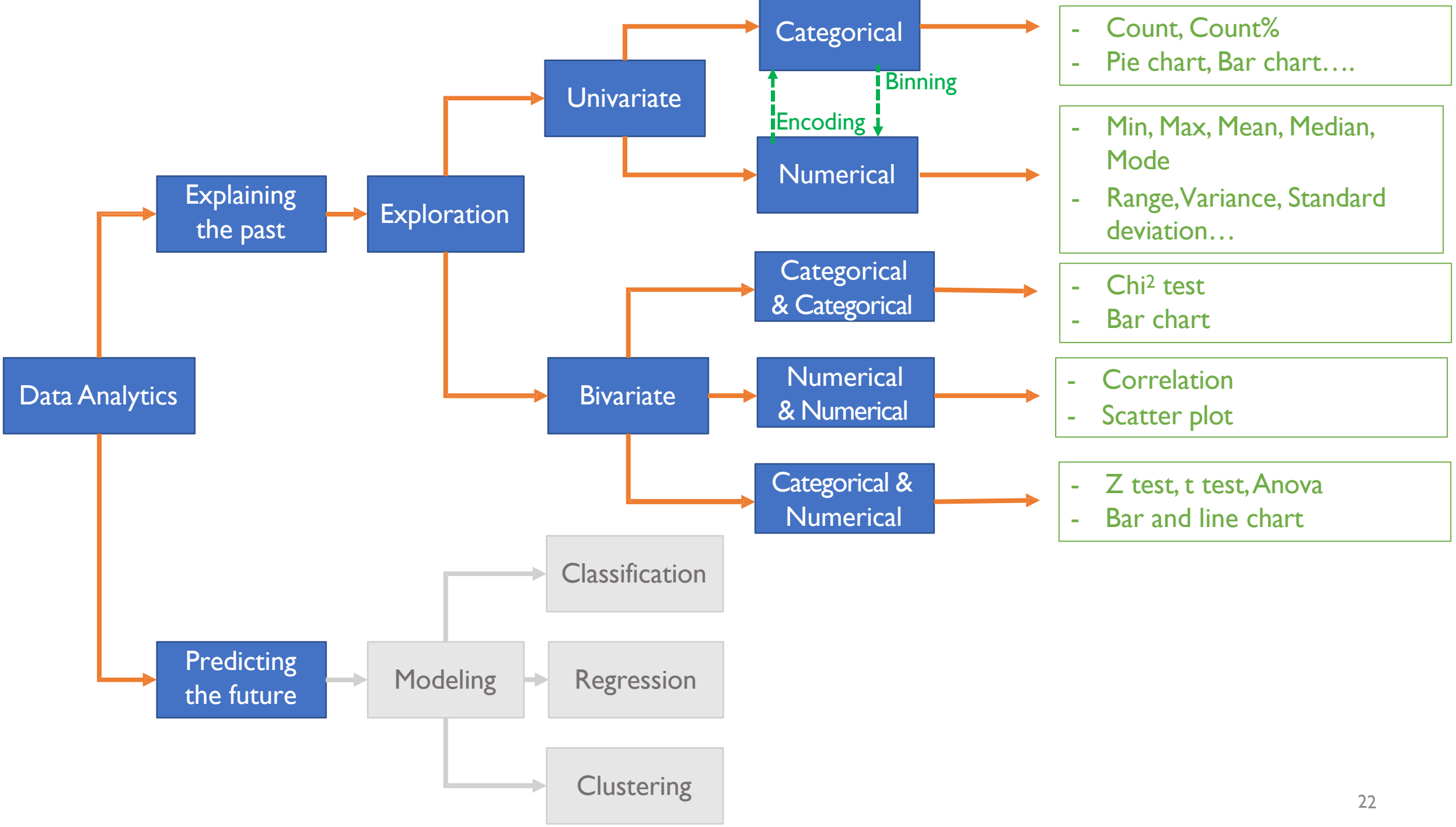
# Data Exploration (part 1)

Summer 2024


Data Analytics
Explaining the past
Exploration
Univariate
Categorical
Binning
Encoding
Numerical
Bivariate
Categorical & Categorical
Numerical & Numerical
Categorical & Numerical
Predicting the future
Modeling
Classification
Regression
Clustering
- Count, Count%
- Pie chart, Bar chart….
- Min, Max, Mean, Median, Mode
- Range, Variance, Standard deviation…
- Chi² test
- Bar chart
- Correlation
- Scatter plot
- Z test, t test, Anova
- Bar and line chart

# Data Exploration

Data Exploration is about describing the data by means of statistical and visualization techniques.

We explore data in order to bring important aspects of that data into focus for further analysis.

1. Univariate analysis
2. Bivariate analysis

# UNIVARIATE ANALYSIS

# UNIVARIATE ANALYSIS

- Univariate analysis explores variables (attributes) one by one.

- Variables could be either *categorical* or *numerical*.

- There are different statistical and visualization techniques of investigation for each type of variable.

  - Numerical variables can be transformed into categorical counterparts by a process called binning or discretization.
  - It is also possible to transform a categorical variable into its numerical counterpart by a process called encoding.

# Categorical attributes

- We present methods to analyze categorical attributes.

- Because categorical attributes have only symbolic values, many of the arithmetic operations cannot be performed directly on the symbolic values.

- However, we can compute the frequencies of these values and use them to analyze the attributes.

# Categorical attributes

- **Mode** : Most frequently occurring value in the given data
  - Example:

```
Data = ["Car", "Bat", "Bat", "Car", "Bat", "Bat", "Bat", "Bike"]
Mode = "Bat”
```

# Categorical attributes

- **Count, Count% :** A categorical variable is summarized by a table showing the count or the percentage of cases in each category

```
Data = ["Car", "Bat", "Bat", "Car", "Bat", "Bat", "Bat", "Bike"]
```

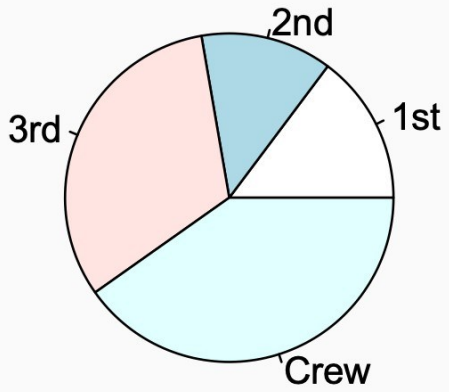
| Data | Count | Count % |
|---|---|---|
| Car | 2 | 25% |
| Bat | 5 | 62.5% |
| Bike | 1 | 12.5% |

# Categorical attributes

- Categorical variables are often displayed by a bar plot or a pie chart.

