical*
- `format` ..... *categorical, nominal*
- `number`

Types of variables (cont.)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

- `spam` ..... *categorical*
- `num_char` ..... *numerical*
- `line_breaks` ..... *numerical*
- `format` ..... *categorical, nominal*
- `number` ..... *categorical, ordinal*

Sometimes a ordinal categorical variable can also be regarded as a numerical variable, e.g.,

- rating of a movie from 1 star to 5 stars
- stage of cancer from 0 to 4

Histograms

How to Make Histograms

Data: Infant mortality rates (number of deaths under one year of age per 1000 live births) of 201 countries/regions in 2010-2015 (Data file: `infmort2015.txt`):

```
    Country.Region Continent X2010.2015
1          Burundi    Africa       77.9
2          Comoros    Africa       58.1
3         Djibouti    Africa       55.3
...
200          Samoa   Oceania       19.7
201          Tonga   Oceania       20.4
```

Step 1: Divide the range of values into *class intervals*.

Step 2: Count the number of values in each class interval.

Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Count	75	38	18	16	18	11	10	9	1	5

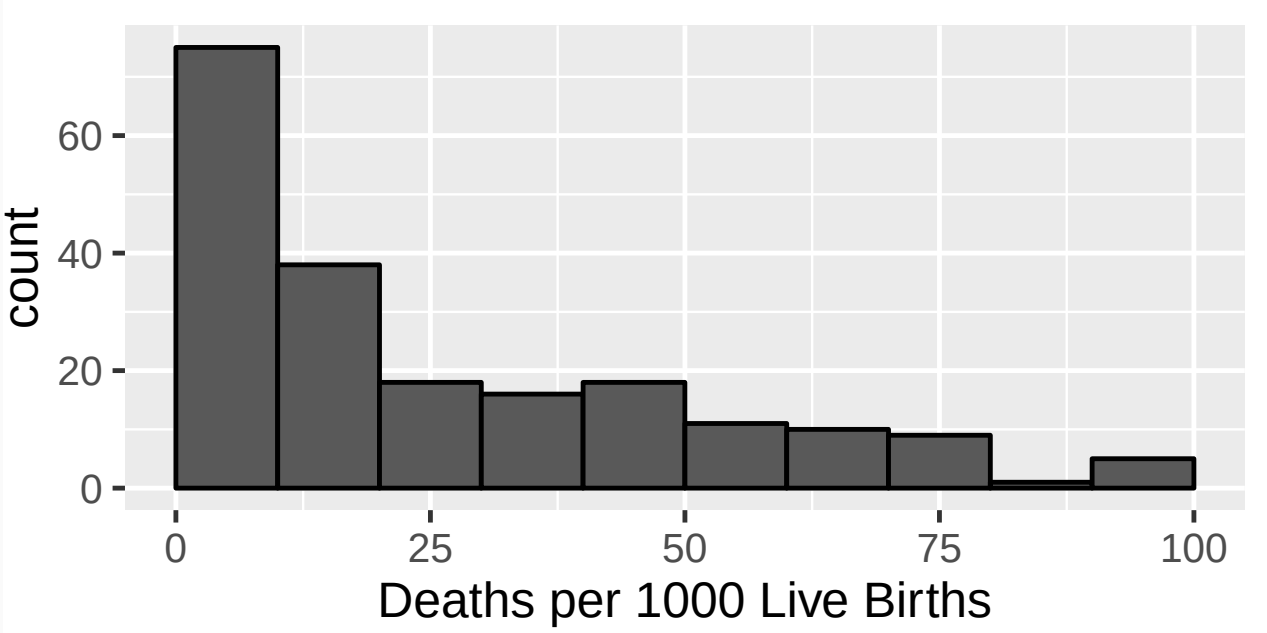
https://en.wikipedia.org/wiki/List_of_countries_by_infant_mortality_rate

How to Make Histograms (Cont'd)

Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Count	75	38	18	16	18	11	10	9	1	5

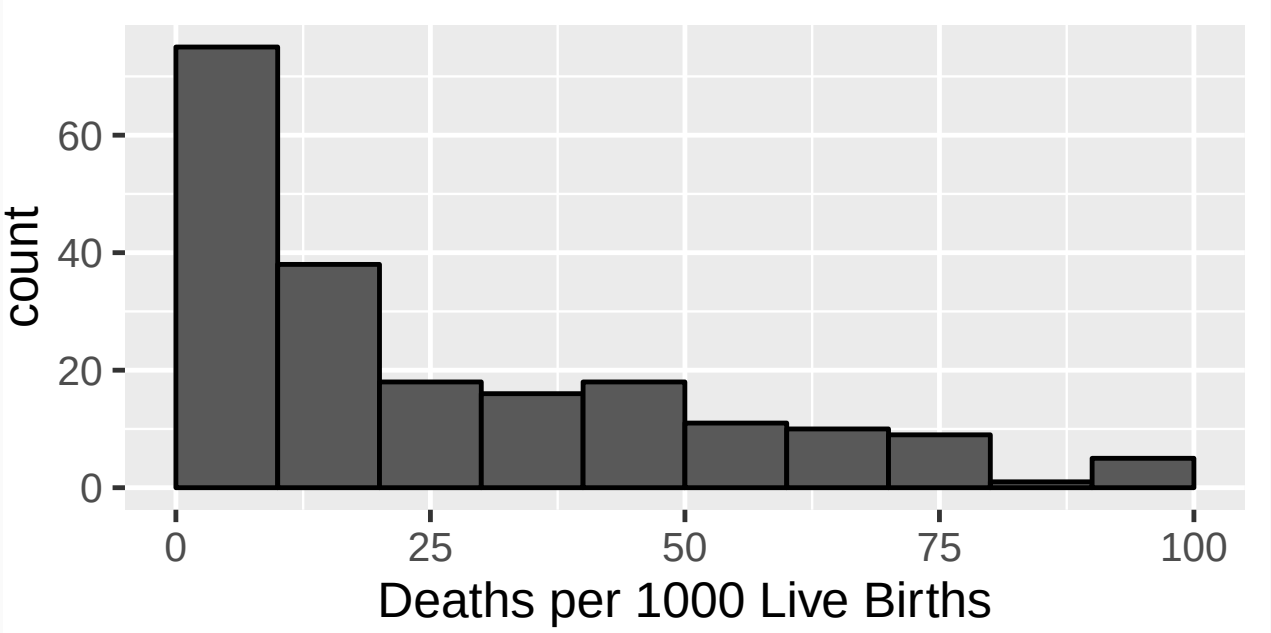
Step 3: Draw the histogram

- No space between bars.
- Label the horizontal axes (with units)!



Histograms

- Histograms provide a view of the *data density*. Higher bars indicate regions with more observations.
- Histograms are especially convenient for describing the *shape* of the data distribution.
- The selection of *bin width* can alter the shape of histogram

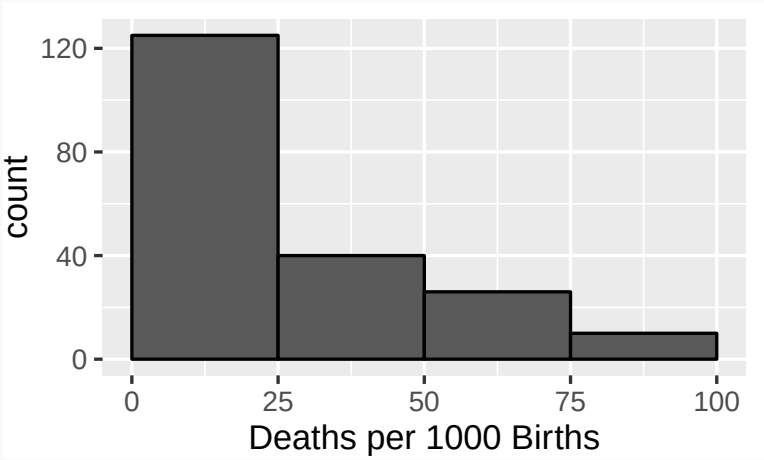


Try Different Binwidth

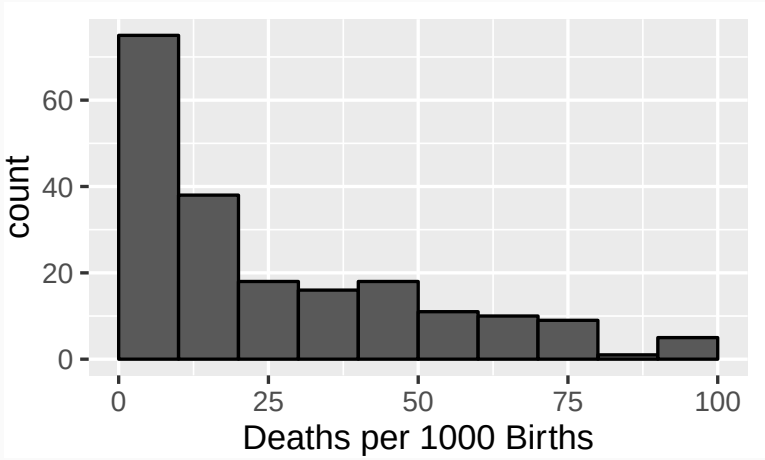
Which one(s) of these histograms reveal too much about the data?

Which hide too much?

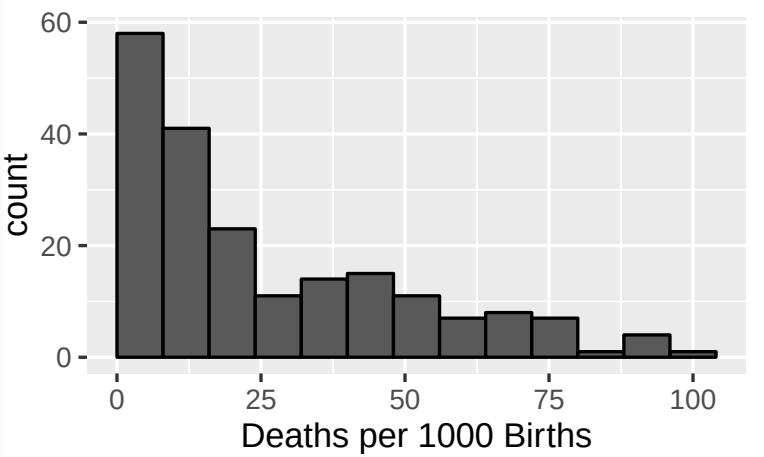
Binwidth = 25



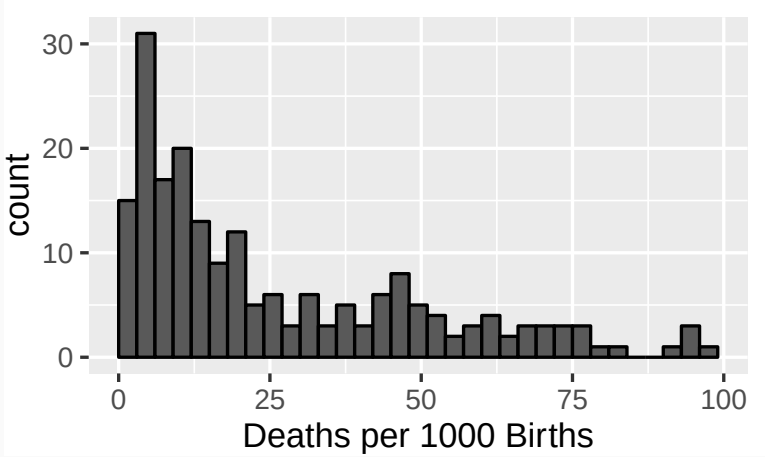
Binwidth = 10



Binwidth = 8



Binwidth = 3



Selection of Binwidth

It is an iterative process — try and try again.

What bin with should you use?

- Not too small that most bins have either 0 or 1 counts
- Not too big that you lose the details in a bin
- (There may not be a unique “perfect” bin size)

General rule: **the more observations, the more bins.**

What to Look in a Histogram?

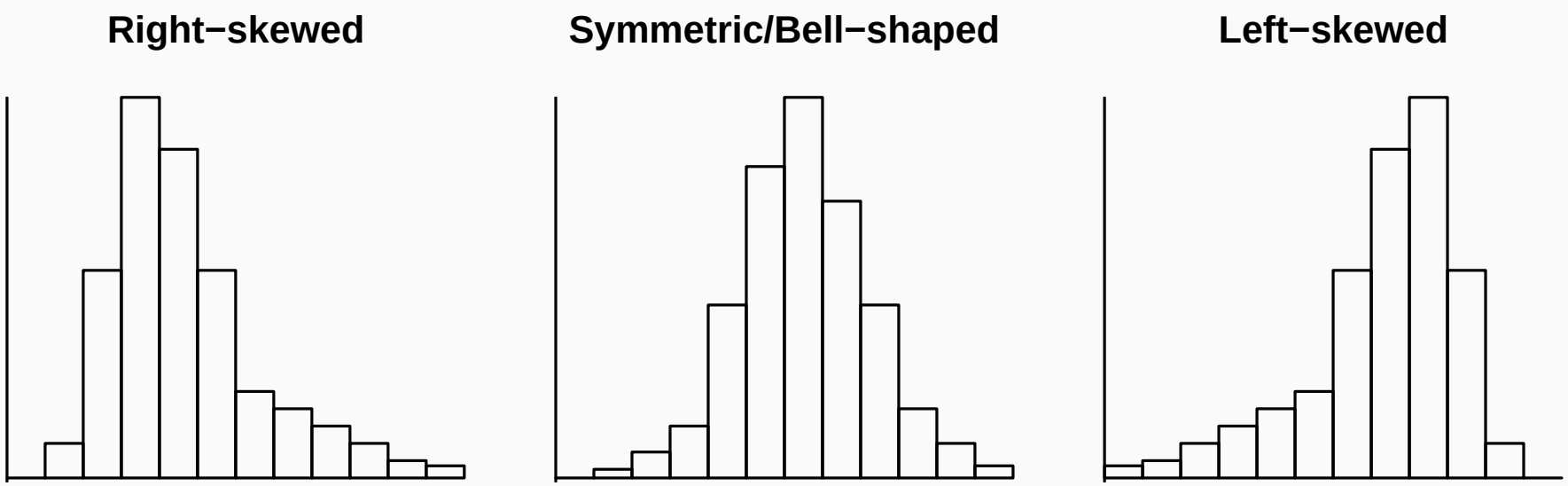
- **Shape**
  - symmetric or skewed (lopsided)
  - number of modes (peaks)

- **Outliers**: Are there any observations that lie outside the overall pattern? They could be unusual observations, or they could be mistakes. Check them!