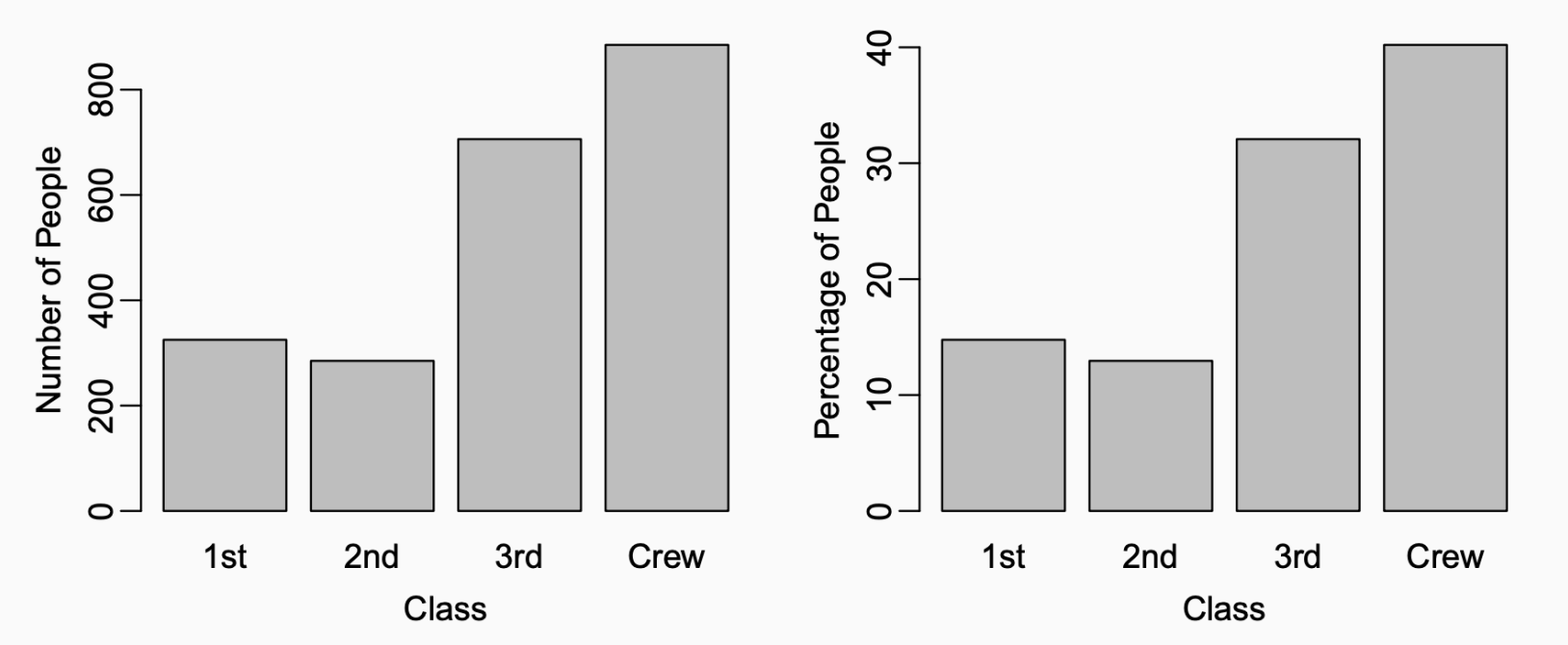
Ex: Passengers on Titanic

| Class | Freq | Percent |
|---|---|---|
| 1st | 325 | 14.8% |
| 2nd | 285 | 12.9% |
| 3rd | 706 | 32.1% |
| Crew | 885 | 40.2% |
| Total | 2201 | 100% |

# Bar plots

- A *bar plot* is a common way to display a single categorical variable.
- A bar plot where proportions instead of frequencies are shown is called *a relative frequency bar plot*.



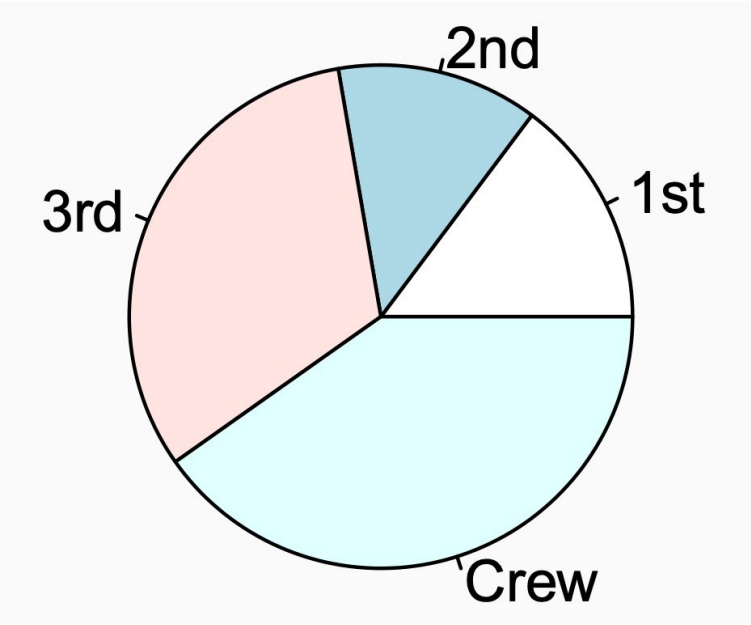
# How are Bar Plots Different From Histograms?

- Bar plots are used for displaying distributions of categorical variables, *while* histograms are used for numerical variables.

- The horizontal axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order
  - (though some orderings make more sense than others, especially for ordinal variables.)

# Why are Bar Plots Recommend Over Pie Charts?

- In a pie chart, the areas of slices represents the percentages of categories.
- However, it is generally more difficult to compare group sizes in a pie chart than in a bar plot, especially when categories have nearly identical counts or proportions




Without looking at the counts, can you tell which class have fewest people from the pie?

# Numerical Attributes

# Histograms

- Histograms are commonly used to explore numerical attributes.

- 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

| | 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 |

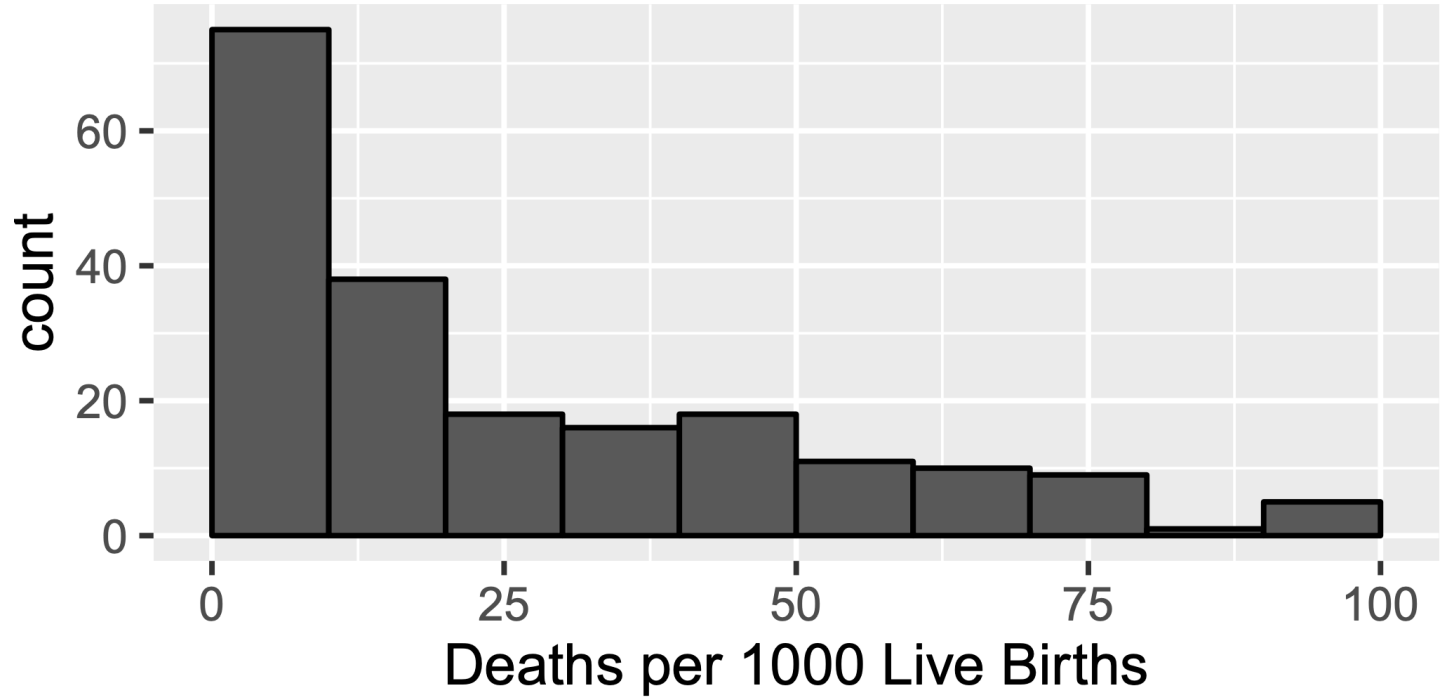
| | Country.Region | Continent | X2010.2015 |
|---|---|---|---|
| 1 | Burundi | Africa | 77.9 |
| 2 | Comoros | Africa | 58.1 |
| 3 | Djibouti | Africa | 55.3 |
| ... | | | |
| 200 | Samoa | Oceania | 19.7 |
| 201 | Tonga | Oceania | 20.4 |

- 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 |

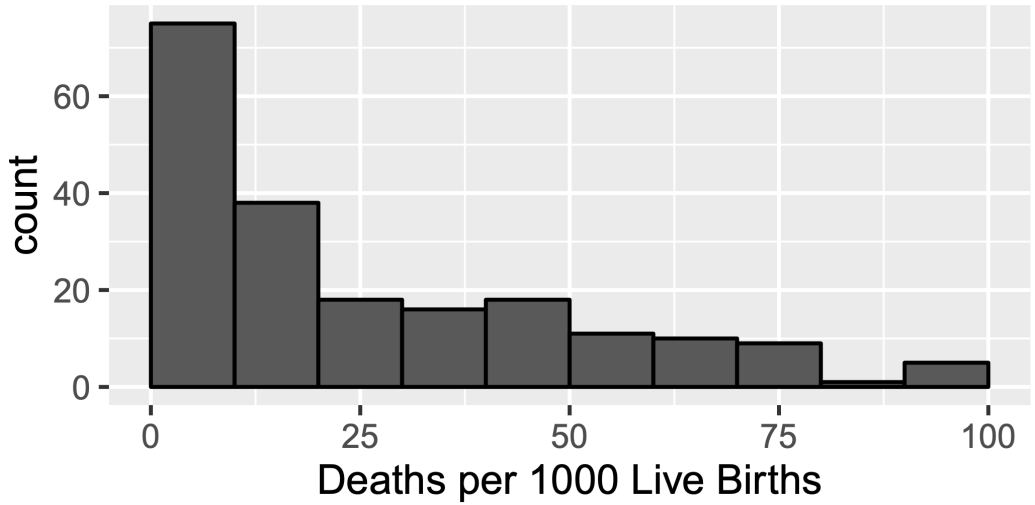
| Interval | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 | 90-100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Count | 75 | 38 | 18 | 16 | 18 | 11 | 10 | 9 | 1 | 5 |

- Step 3: Draw the histogram
  - No space between bars.
  - Label the horizontal axes (with units)



# Histograms

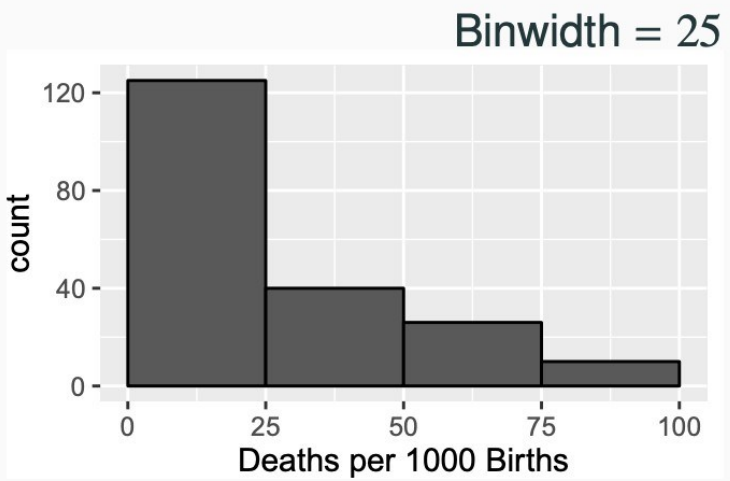
- Histograms provide a view of the *data density*. Higher bars indicate regions with more observations.
- Histograms are especially convenient for describing the *shape* of the data distribution.
- The selection of *bin width* can alter the shape of histogram

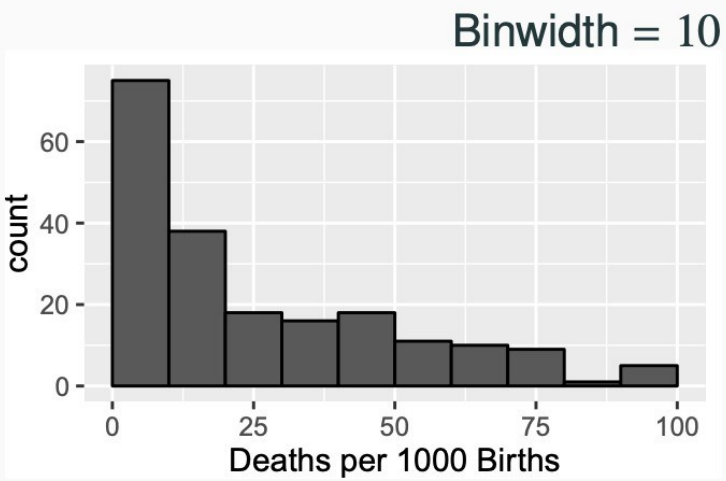


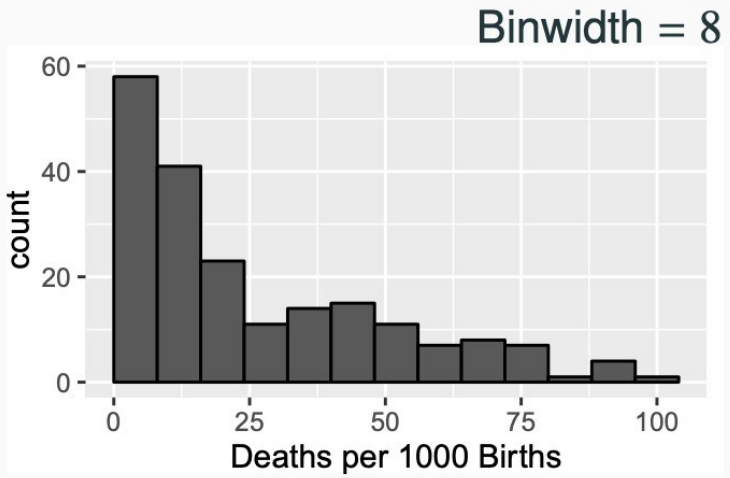
# Try Different Binwidth

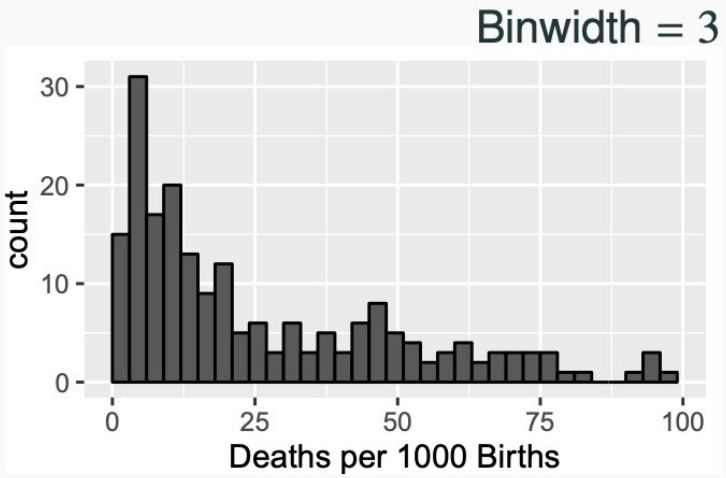
Which one(s) of these histograms reveal too much about the data?

Which hide too much?



# Selection of Bin width

- 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?

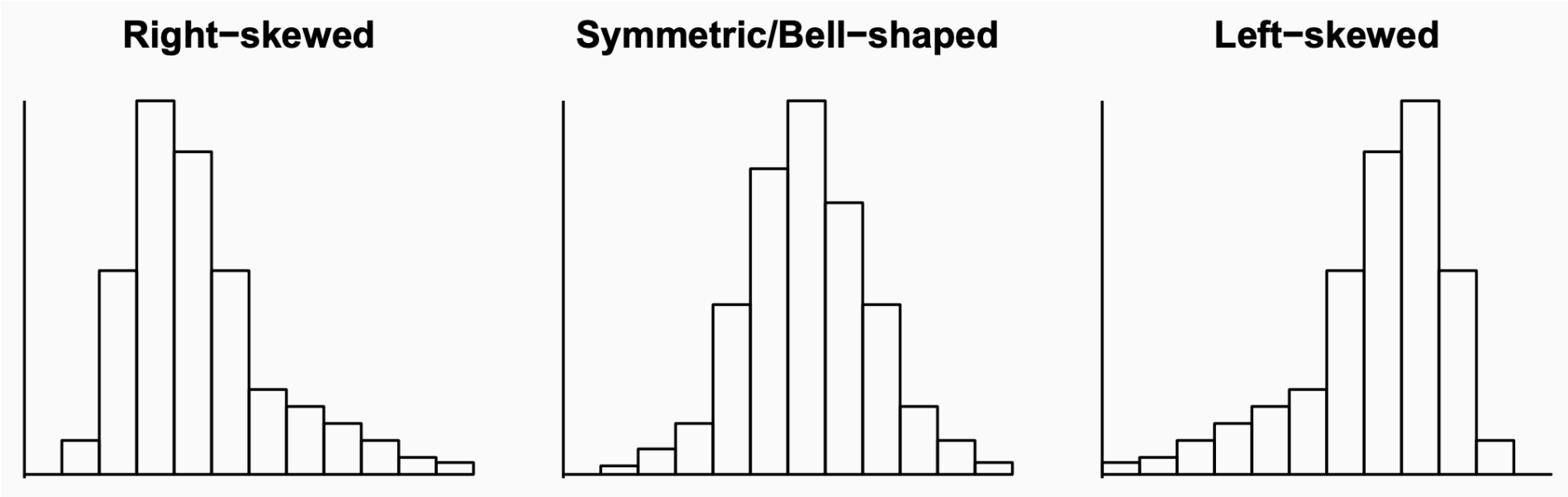
Once you have histogram, interpret the distribution of the data.