- **Center**: Where is the "middle" of the histogram?
  - typically represented by *mean* and *median*

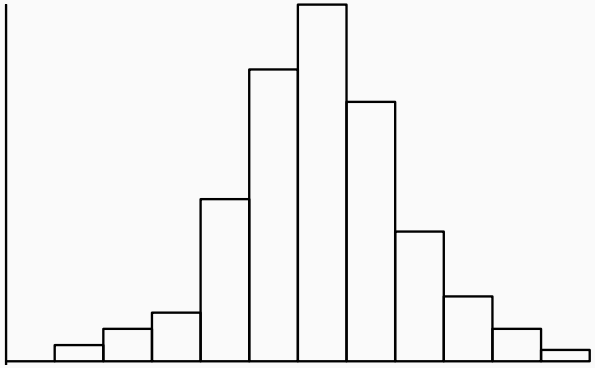
- **Spread**: What is the range of data?
  - typically represented by *SD* and *IQR* (will introduce soon)

Skewness of Histograms

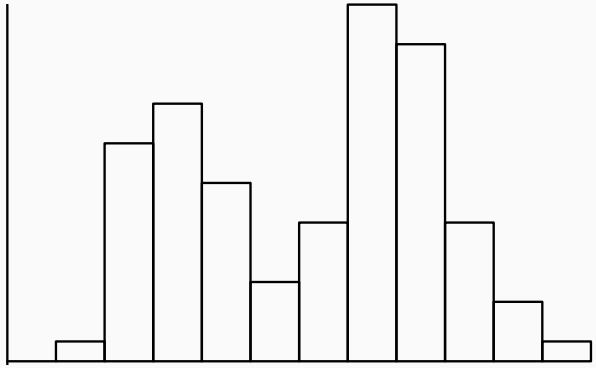


Mode of Histograms (= Number of Peaks)

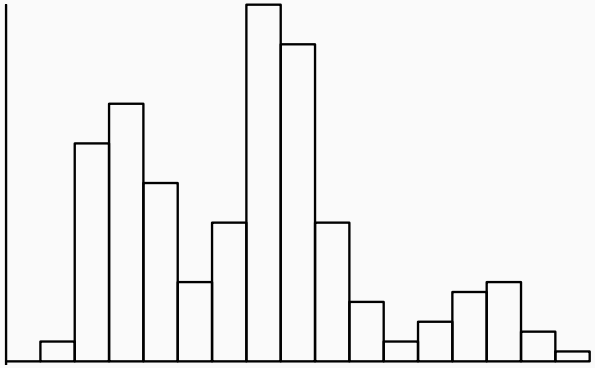
Unimodal



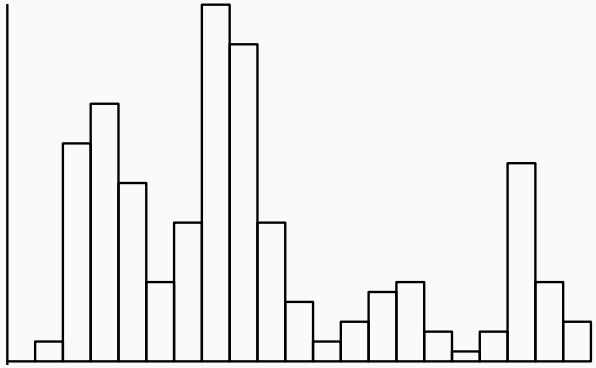
Bimodal



Trimodal



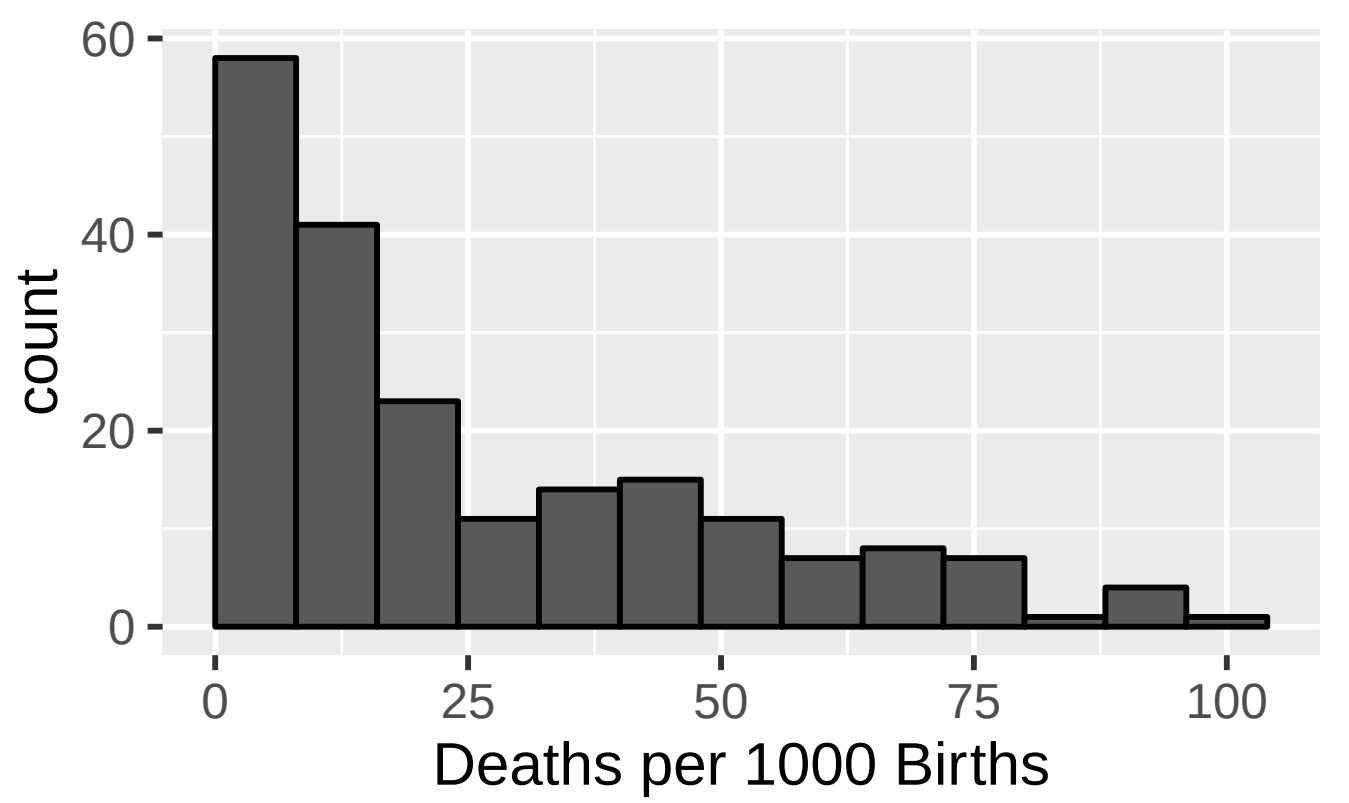
Multimodal



A histogram with two or more modes may indicate that the data is a mixture of two or more distinct populations.

Example (Infant Mortality Rates)

In addition to the major peak near 0, there appears to be a secondary peak around 40-50.

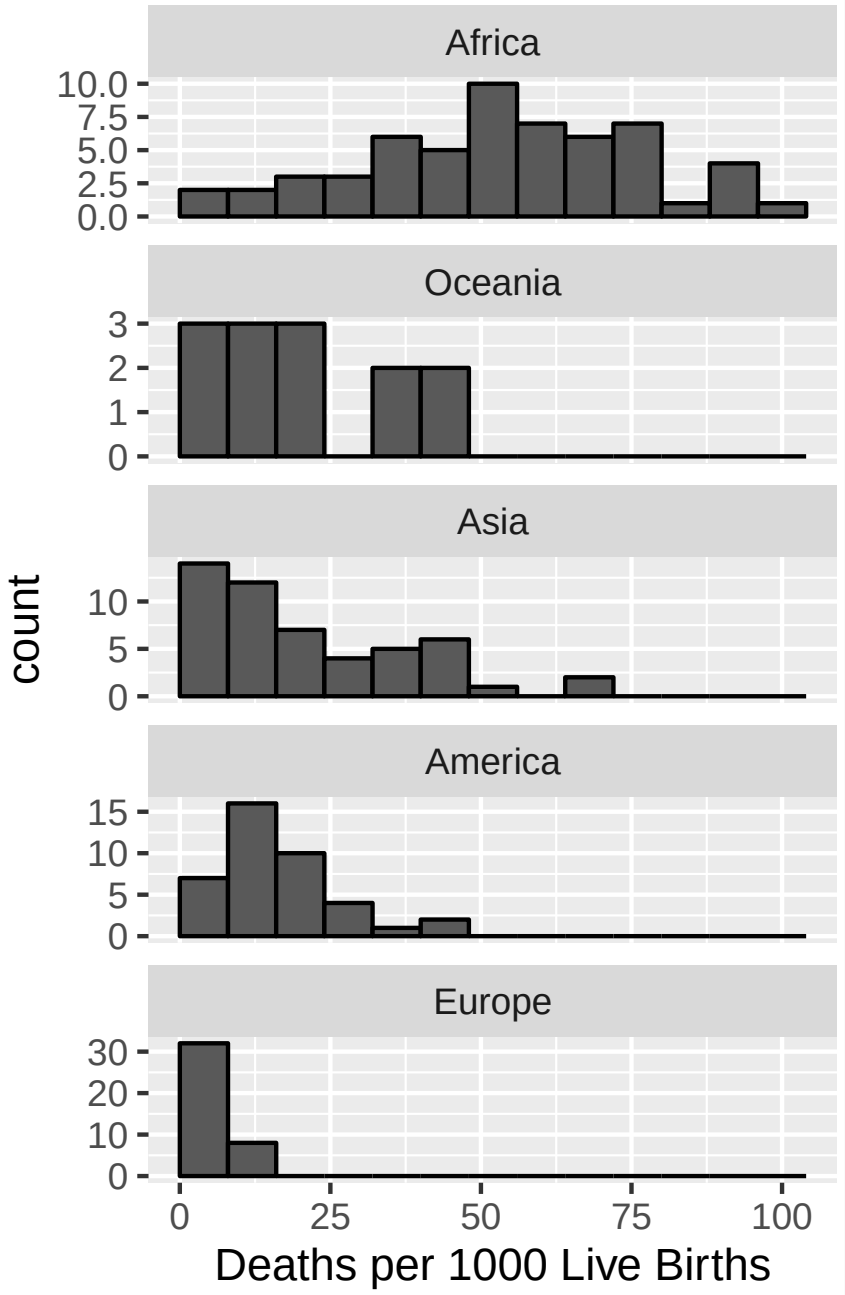


Side-by-Side Histograms — Infant Mortality Rate Data

If countries are grouped by continent and histograms are made separately on the same horizontal axis, we can compare the infant mortality rates of countries in the 5 continents by the location of the histograms, which were

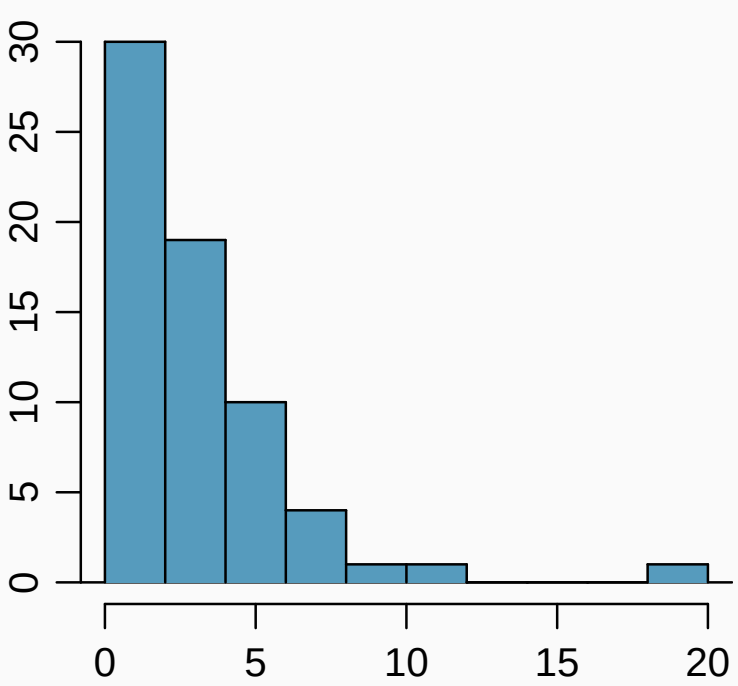
- uniformly low in Europe
- much higher and with greater variability in Africa.

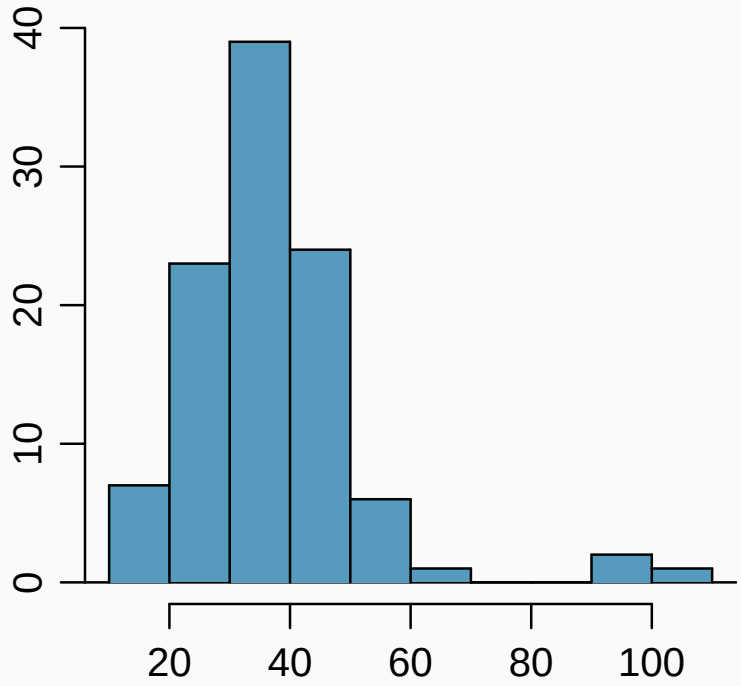
This explains why the histogram for the whole world to be bimodal.



Outliers

Another thing we look at a histogram is whether there are any unusual observations or potential *outliers*?





Practice: Shapes of Distributions

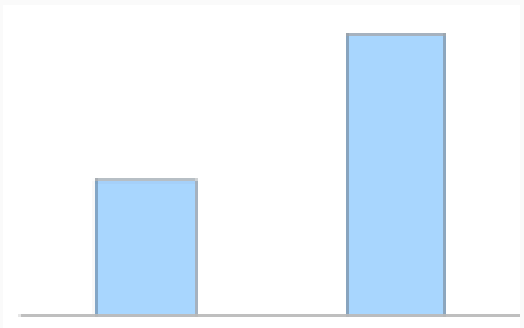
Match the following variables with the histograms and bar graphs given below. Suppose the data represent STAT 220 students.

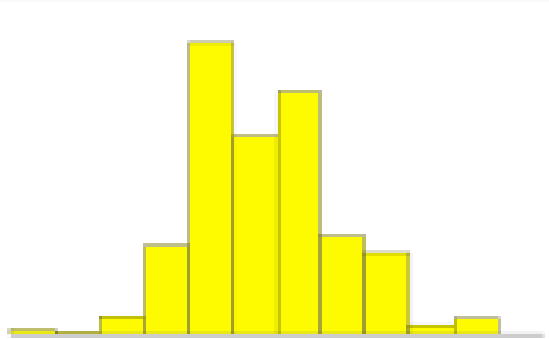
[Hint: Think about how each variable should behave.]

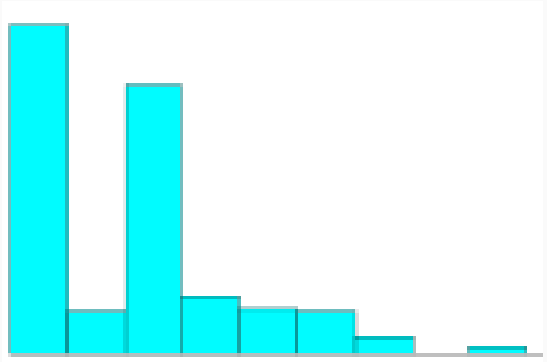
(a) the height of students

(b) gender breakdown of students

(c) the number of piercings students have







Mean and median

Mean

The *mean* of a numerical variable is computed as the sum of all of the observations divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

Example. Suppose a variable has 5 observed values:

$$4, 8, 3, 5, 13.$$

The mean of the variable is given by:

$$\bar{x} = \frac{4 + 8 + 3 + 5 + 13}{5} = \frac{33}{5} = 6.6.$$

Median

The **median** of a numerical variable is a number such that half of the observed value are smaller than it and half are larger than it.

Ex 1: Suppose a variable has 5 observed values: 4, 8, 3, 5, 13.

data	$\rightarrow$	4	8	3	5	13
sorted	$\rightarrow$	3	4	5	8	13

The median is 5.

Ex 2: Suppose a variable has 5 observed values: 4, 8, 3, 5, 13, 12.

data	$\rightarrow$	4	8	3	5	13	12
sorted	$\rightarrow$	3	4	5	8	12	13

The median is thus $= \dfrac{5+8}{2} = 6.5$.

Median

The **median** of a numerical variable is a number such that half of the observed value are smaller than it and half are larger than it.

Ex 1: Suppose a variable has 5 observed values: 4, 8, 3, 5, 13.

data	$\rightarrow$	4	8	3	5	13
sorted	$\rightarrow$	3	4	5	8	13

The median is 5.

Ex 2: Suppose a variable has 5 observed values: 4, 8, 3, 5, 13, 12.

data	$\rightarrow$	4	8	3	5	13	12
sorted	$\rightarrow$	3	4	5	8	12	13

The median is thus $= \dfrac{5+8}{2} = 6.5$.

Median

The **median** of a numerical variable is a number such that half of the observed value are smaller than it and half are larger than it.

Ex 1: Suppose a variable has 5 observed values: 4, 8, 3, 5, 13.

data	$\rightarrow$	4	8	3	5	13
sorted	$\rightarrow$	3	4	5	8	13

The median is 5.

Ex 2: Suppose a variable has 5 observed values: 4, 8, 3, 5, 13, 12.

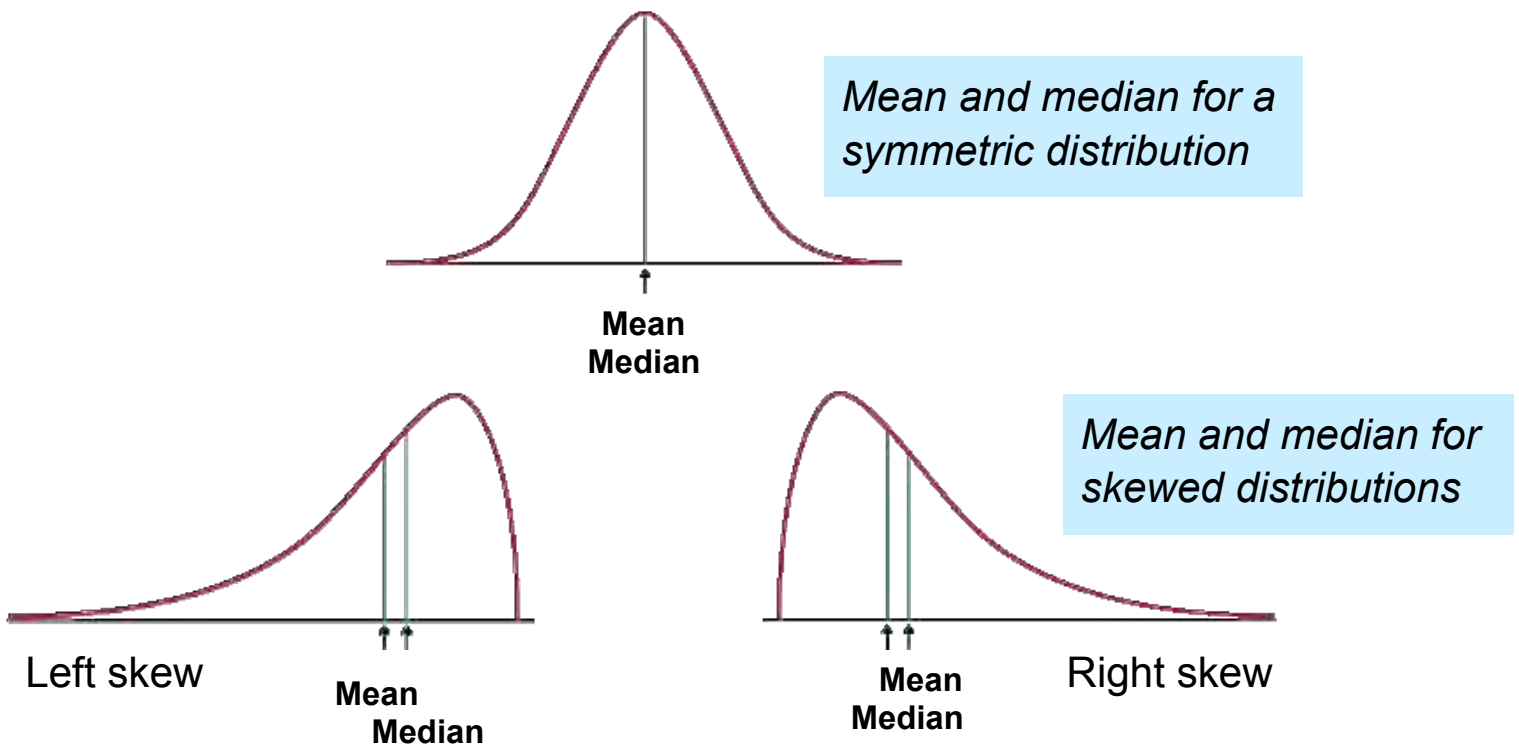
data	$\rightarrow$	4	8	3	5	13	12
sorted	$\rightarrow$	3	4	5	8	12	13

The median is thus $= \dfrac{5+8}{2} = 6.5$.

Mean vs. Median

- In a symmetric distribution, mean $\approx$ median.

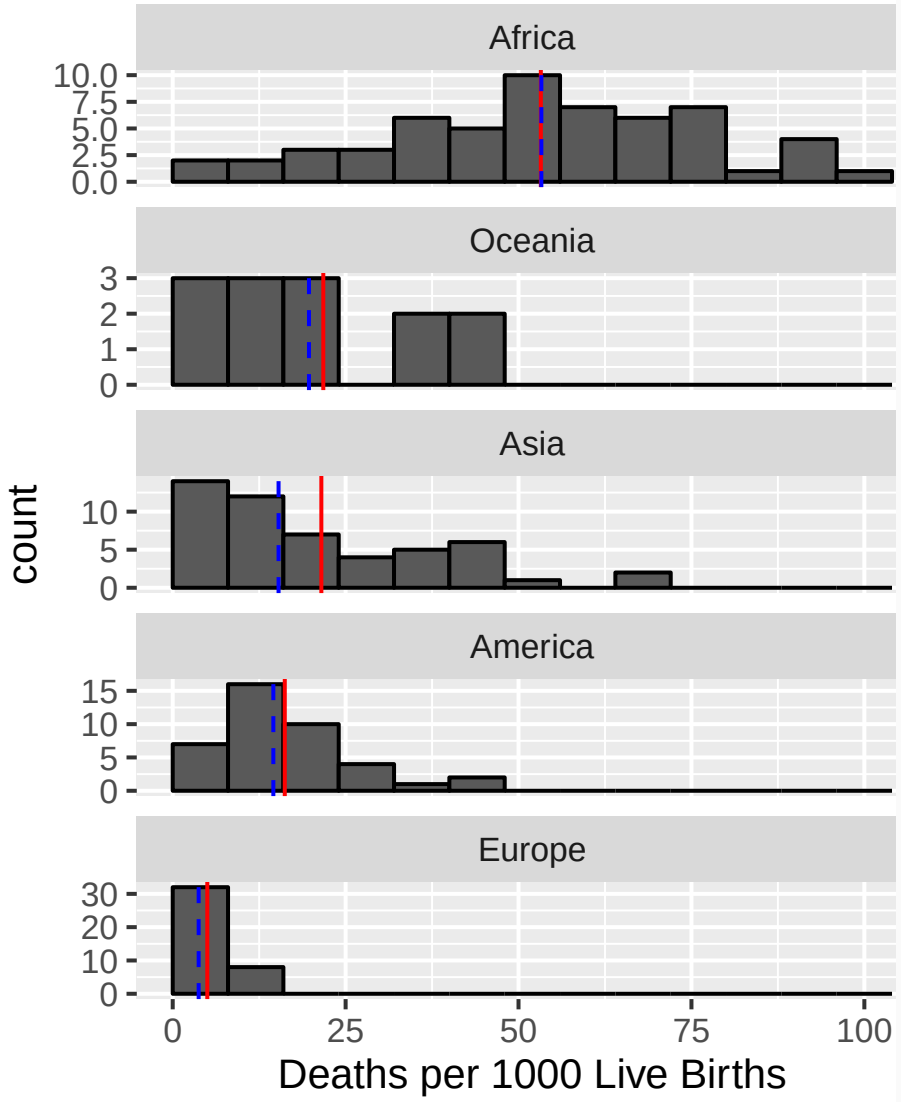
  If exactly symmetric, then mean = median.
- In a skewed distribution, the mean is pulled toward the longer tail.



Example (Infant Mortality Rates)

Which of the 5 histograms are symmetric? Which are skewed? How are their means compared with their medians?

Continent	Mean	Median
Africa	53.22	53.30
Oceania	21.77	19.70
Asia	21.49	15.30
America	16.20	14.55
Europe	5.00	3.75



(red solid — mean, blue dash — median)

Robustness of the Median

Consider the data w/ six observations: -2, -1, 0, 0, 2, 4. If the number '2' in the data is wrongly recorded as 20,

- The mean is increased by $\frac{20-2}{6} = 3$.
- The median is unaffected.

Median is more resistent, i.e., less sensitive to extreme values or outliers than the mean. We say the median is more **robust**.

- Example: Housing sales price in Hyde Park

	Mean	Median
Jun – Aug, 2011	\$525,384	\$227,000
Jun – Aug, 2013	\$423,528	\$291,750
May – Aug, 2017	\$259,542	\$226,750

Source: `http://www.trulia.com/home_prices/Illinois/Chicago-heat_map/`

Five Number Summary

Quartiles, IQR, Five-Number Summary

- **Quartiles** divide data into 4 even parts
  - **first quartile $\mathbf{Q_1}$** = 25th percentile:
    25% of data fall below it and 75% above it
  - **second quartile $\mathbf{Q_2}$** = median = 50th percentile
  - **third quartile $\mathbf{Q_3}$** = 75th percentile
    75% of data fall below it and 25% above it
- **Interquartile Range** $(IQR) = Q_3 - Q_1$
- **Five-Number Summary**:

$$\min, \quad Q_1, \quad \text{Median}, \quad Q_3, \quad \max$$

Example 1

For the 9 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34

$$\begin{array}{cc} 43 & 27 \\ 35 & 33 \\ 43 & 34 \\ 33 & 35 \\ 38 & \xrightarrow{\text{sort}} \quad 38 \\ 53 & 43 \\ 64 & 43 \\ 27 & 53 \\ 34 & 64 \end{array}$$

- IQR $= Q_3 - Q_1 = 48 - 33.5 = 14.5$
- Five number summary: 27, 33.5, 38, 48, 64

Example 1

For the 9 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34

43		27		
35		33		
43		34		
33		35		
38	$\xrightarrow{\text{sort}}$	38	$\leftarrow$ overall median	$= Q_2$
53		43		
64		43		
27		53		
34		64		

- IQR $= Q_3 - Q_1 = 48 - 33.5 = 14.5$
- Five number summary: 27, 33.5, 38, 48, 64

Example 1

For the 9 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34

43		27
35		33
43		34
33		35
38	sort $\longrightarrow$	38
53		43
64		43
27		53
34		64

$\leftarrow$ median of this half $= \frac{33 + 34}{2} = 33.5 = Q_1$

$\leftarrow$ overall median $= Q_2$

- IQR $= Q_3 - Q_1 = 48 - 33.5 = 14.5$
- Five number summary: 27, 33.5, 38, 48, 64

Example 1

For the 9 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34

43		27	
35		33	
43		34	← median of this half $= \frac{33+34}{2} = 33.5 = Q_1$
33		35	
38	sort →	38	← overall median $= Q_2$
53		43	
64		43	
27		53	← median of this half $= \frac{43+53}{2} = 48 = Q_3$
34		64	

- IQR $= Q_3 - Q_1 = 48 - 33.5 = 14.5$
- Five number summary: 27, 33.5, 38, 48, 64

Example 2

For the 10 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34, 27

$$\begin{array}{cc} 43 & 27 \\ 35 & 27 \\ 43 & 33 \\ 33 & 34 \\ 38 & 35 \\ 53 & 38 \\ 64 & 43 \\ 27 & 43 \\ 34 & 53 \\ 27 & 64 \end{array} \quad \xrightarrow{\text{sort}}$$

- IQR $= Q_3 - Q_1 = 43 - 33 = 10$
- Five number summary: 27, 33, 36.5, 43, 64

Example 2

For the 10 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34, 27

$$\begin{array}{cc} 43 & 27 \\ 35 & 27 \\ 43 & 33 \\ 33 & 34 \\ 38 & 35 \\ 53 & 38 \\ 64 & 43 \\ 27 & 43 \\ 34 & 53 \\ 27 & 64 \end{array} \xrightarrow{\text{sort}} \quad \leftarrow \text{overall median} = \frac{35+38}{2} = 36.5 = Q_2$$

- IQR $= Q_3 - Q_1 = 43 - 33 = 10$
- Five number summary: 27, 33, 36.5, 43, 64

Example 2

For the 10 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34, 27

43		27
35		27
43		33
33		34
38	sort $\longrightarrow$	35
53		38
64		43
27		43
34		53
27		64

$\leftarrow$ median of this half $= 33 = Q_1$

$\leftarrow$ overall median $= \dfrac{35 + 38}{2} = 36.5 = Q_2$

- IQR $= Q_3 - Q_1 = 43 - 33 = 10$
- Five number summary: 27, 33, 36.5, 43, 64

Example 2

For the 10 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34, 27

43		27	
35		27	
43		33	← median of this half $= 33 = Q_1$
33		34	
38	sort →	35	← overall median $= \frac{35+38}{2} = 36.5 = Q_2$
53		38	
64		43	
27		43	← median of this half $= 43 = Q_3$
34		53	
27		64	

- IQR $= Q_3 - Q_1 = 43 - 33 = 10$
- Five number summary: 27, 33, 36.5, 43, 64

Calculation of Quartiles (1)

In fact, statisticians don't have a consensus on the calculation of quartiles. There are several formulas for quartiles, varying from book to book, software to software.