Shape: to understand the distribution

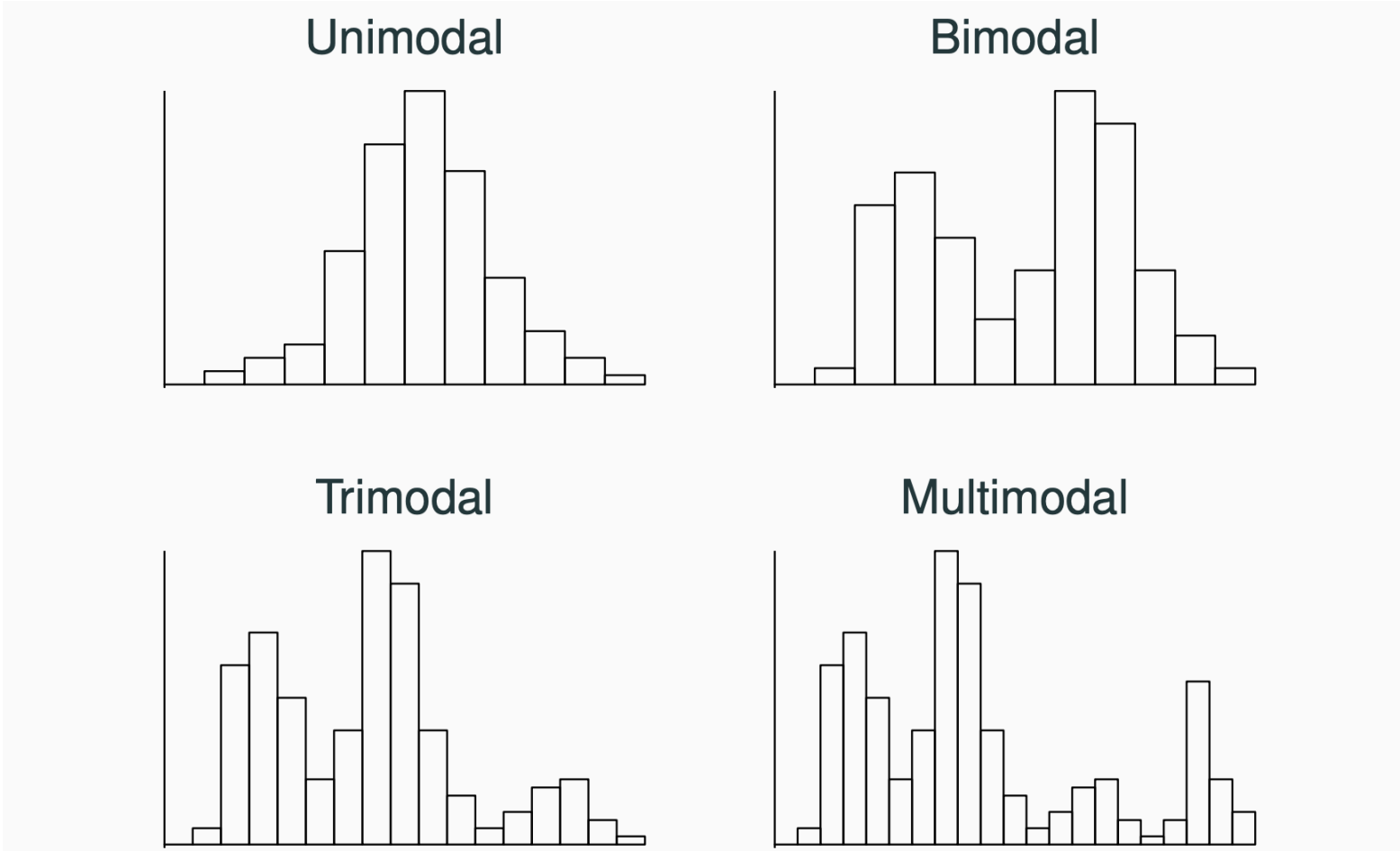
- symmetric or skewed (lopsided)
- number of modes (peaks)

- Outliers or gap in the data: 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 (will introduce soon)

# Skewness of Histograms



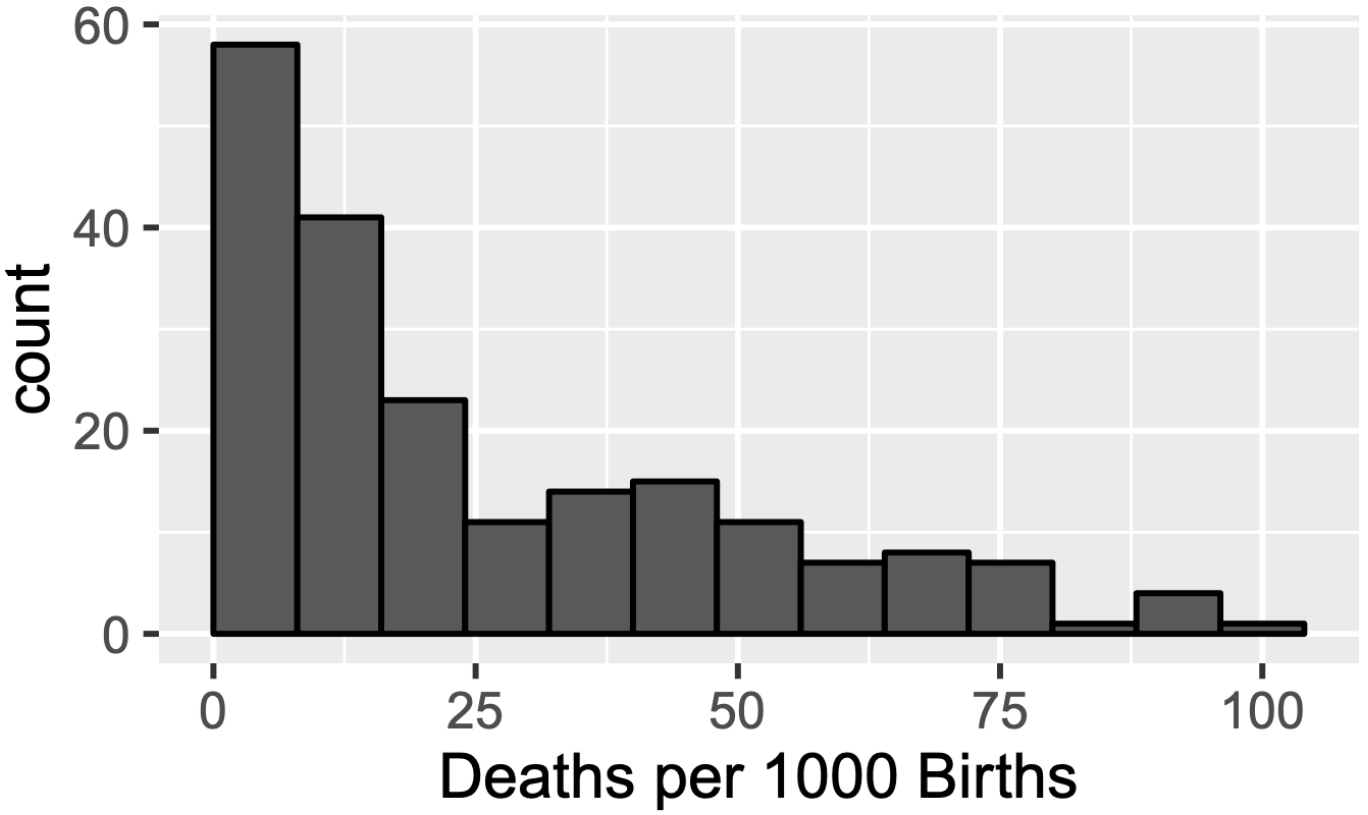
# Mode of Histograms (= Number of Peaks)