E.g., for the 9 numbers in Example 1

```
> x = c(43,35,43,33,38,53,64,27,34)
```

our computation gives $Q_1 = 33.5$, $Q_3 = 48$, but R gives $Q_1 = 34$, $Q_3 = 43$.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   34.00   38.00   41.11   43.00   64.00
> fivenum(x)
[1] 27 34 38 43 64
> IQR(x)
[1] 9
```

Calculation of Quartiles (2)

Sometimes even different commands in R give different quartiles.

E.g., for the 10 numbers in Example 2,

```
> y = c(43,35,43,33,38,53,64,27,34,27)
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   33.25   36.50   39.70   43.00   64.00
> fivenum(y)
[1] 27.0 33.0 36.5 43.0 64.0
> IQR(y)
[1] 9.75
```

Don't worry about the formula. Just keep in mind that

quartiles divide data into 4 even parts

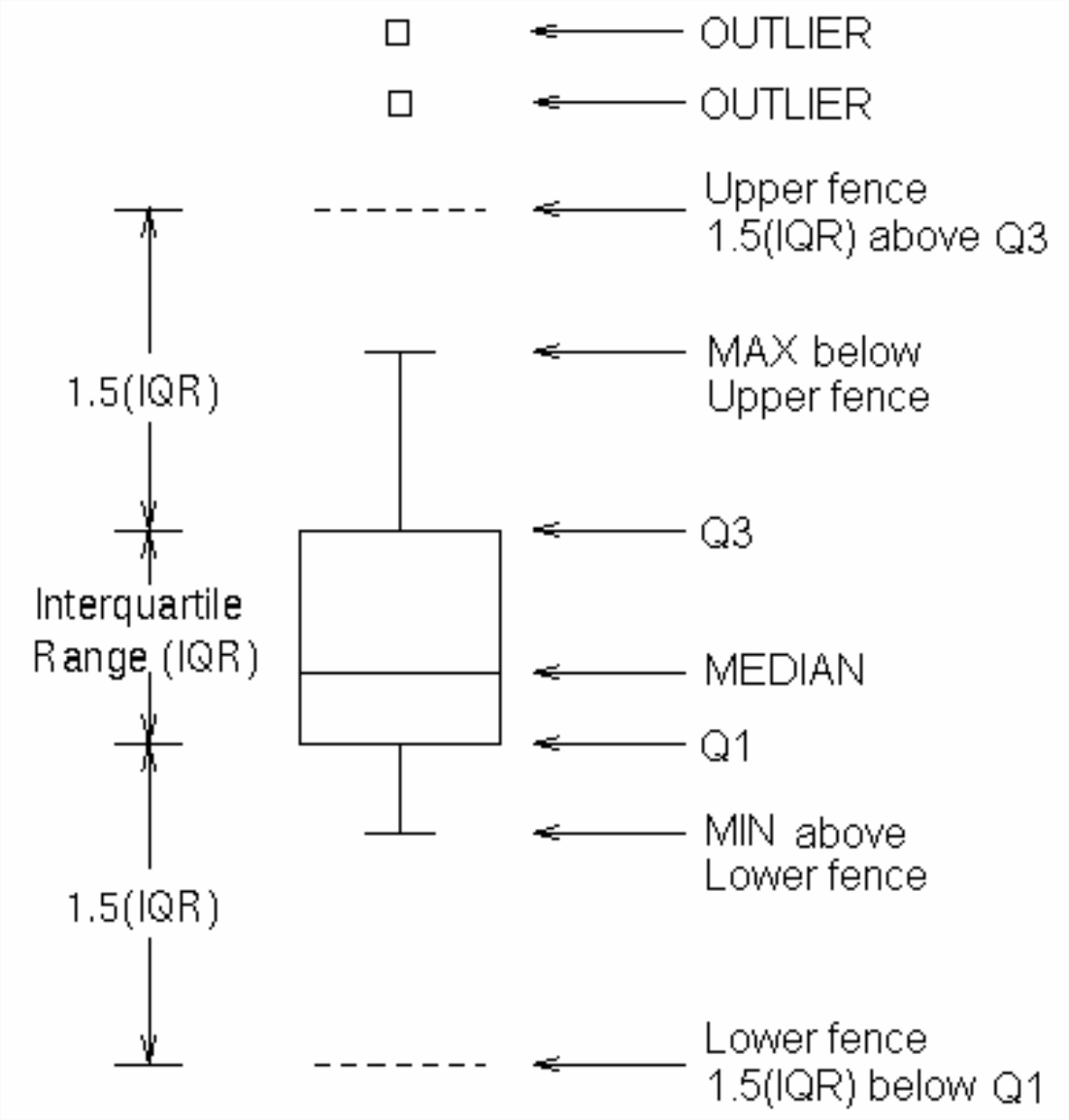
In HWs, just report whatever values your software gives.

Boxplots

1.5 IQR Rule for Identifying Potential Outliers

The 1.5 IQR Rule tags an observation as a *potential outlier* if it lies more than $1.5 \times$ IQR below Q_1 above above Q_3.

Box-and-Whiskers Plot (also called Boxplot)

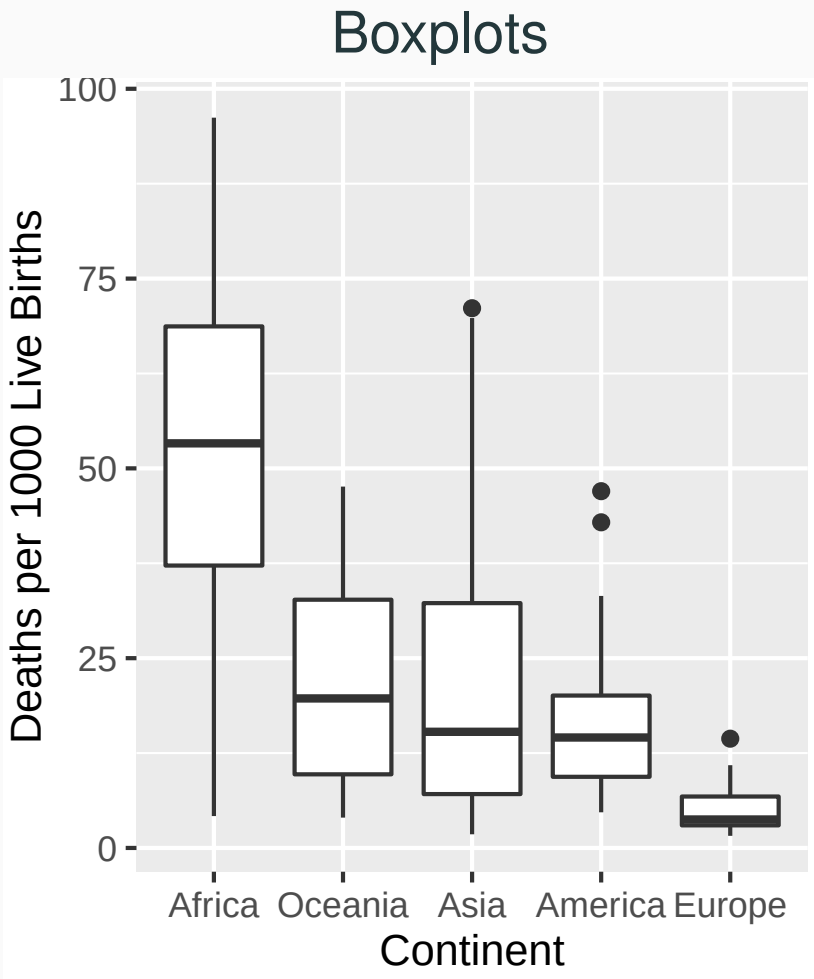


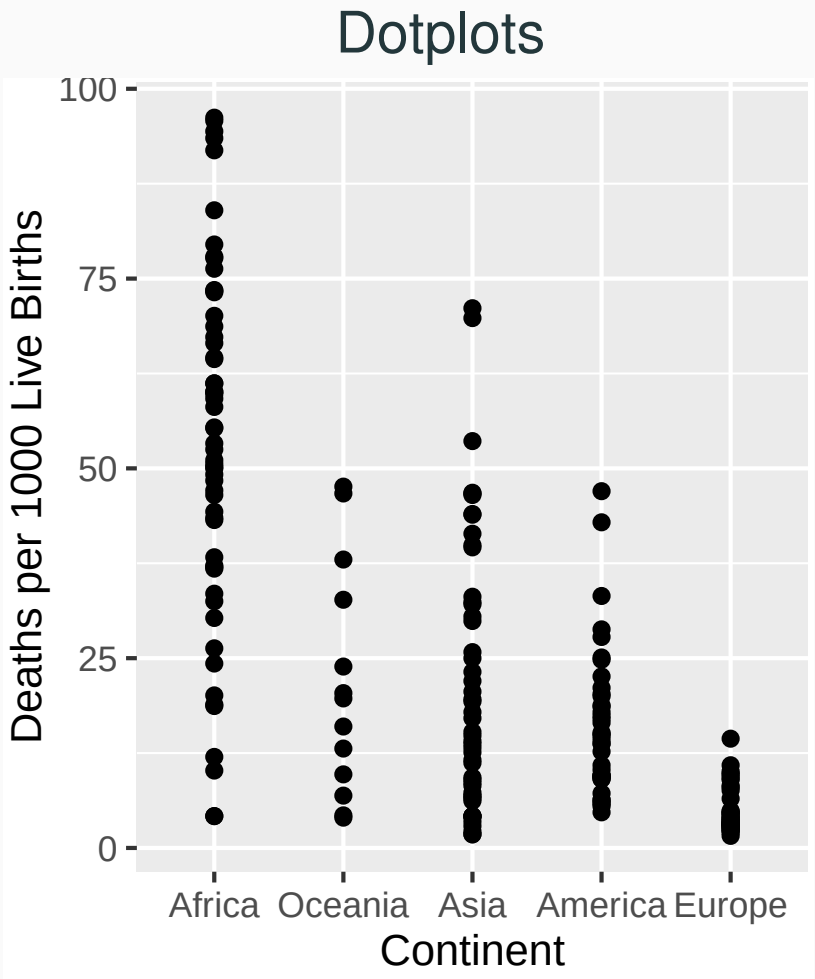
Think About It ...

- What does a boxplot look like if the distribution is symmetric?
- Ditto, if right-skewed?
- Can you tell from a boxplot whether the distribution is unimodal or bimodal?

Side by Side Boxplots

Just like histograms, boxplots of related distributions are often placed side-by-side for comparison.





Standard Deviation

Standard Deviation

Another common way to describe how spread out are the observations is the *standard deviation (SD)*.

To understand how SD works, let's use a small data set $\{1, 2, 2, 7\}$ as an example.

- Each of these numbers deviates from the mean $\frac{1+2+2+7}{4} = 3$ by some amount:

$$1 - 3 = -2, \quad 2 - 3 = -1,$$
$$2 - 3 = -1, \quad 7 - 3 = 4.$$

- How should we measure the overall size of these deviations?
- Taking their mean isn't going to tell us anything (why not?)

Standard Deviation (Cont'd)

- One sensible way is take the average of their absolute values:

$$\frac{|-2| + |-1| + |-1| + |4|}{4} = 2$$

  This is called the mean absolute deviation (MAD), not the SD.

- But for a variety of reasons, statisticians prefer using the *root-mean-square* as a measure of overall size:

$$\sqrt{\frac{(-2)^2 + (-1)^2 + (-1)^2 + 4^2}{4}} \approx 2.35$$

  but this is still not the (sample) SD.

Standard Deviation (Cont'd)

The formula for the (sample) *standard deviation (SD)* is

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