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

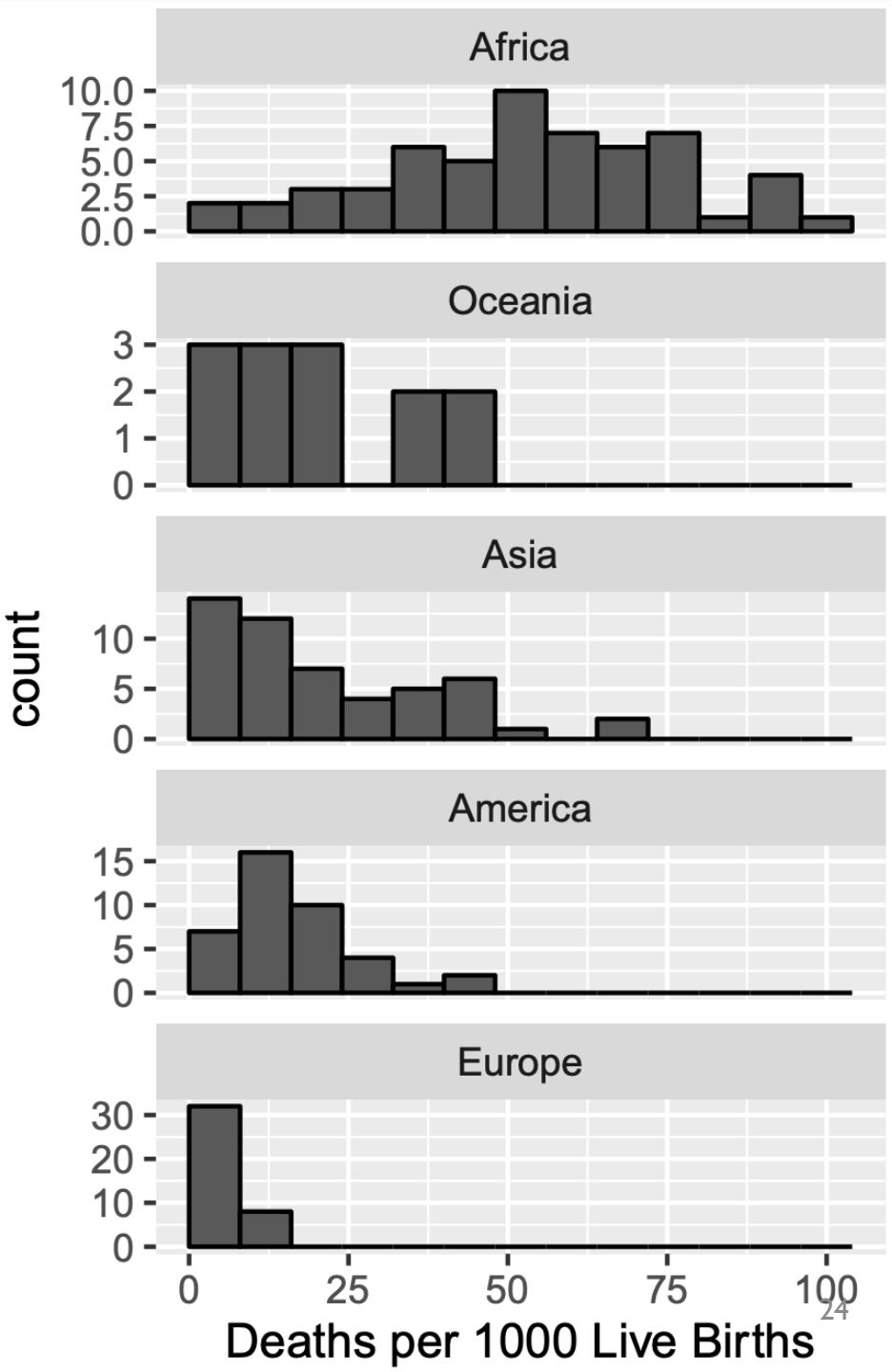
# Example (Infant Mortality Rates)

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



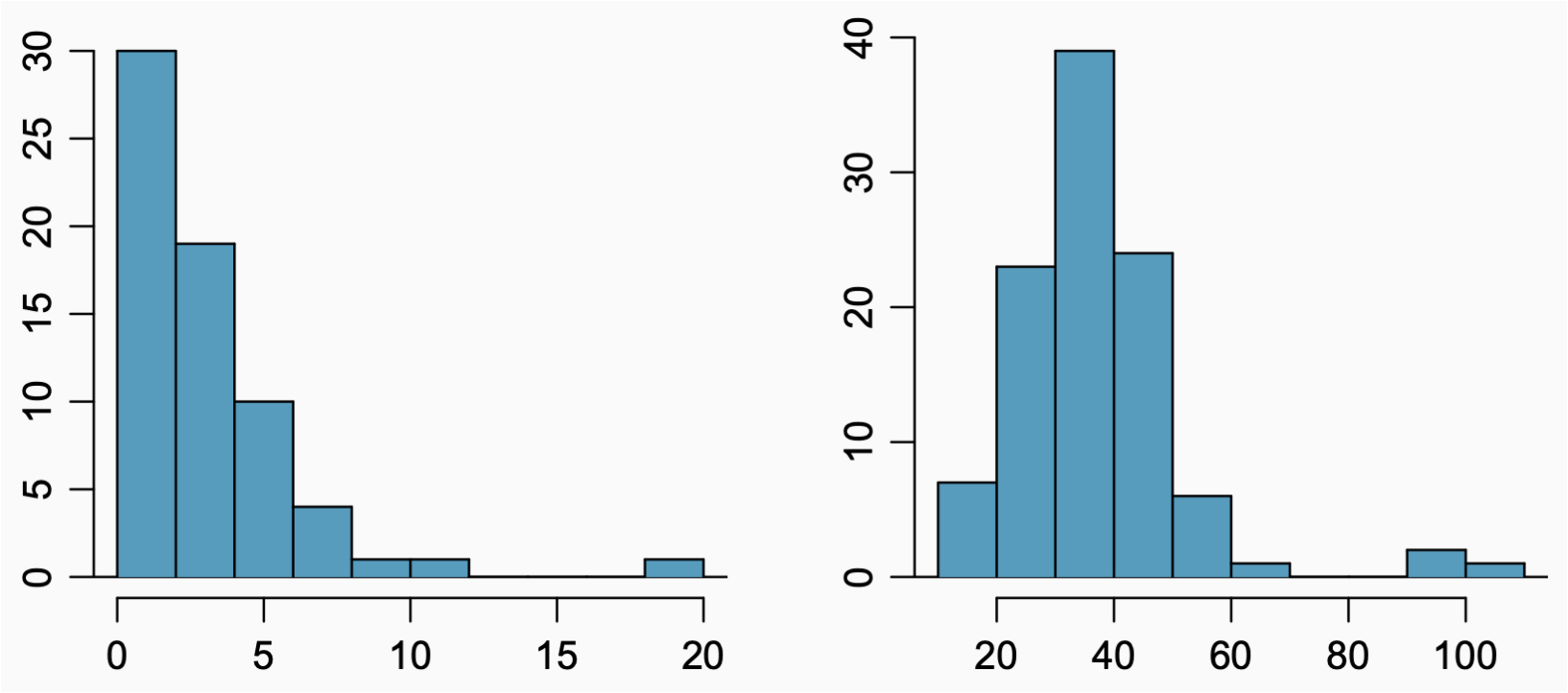
# Side-by-Side Histograms — Infant Mortality Rate Data

- 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

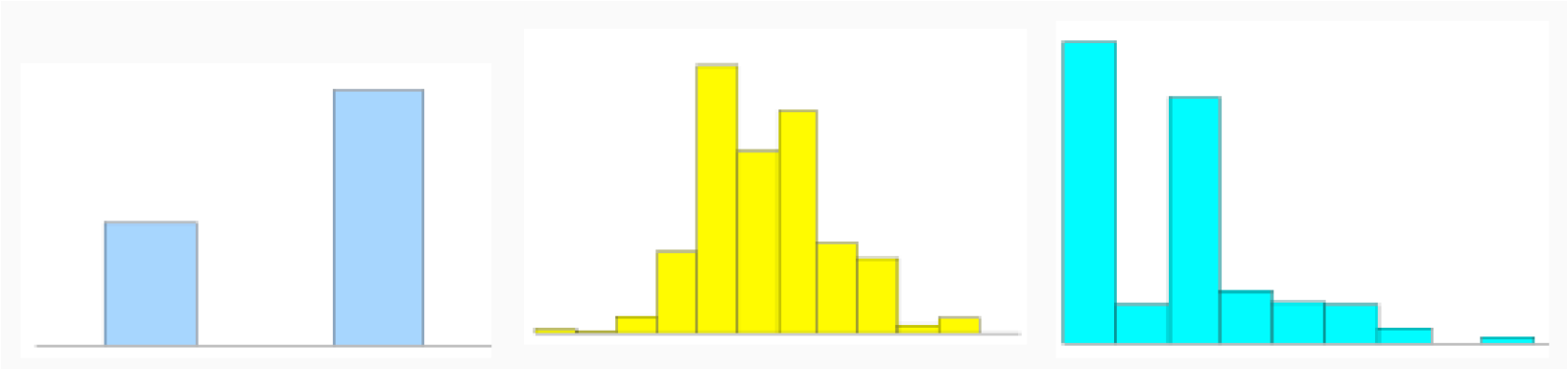
- Match the following variables with the histograms and bar graphs given below. Suppose the data represent students in a course.

  [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



# Answer

(a)The height of students: Heights typically follow a normal distribution because most students will have average heights with fewer students having extremely short or tall heights. Look for a histogram that is bell-shaped, indicating a normal distribution.

(b) Gender breakdown of students: Gender breakdown will typically be represented by a bar graph, not a histogram, showing the count of students in each gender category. Look for a bar graph with distinct bars for different genders.

(c)The number of piercings students have: The number of piercings can be represented by a histogram, likely showing a skewed distribution where most students have zero or few piercings, and fewer students have many piercings. Look for a histogram that shows this kind of skewed distribution.

# Mean and Median

# Numerical Attributes

- 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

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

# Numerical Attributes

- 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 → 4 8 3 5 13

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

data → 4 8 3 5 13 12

# Numerical Attributes

- 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 | → | 4 | 8 | 3 | 5 | 13 |
| sorted | → | 3 | 4 | 5 | 8 | 13 |

The median is 5.

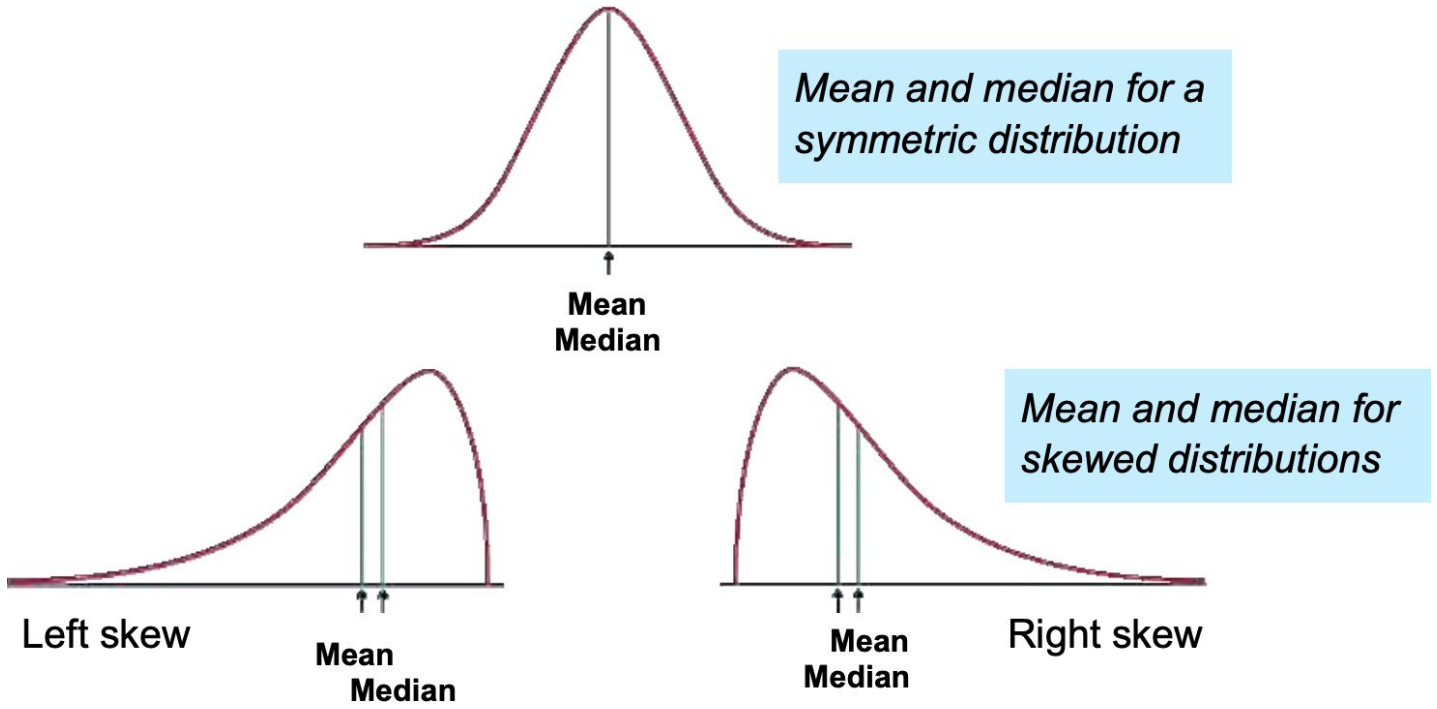
**Ex 2**: Suppose a variable has 5 observed values: 4, 8, 3, 5, 13, 12.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| data | → | 4 | 8 | 3 | 5 | 13 | 12 |
| sorted | → | 3 | 4 | 5 | 8 | 12 | 13 |

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

# Mean vs. Median

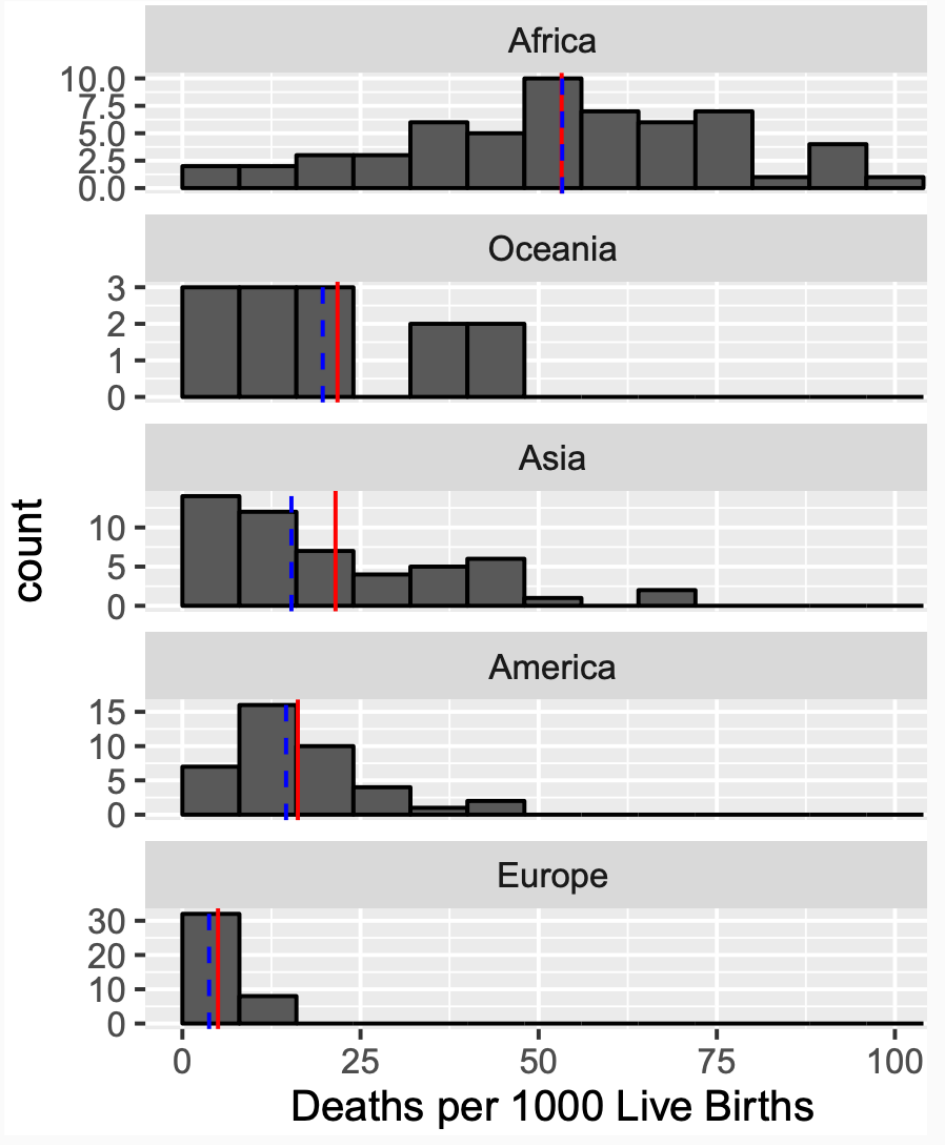
- In a symmetric distribution, mean ≈ median.
  - If exactly symmetric, then mean = median.
- In a skewed distribution, the mean is pulled toward the longer tail



# 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 (20 – 2) / 6 = 3.
  - The median is unaffected

- Median is more *resistant*, 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 |

# Five Number Summary

# Quartiles, IQR, Five-Number Summary

- Quartiles divide data into 4 even parts
  - **first quartile Q1** = 25th percentile:
    25% of data fall below it and 75% above it
  - **second quartile Q2** = median = 50th percentile
  - **third quartile Q3** = 75th percentile
    75% of data fall below it and 25% above it