- Wait a minute; why divide by $n-1$? Not n?
- The reason (which we will discuss further in a few weeks) is that dividing by n turns out to underestimate the true (population) standard deviation. Dividing by $n-1$ instead of n corrects some of that bias
- The standard deviation of $\{1, 2, 2, 7\}$ is

$$\sqrt{\frac{(-2)^2 + (-1)^2 + (-1)^2 + 4^2}{4-1}} \approx 2.71$$

(recall we get 2.35 when divided by $n = 4$)

Variance

The square of the (sample) standard deviation is called the *(sample) variance*, denoted as

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

which is roughly the average squared deviation from the mean.

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.”

The 68% and 95% Rule (Section 3.1.5 in Text)

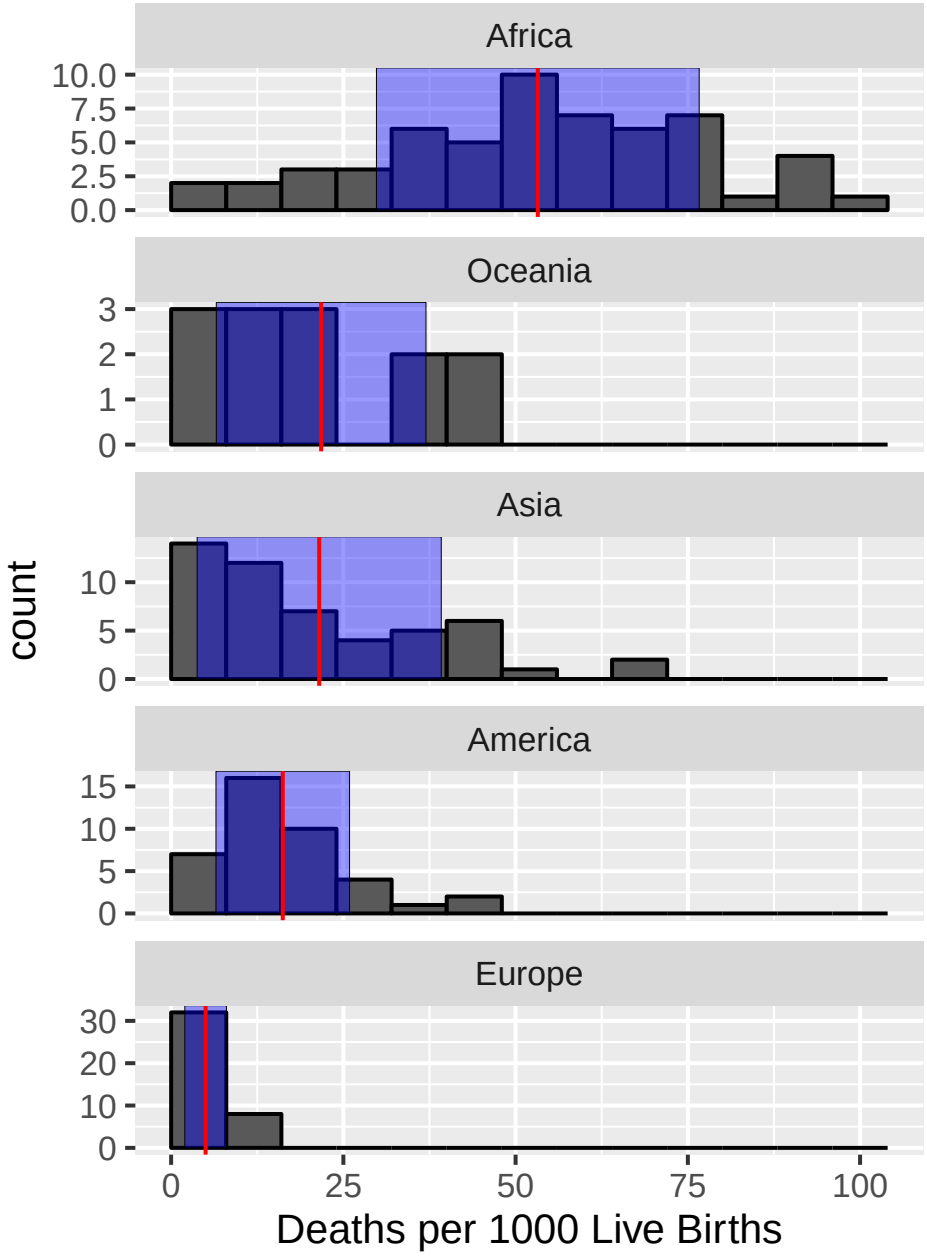
- Roughly 68% of the observations will be within 1 SD away from the mean
- Roughly 95% will be with 2 SD away from the mean

The 68% and 95% rules work very well for bell-shaped data, and reasonably well for unimodal and not seriously skewed data, but not for all data.

The 68% and 95% Rule (Section 3.1.5 in Text)

Continent	Mean	SD
Africa	53.2	23.4
Oceania	21.8	15.2
Asia	21.5	17.7
America	16.2	9.7
Europe	5.0	3.0

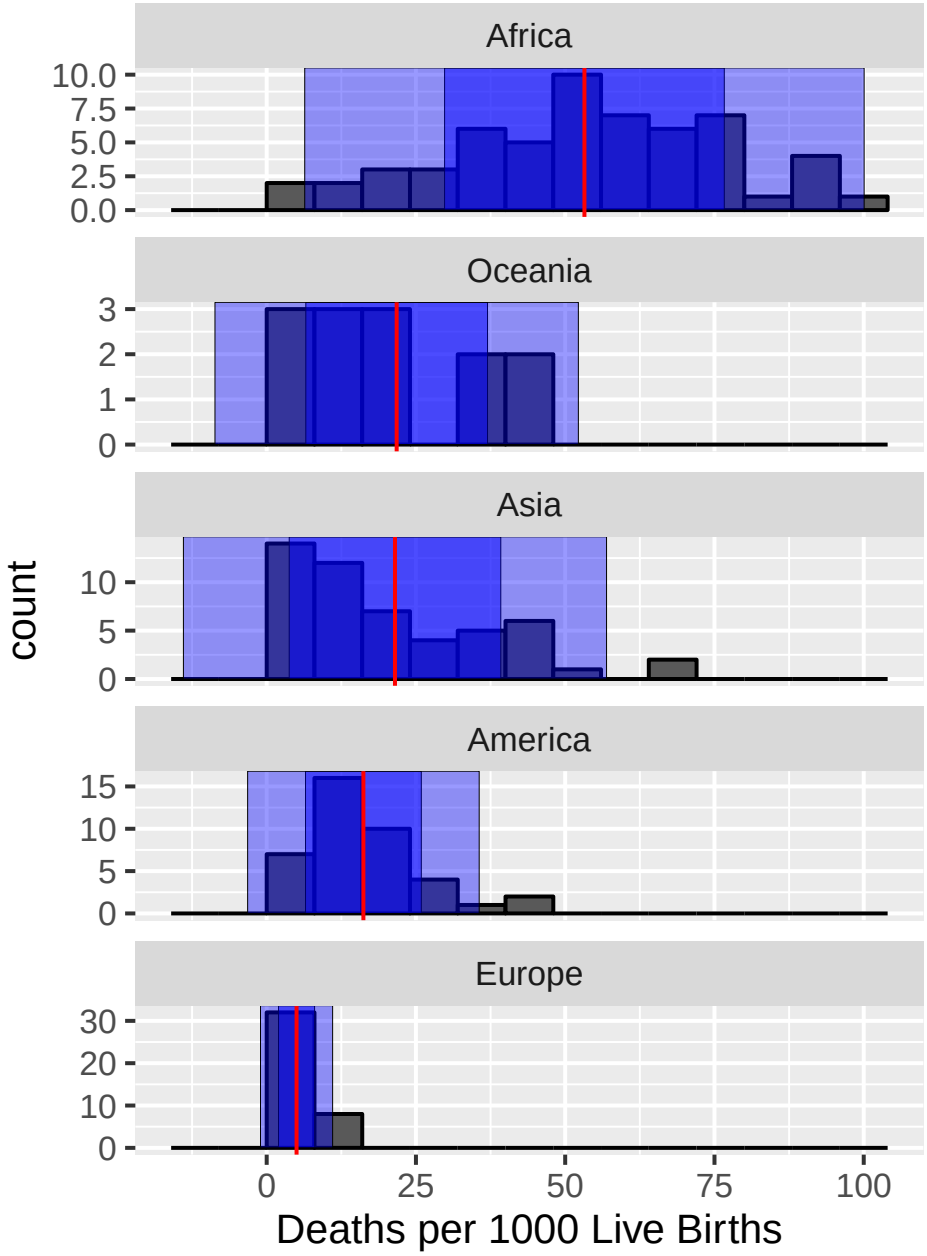
Continent	proportion within 1 SD from mean
Africa	$39/57 \approx 68\%$
Oceania	$8/13 \approx 62\%$
Asia	$35/51 \approx 69\%$
America	$29/40 \approx 72\%$
Europe	$31/40 \approx 78\%$



The 68% and 95% Rule (Section 3.1.5 in Text)

Continent	Mean	SD
Africa	53.2	23.4
Oceania	21.8	15.2
Asia	21.5	17.7
America	16.2	9.7
Europe	5.0	3.0

Continent	proportion within 2 SD from mean
Africa	$55/57 \approx 96\%$
Oceania	$13/13 = 100\%$
Asia	$49/51 \approx 96\%$
America	$38/40 = 95\%$
Europe	$39/40 = 97.5\%$



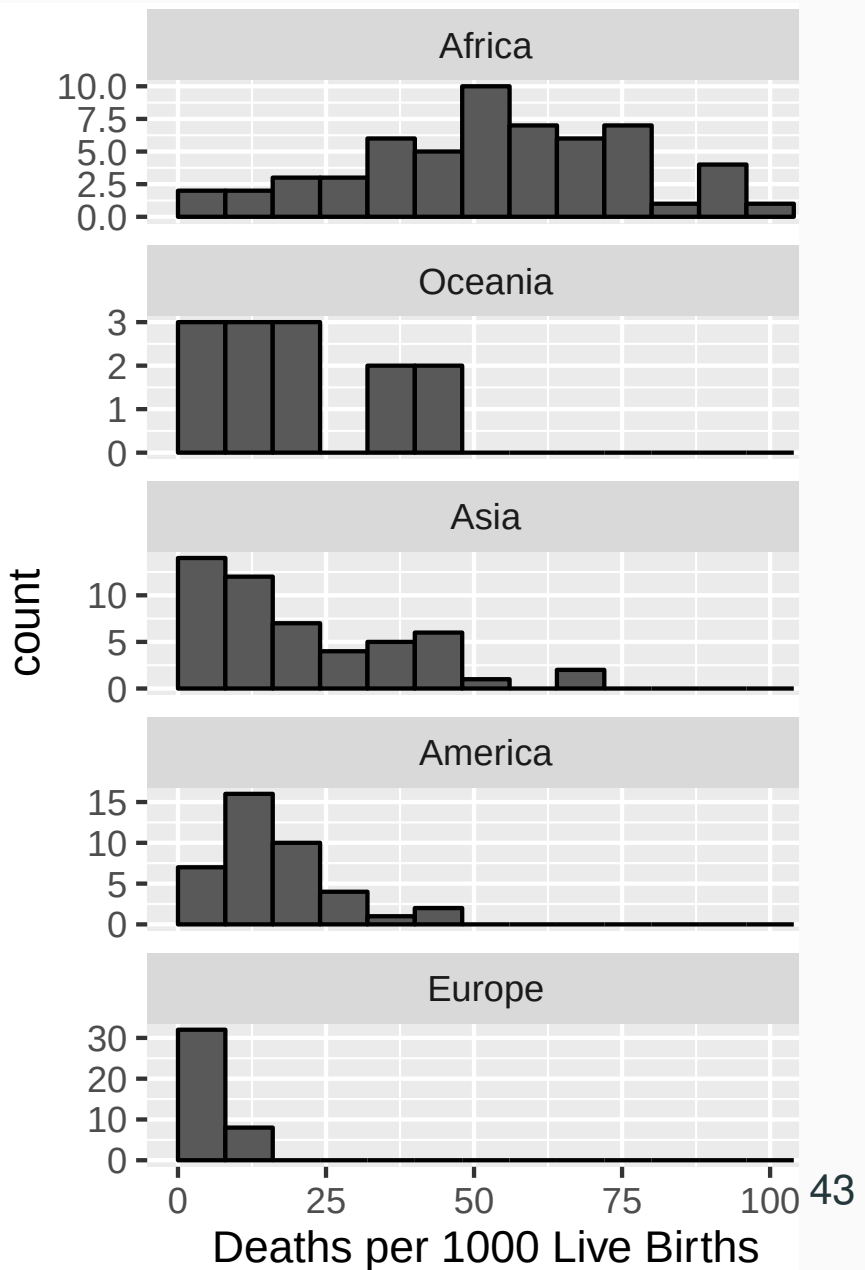
Properties of Standard Deviation (SD)

- SD measures spread about the mean and should be used only when the mean is the measure of center.
- When SD $= 0$, what are the observations look like?
  - and what if IQR $= 0$?
- and what if SD < 0?
- SD is very sensitive to outliers.
- SD has the same units of measurement as the original observations, while the variances in the square of these units.

Recap: Common Numerical Summaries of Numerical Variables

When comparing histograms, we often compare their *center* and *spread*.

- Common measure of center:
  - Mean
  - Median
- Common measure spread:
  - Range: max – min
  - Standard deviation (SD)
  - Interquartile range (IQR)

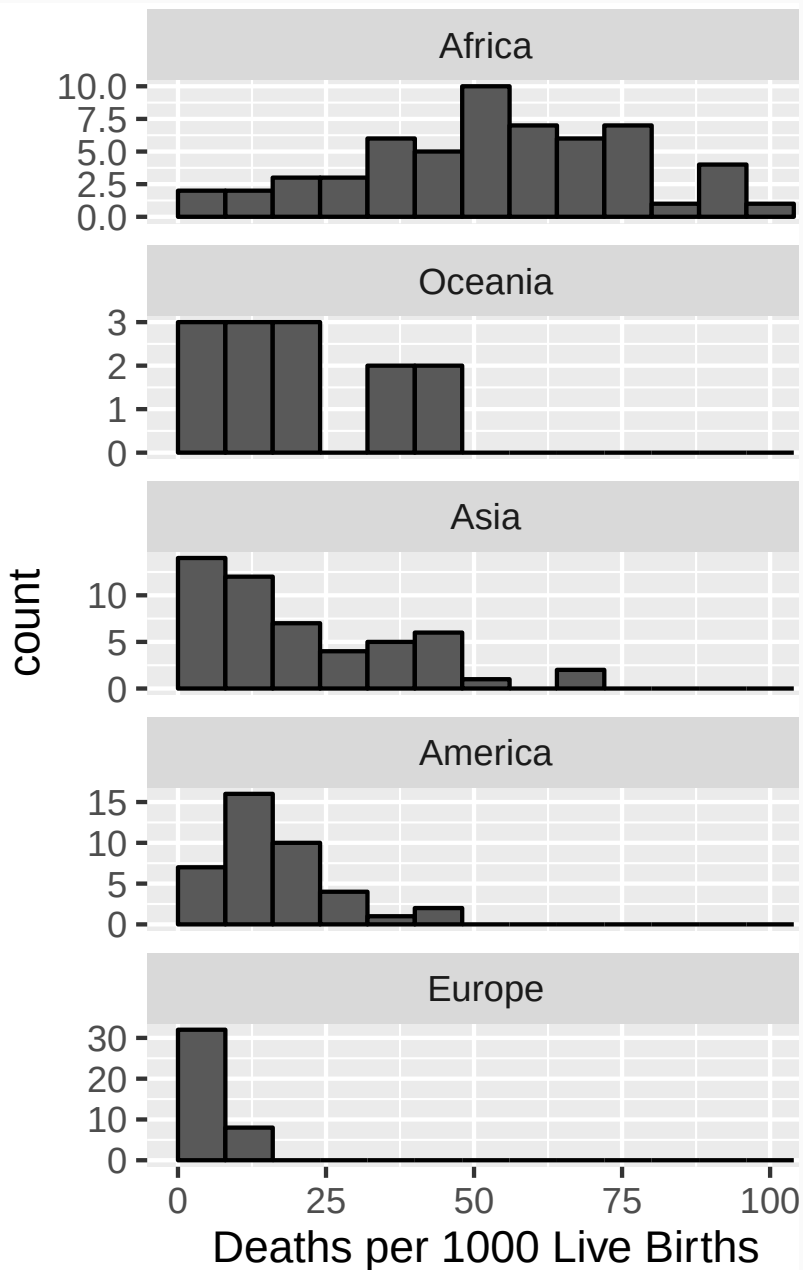


Recap: Common Numerical Summaries of Numerical Variables