- **Interquartile Range (IQR)** = Q3 – Q1

- **Five-Number Summary**: min, Q1, Median, Q3, max

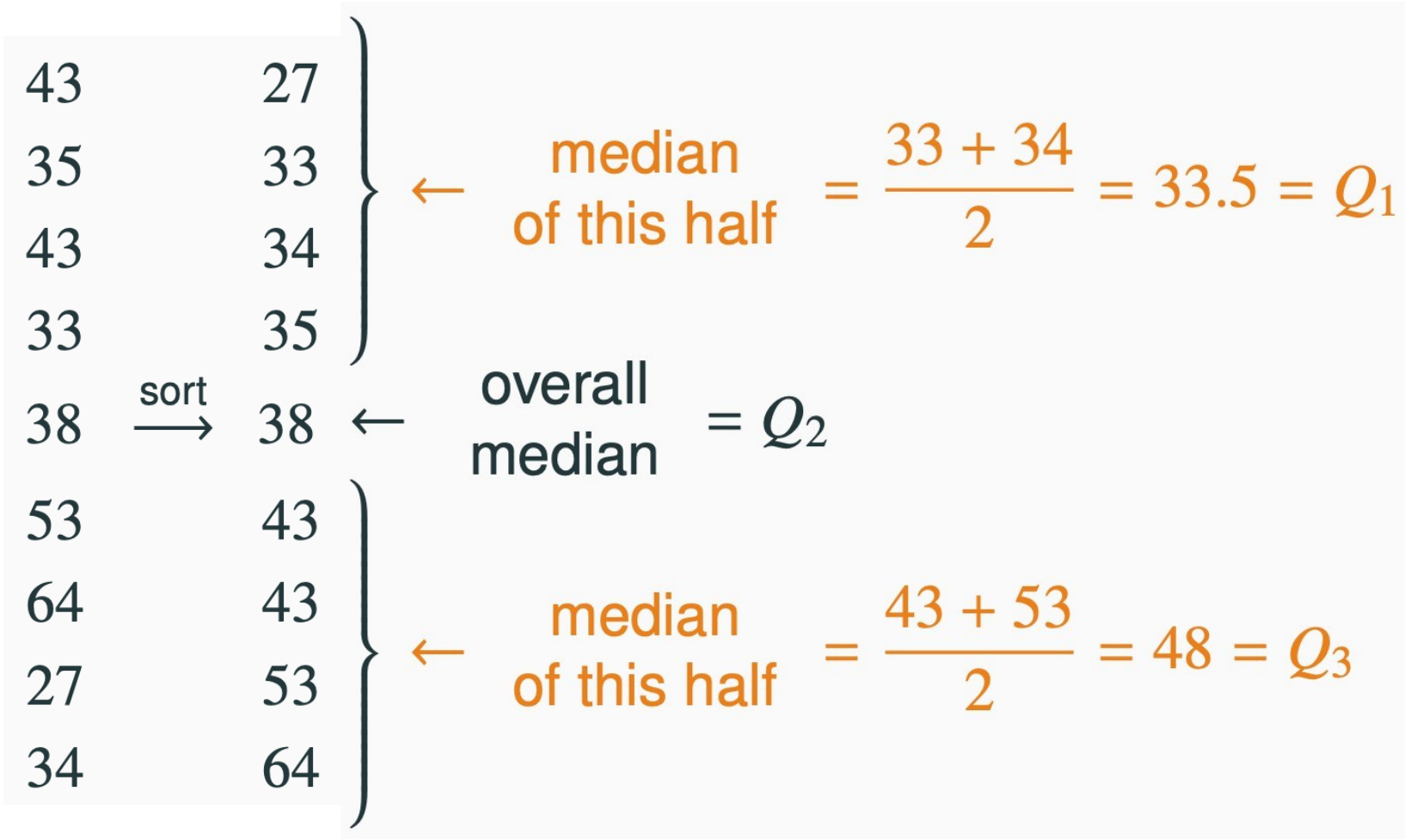
# Example 1

- 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 |

# Example 1

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



# 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 |

IQR = Q3 – Q1 = 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

| | | |
|---|---|---|
| 43 | | 27 |
| 35 | | 27 |
| 43 | | 33 |
| 33 | | 34 |
| 38 | sort ⟶ | 35 |
| 53 | | 38 |
| 64 | | 43 |
| 27 | | 43 |
| 34 | | 53 |
| 27 | | 64 |

IQR = ?

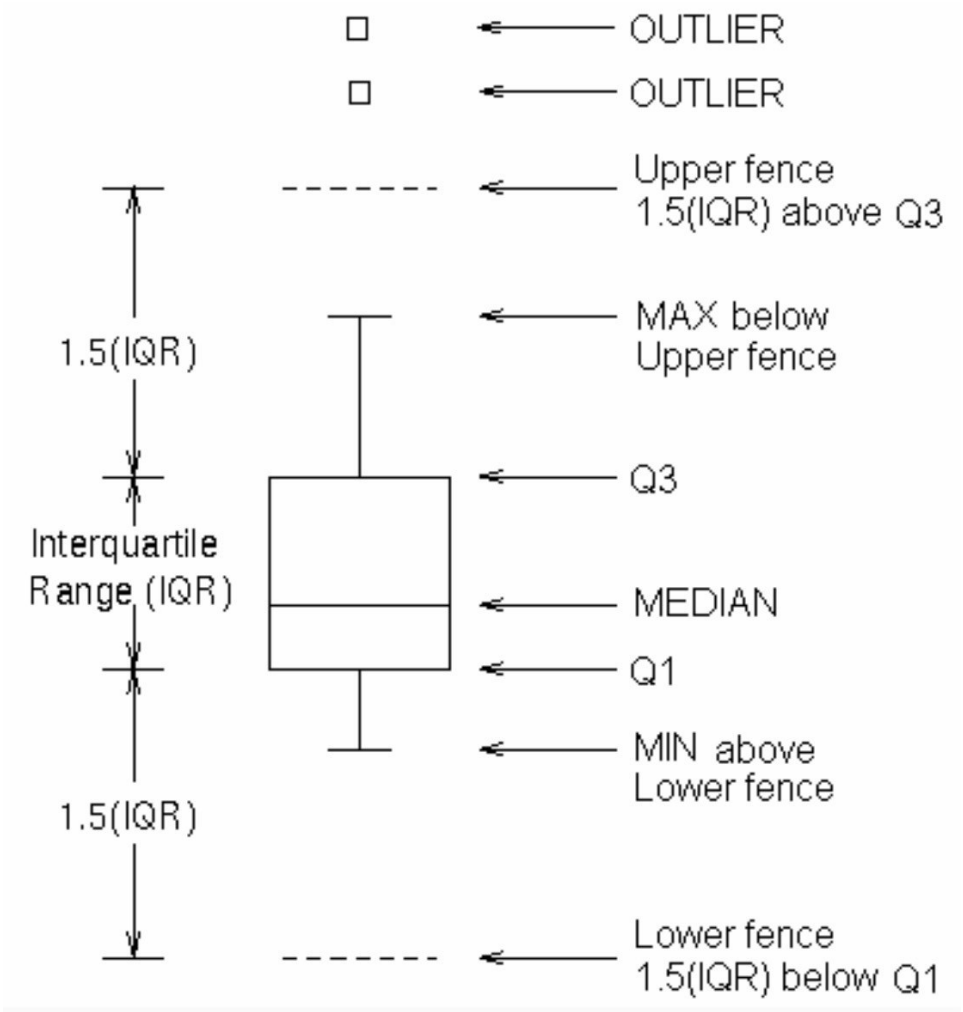
Five number summary: ?