When comparing histograms, we often compare their *center* and *spread*.

- Common measure of center:
  - Mean
  - Median
- Common measure spread:
  - Range: max – min
  - Standard deviation (SD)
  - Interquartile range (IQR)

Measures of center and spread are important summaries of a distribution. But they don't tell about modality, skewness, and whether there are outliers. **Always check the histogram!**



Recap: Common Graphical Displays for Numerical Variables

- **Histograms**
  A histogram breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class.

- **Dotplots**: See Section 1.6.2
  - Pros: Shows each individual data point
  - Cons: Hard to read even for moderately big data

- **Box-plots**

Scatterplots

Example (Diamonds Data)

Diamonds Data: records of 53940 diamonds, collected from *http://www.diamondse.info/* in 2008.

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
...										
53940	0.75	Ideal	D	SI2	62.2	55	2757	5.83	5.87	3.64

See Lab 02 for R codes:

http://www.stat.uchicago.edu/~yibi/s220/labs/lab02.html

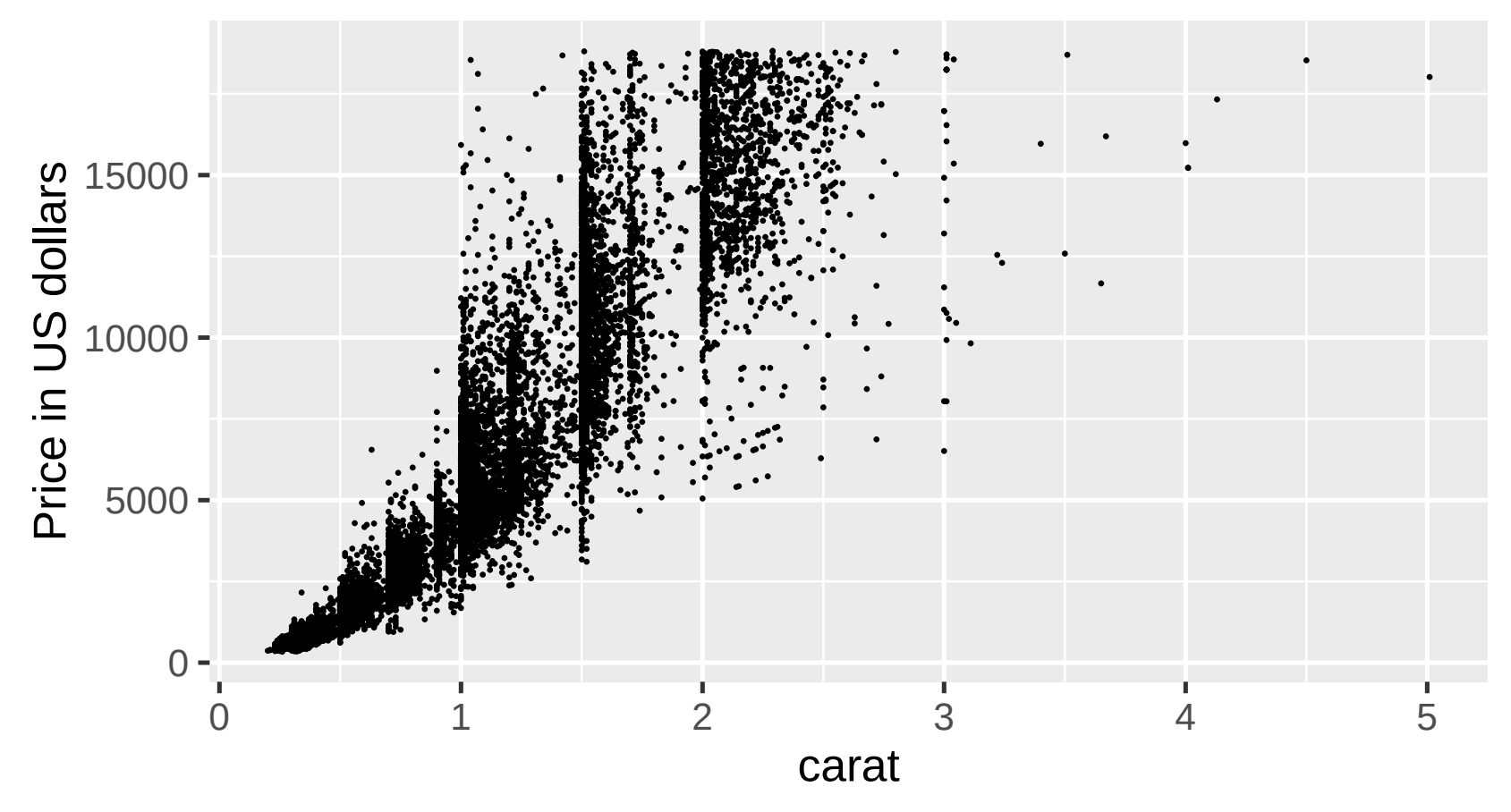
Example (Diamonds Data)

The diamonds data contains 10 variables, including

- `price`: price in US dollars
- `carat`: weight of the diamond
- `cut`: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- `color`: diamond color, from J (worst) to D (best)
- `clarity`: a measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))

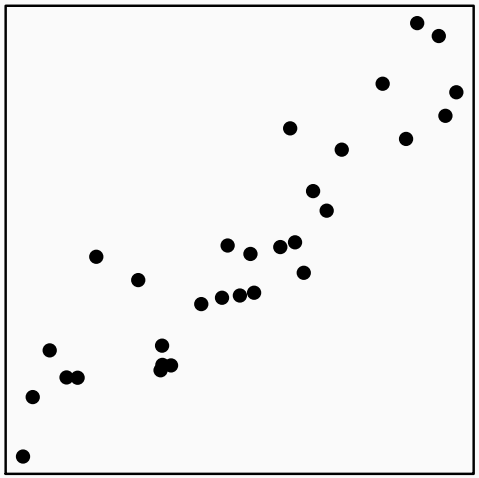
Scatter Plot of Prices and Carats of Diamonds

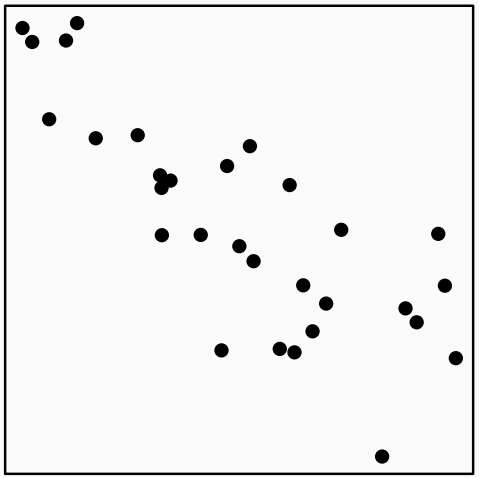
What patterns do you observe from the scatter plot?



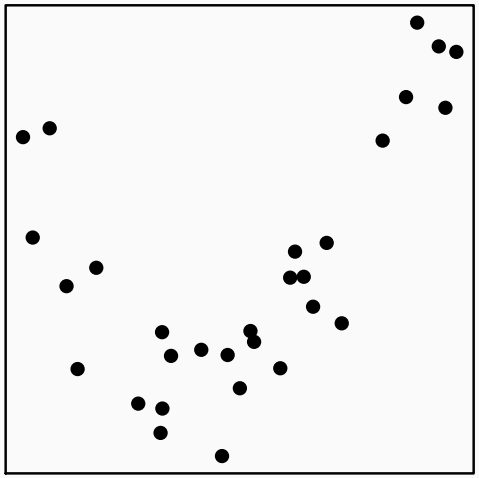
Forms of Relationship

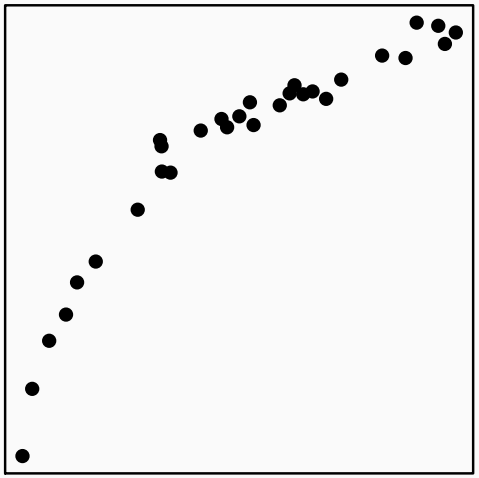
Linear Relationship



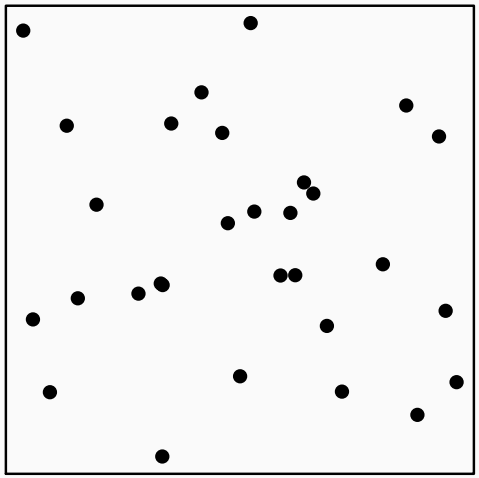


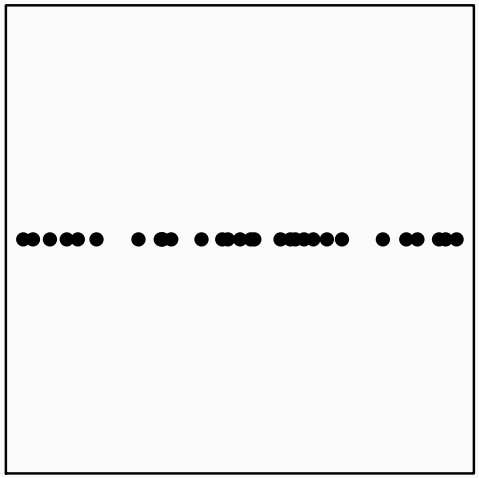
Nonlinear Relationship





No Relationship



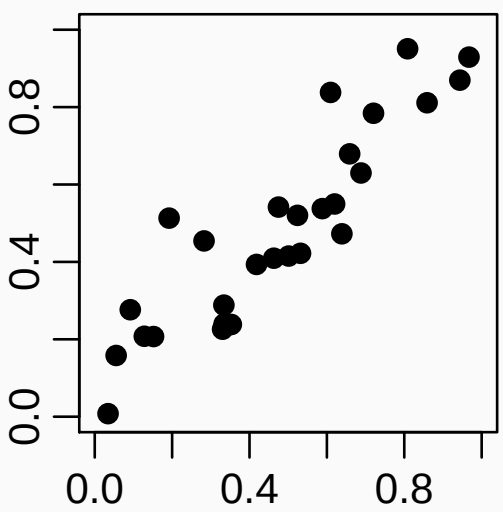


Positive and Negative Relation

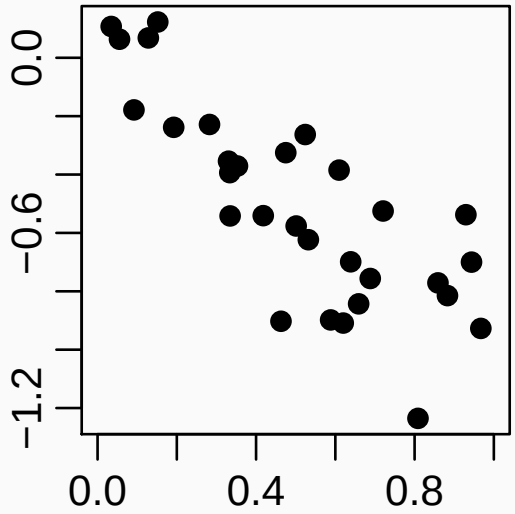
Positive
Negative **association**: High values of one variable tend to occur

together with high / low values of the other variable.

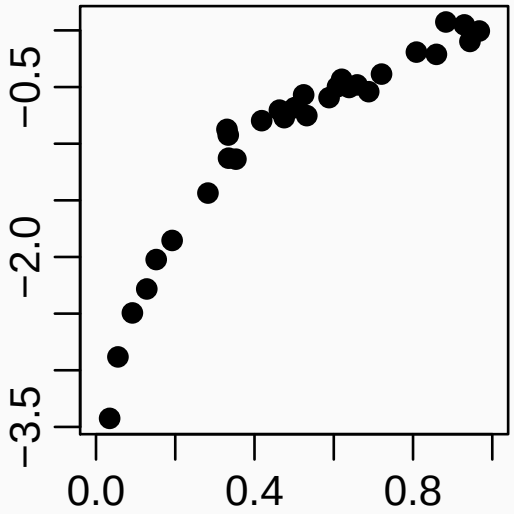
positive linear



negative linear



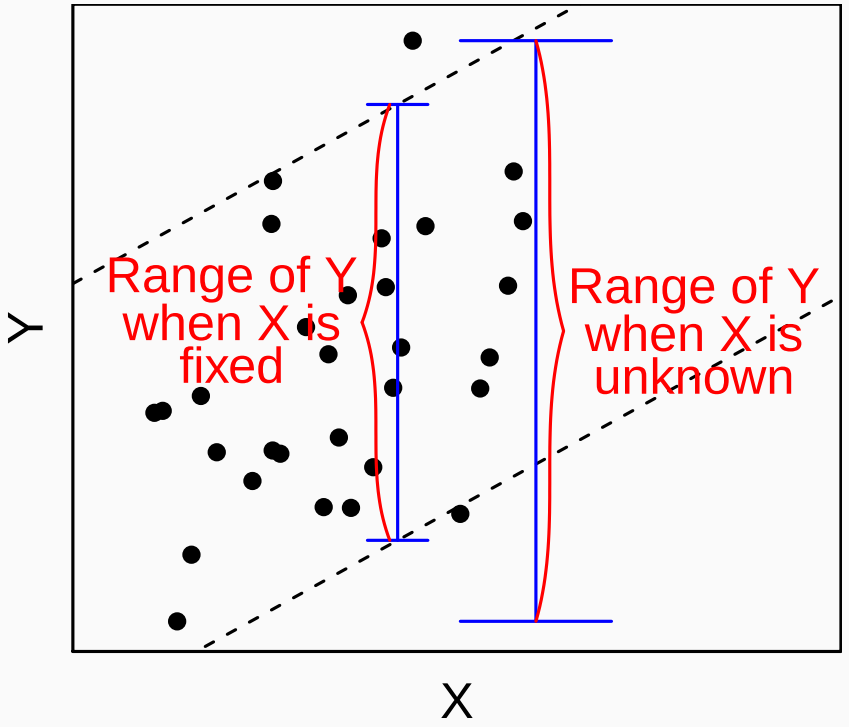
positive nonlinear



Strength of the Relationship

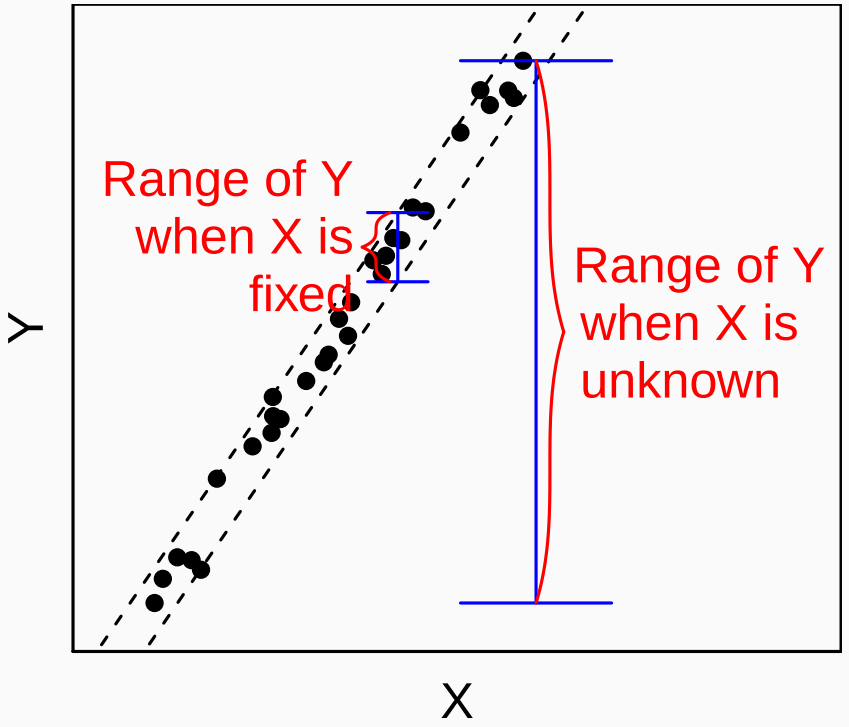
The **strength** of the relationship between the two variables can be seen by how much variation, or **scatter**, there is around the main form.

Weak Association



Large spread of Y when X is known

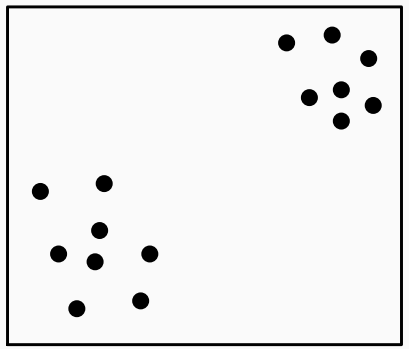
Strong Association



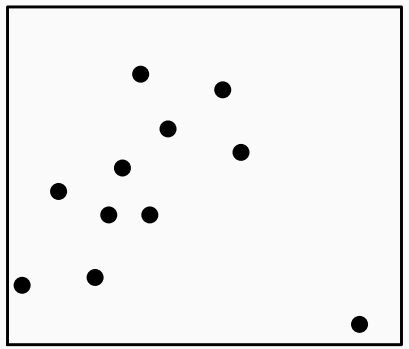
Small spread of Y when X is known

Clusters & Outliers

Sometimes points in a scatter plot form *clusters*, which indicates data may be comprised of several distinct kinds of individuals.



Scatter plots can also be used to spot *outliers*.



Recap: What to Look For in a Scatterplot?

- What is the *form* of the relationship?
  (linear, curved, clustered ...)

Recap: What to Look For in a Scatterplot?

- What is the *form* of the relationship?
  (linear, curved, clustered ...)
- What is the *direction* of the relationship?
  (positive, negative)

Recap: What to Look For in a Scatterplot?

- What is the *form* of the relationship?
  (linear, curved, clustered ...)
- What is the *direction* of the relationship?
  (positive, negative)
- What is the *strength* of the relationship?
  (strong, weak, ...)

Recap: What to Look For in a Scatterplot?

- What is the *form* of the relationship?
  (linear, curved, clustered ...)
- What is the *direction* of the relationship?
  (positive, negative)
- What is the *strength* of the relationship?
  (strong, weak, ...)
- Are the points *clustered*?

Recap: What to Look For in a Scatterplot?

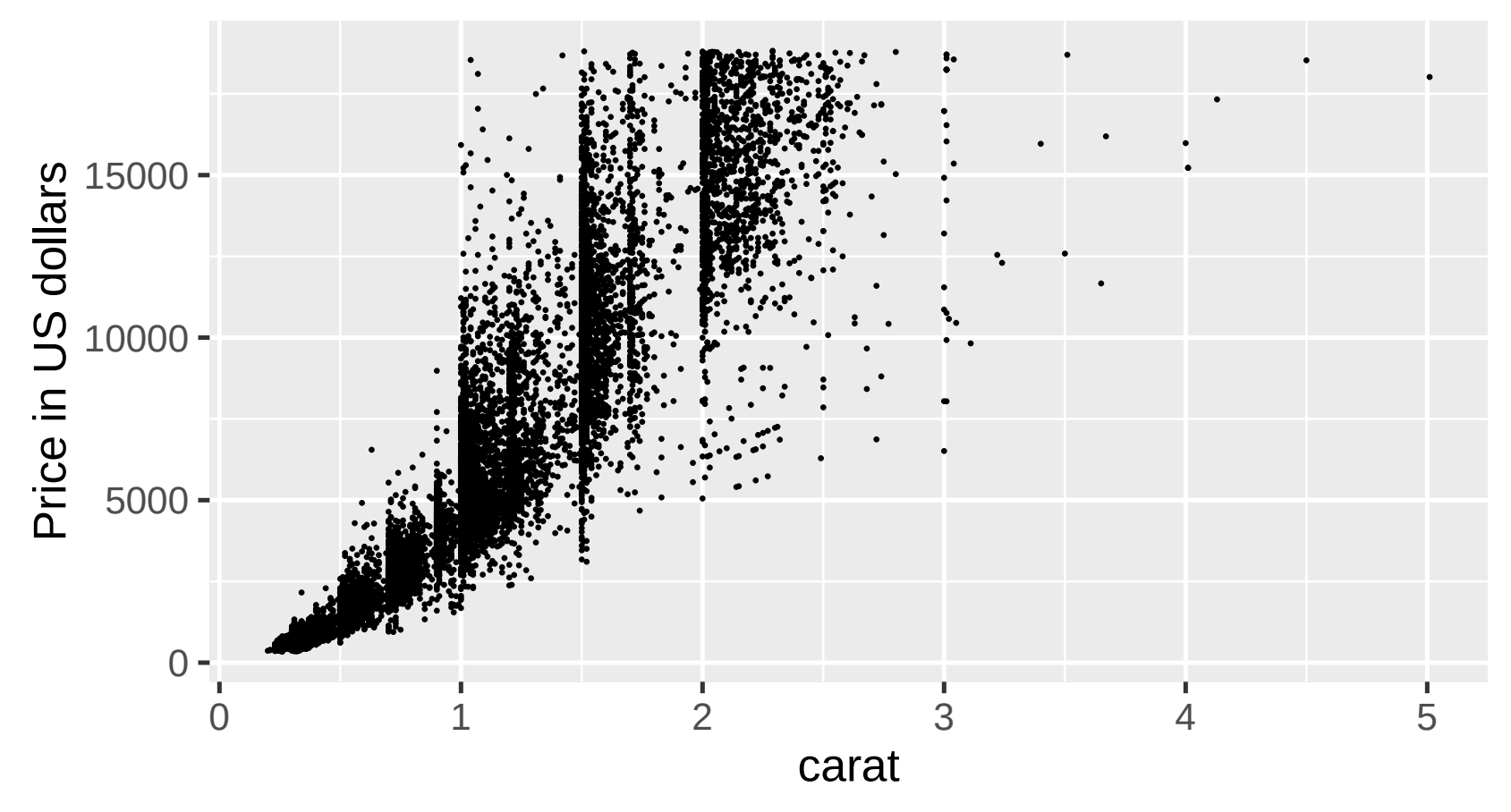
- What is the *form* of the relationship?
  (linear, curved, clustered ...)
- What is the *direction* of the relationship?
  (positive, negative)
- What is the *strength* of the relationship?
  (strong, weak, ...)
- Are the points *clustered*?
- Are there any deviations from the overall pattern, i.e. *outliers*?

Recap: What to Look For in a Scatterplot?

- What is the *form* of the relationship?
  (linear, curved, clustered ...)
- What is the *direction* of the relationship?
  (positive, negative)
- What is the *strength* of the relationship?
  (strong, weak, ...)
- Are the points *clustered*?
- Are there any deviations from the overall pattern, i.e. *outliers*?
- In fact, the list above is not exclusive.
  Any unusual pattern worths investigation.

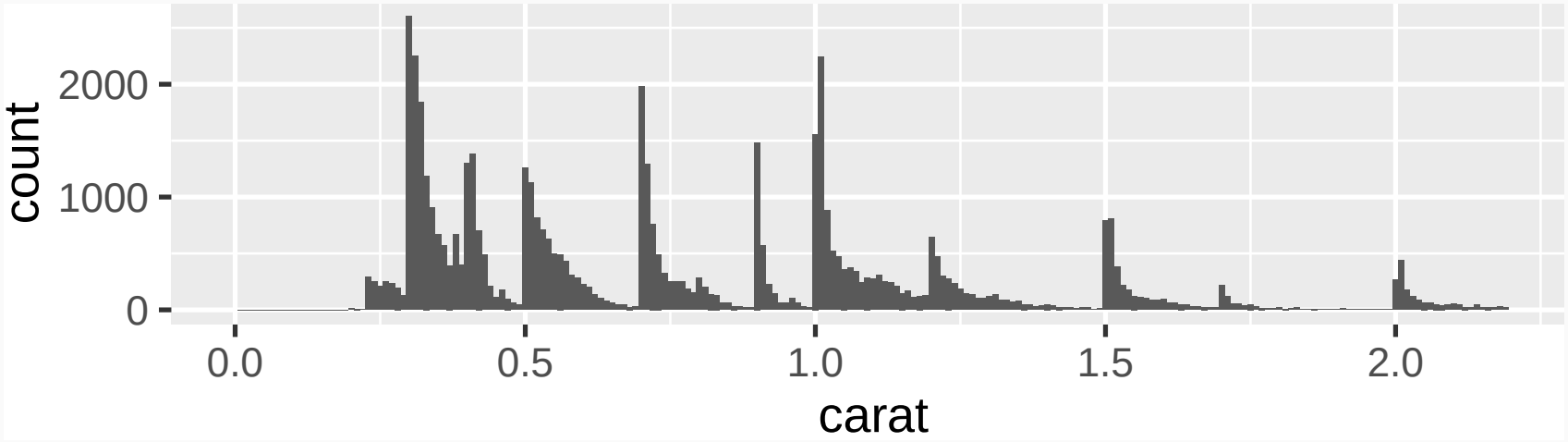
Example: Scatter Plot of Prices and Carats of Diamonds

What do you observe from the scatter plot below?



Weights of diamonds cluster in some benchmark carats (0.3, 0.5, 0.7, 0.9, 1, 1.2, 1.5, 2, ...), e.g.,

- 1981 diamonds weigh 0.7 carat, 1294 weigh 0.71 carat, but only 26 weigh 0.69 carat
- 1558 diamonds weigh 1 carat, 2242 weigh 1.01 carat, but only 23 weigh 0.99 carat
- 793 diamonds weigh 1.5 carat, 807 weigh 1.51 carat, but only 11 weigh 1.49 carat

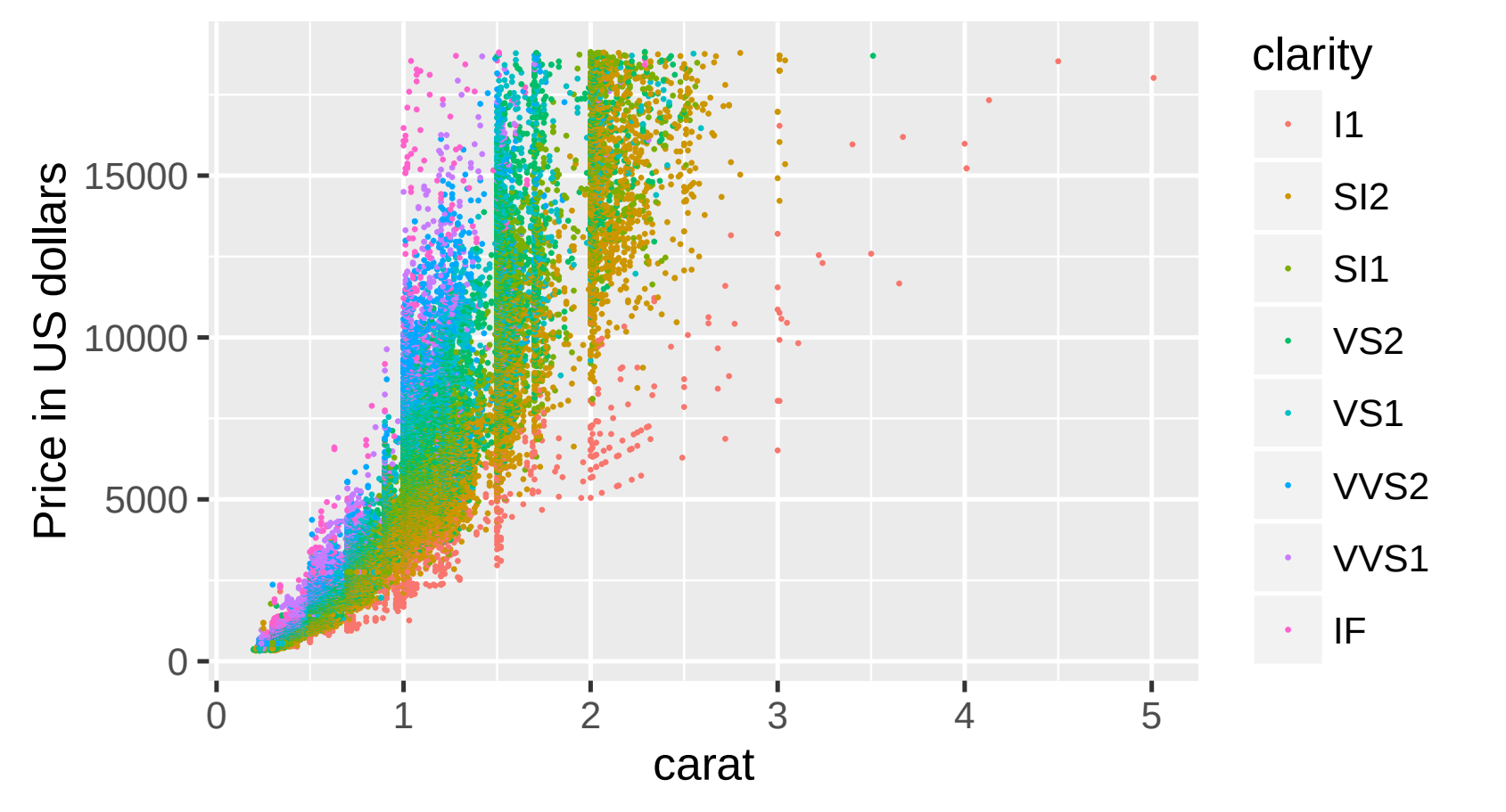


(This histogram is truncated at 2.2 carat. The x axis in fact extends to 5 carat)

Making Scatterplots More Informative

The scatterplot shows relationship between two variables - weight and price, but there are many other attributes of the data. E.g., we might want to see how the quality of the cut, or the color, or the clarity, affects the price.

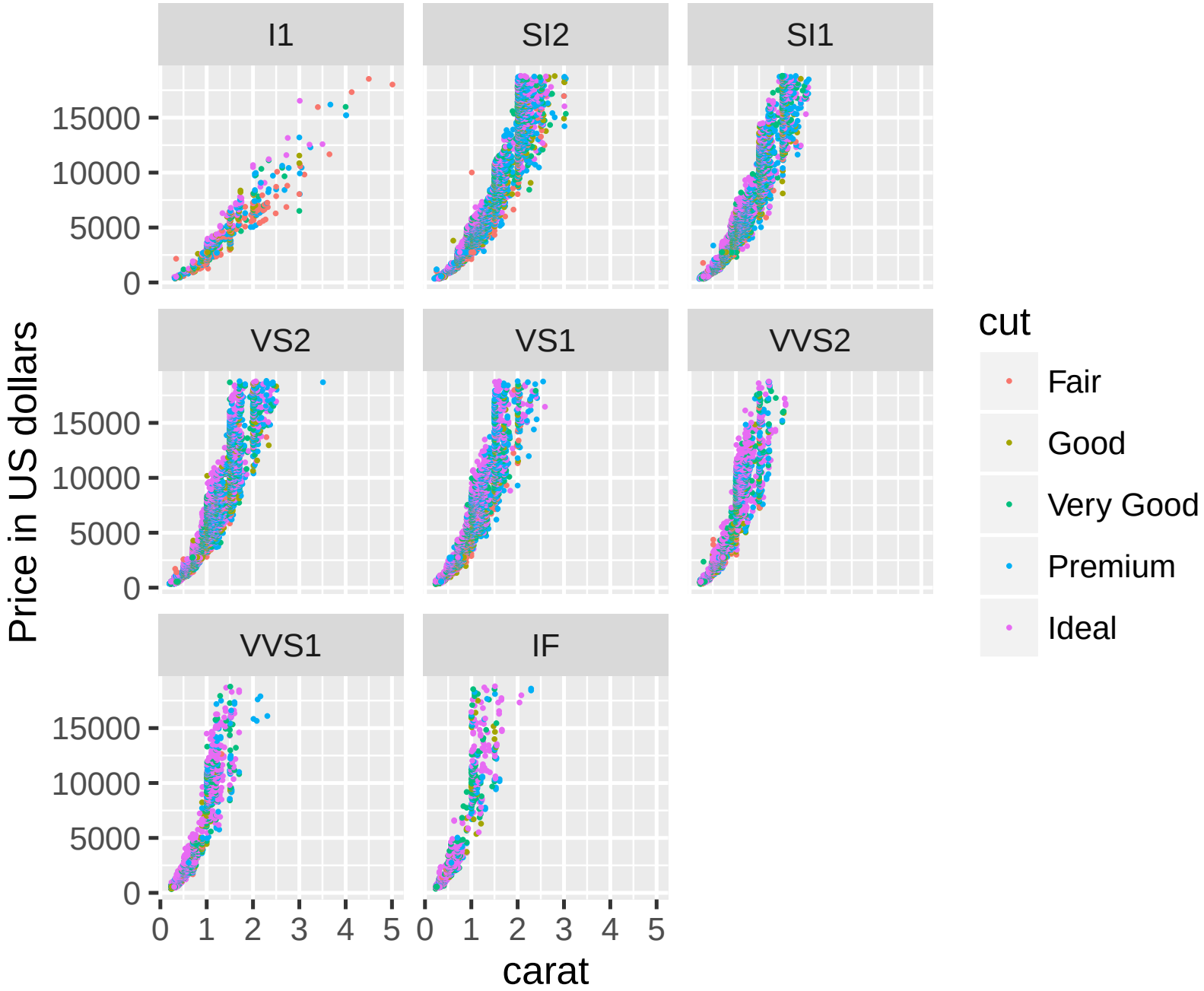
We can make the scatterplot more informative, by using the color of points to represent, e.g., the clarity of diamonds.



For diamonds of the same weight, the price increases with the clarity of the diamond.

Another way to communicate information about a 3rd (categorical) variable in a scatter plot is to divide the plot up into multiple plots, one for each level, letting you view them separately.

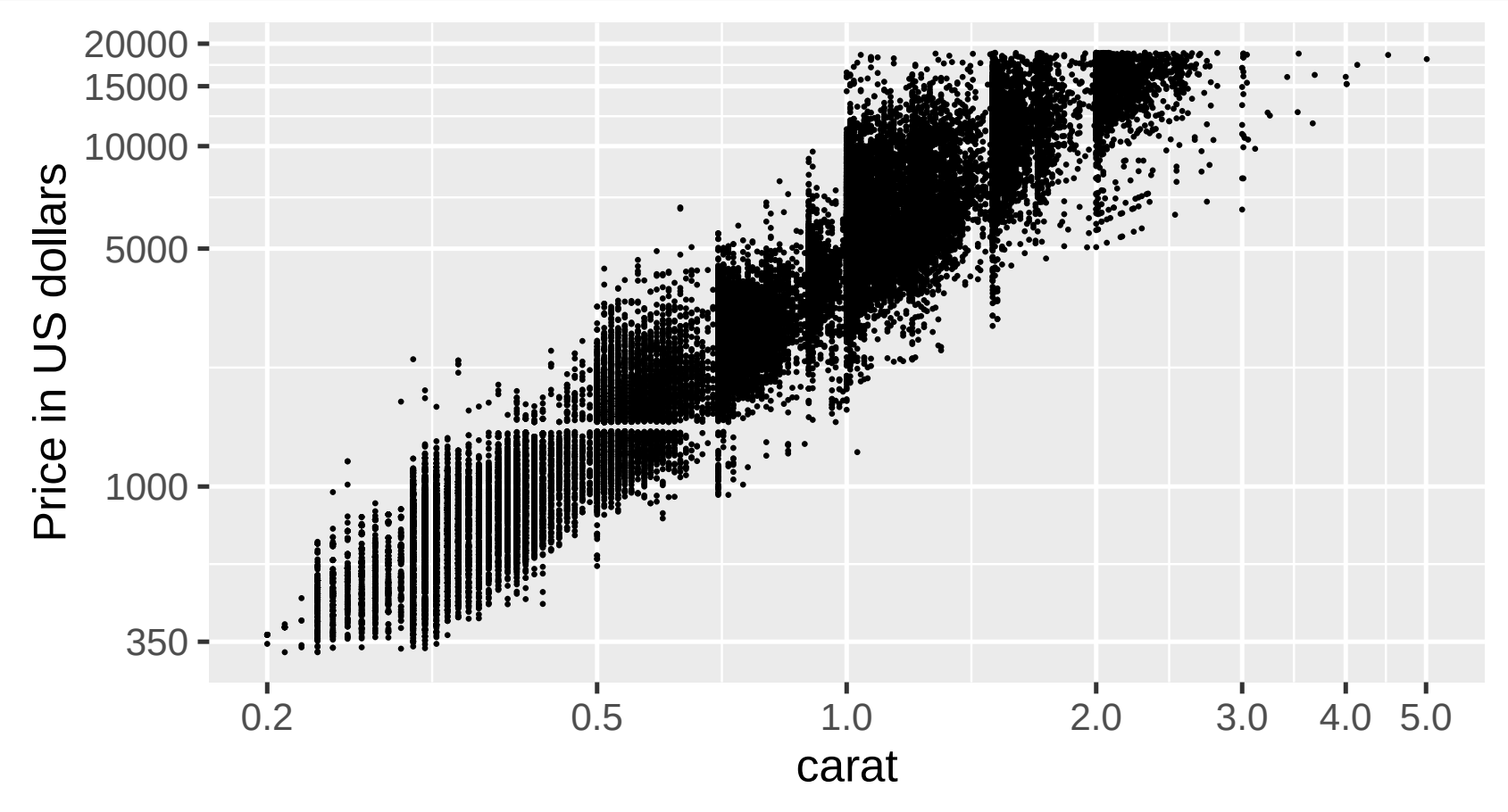
Such a plot is shown in the next slide.


I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
Price in US dollars
15000
10000
5000
0
0 1 2 3 4 5
carat
cut
Fair
Good
Very Good
Premium
Ideal

Transforming Data

Scatterplot with Both X-Y Variables Transformed

Scatterplot when both carat and price are log-transformed.



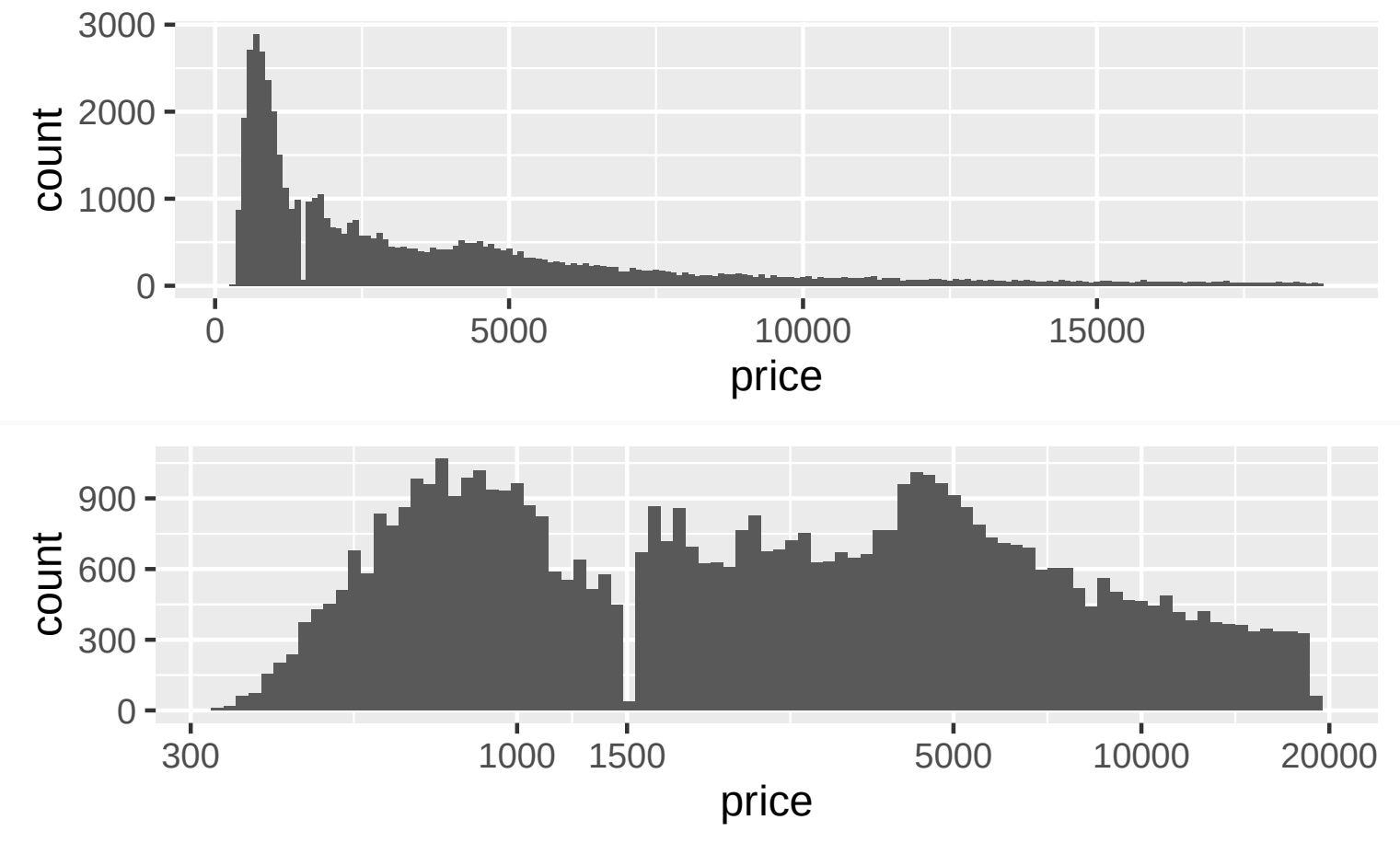
Scatter Plot of Prices and Carats of Diamonds

What's the advantage of transformation (for the diamond data)?

- *After log-transformation, the relationship between* $\log(Price)$ *and* $\log(carat)$ *become a simple linear relationship, and the variability of log(price) stays about the same with carat.*
- *Before transformation, points are too tightly clustered near 0 to reveal the structure of the distribution there. After transformation, the scatterplot shows a clear gap in price around $1500.*

Histogram of Log-Transformed Data

The prices for diamonds are heavily skewed, but on a log scale seem much better behaved. In fact, we can see some evidence of bimodality on the log scale and the gap at $1500.



Transformation

- Sometimes, only one variable in the scatterplot is transformed, instead of both.
- Other commonly used transformations other than log: square-root, cubic-root, inverse, ...
- Why transforming data?
  - Highly skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.
  - Variables may exhibit a simpler relationship after transformation
  - Sometimes, in a scatter plot or a histogram, when observations are too tightly clustered near 0 to look at the pattern there, transformation might help.
- However, the transformed variable might be meaningless, making the result of analysis hard to interpret.