# 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 × IQR below Q1 and above Q3.

# Box-and-Whiskers Plot (also called Boxplot)

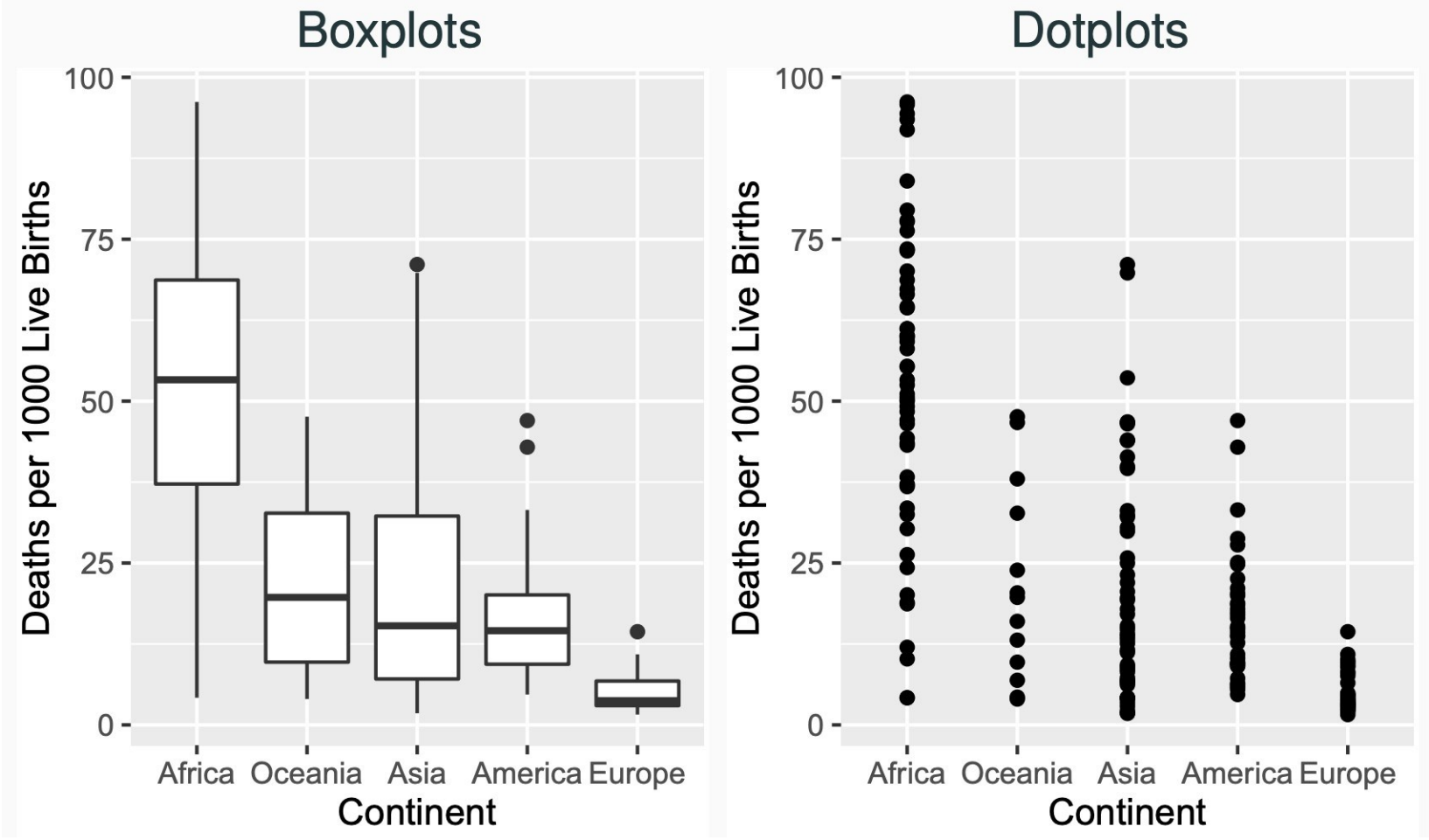


# Think About It ...

- 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

# 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."

- By using standard deviation, data scientists can understand and interpret the spread of data, identify unusual values, and make more accurate predictions and decisions based on data patterns.

# 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 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 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 $n = 4$)

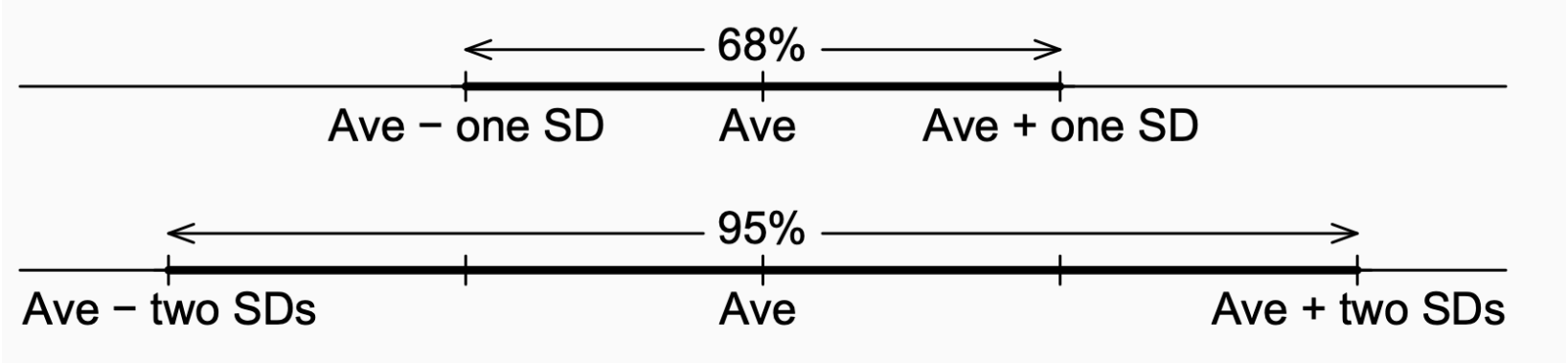
# Variance

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

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

# The 68% and 95% Rule

- 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

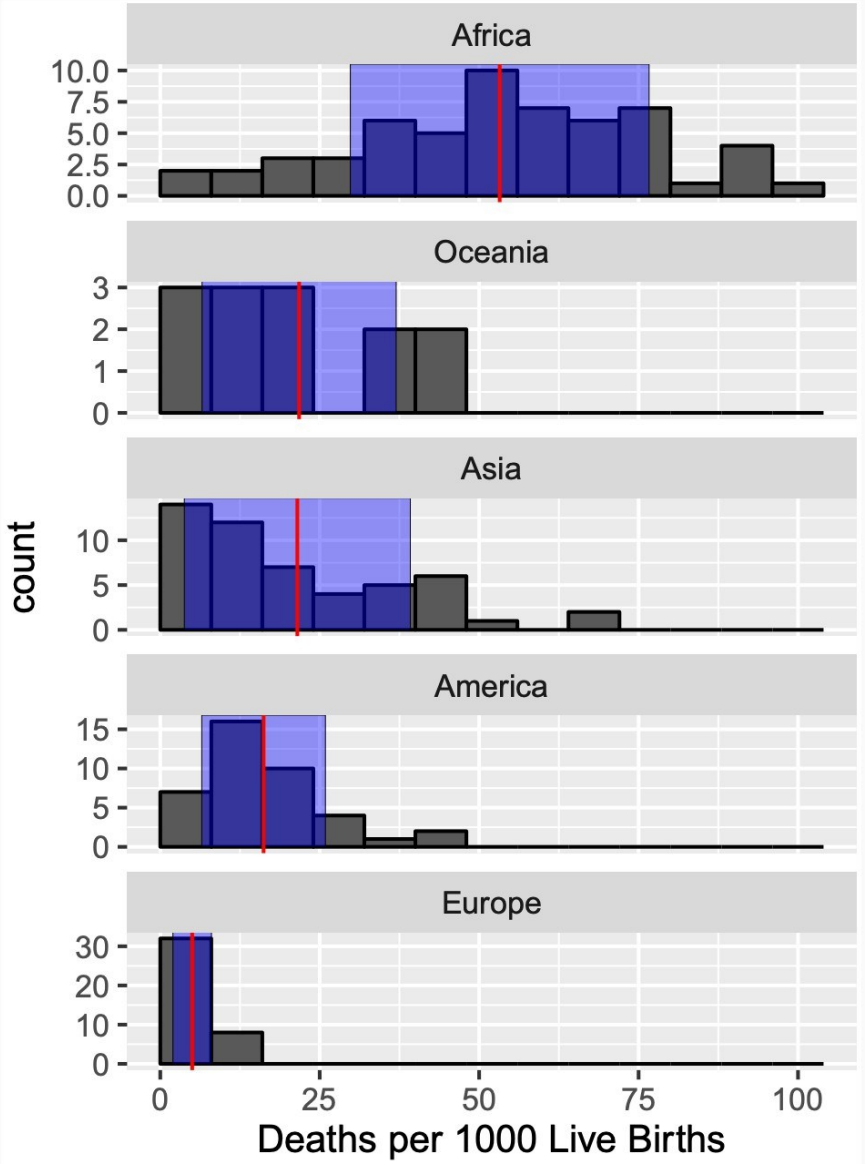



- The 68% and 95% rules work very well for bell-shaped data, and reasonably well for unimodal and not seriously skewed data, but not for all data.

# The 68% and 95% Rule

| 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 |

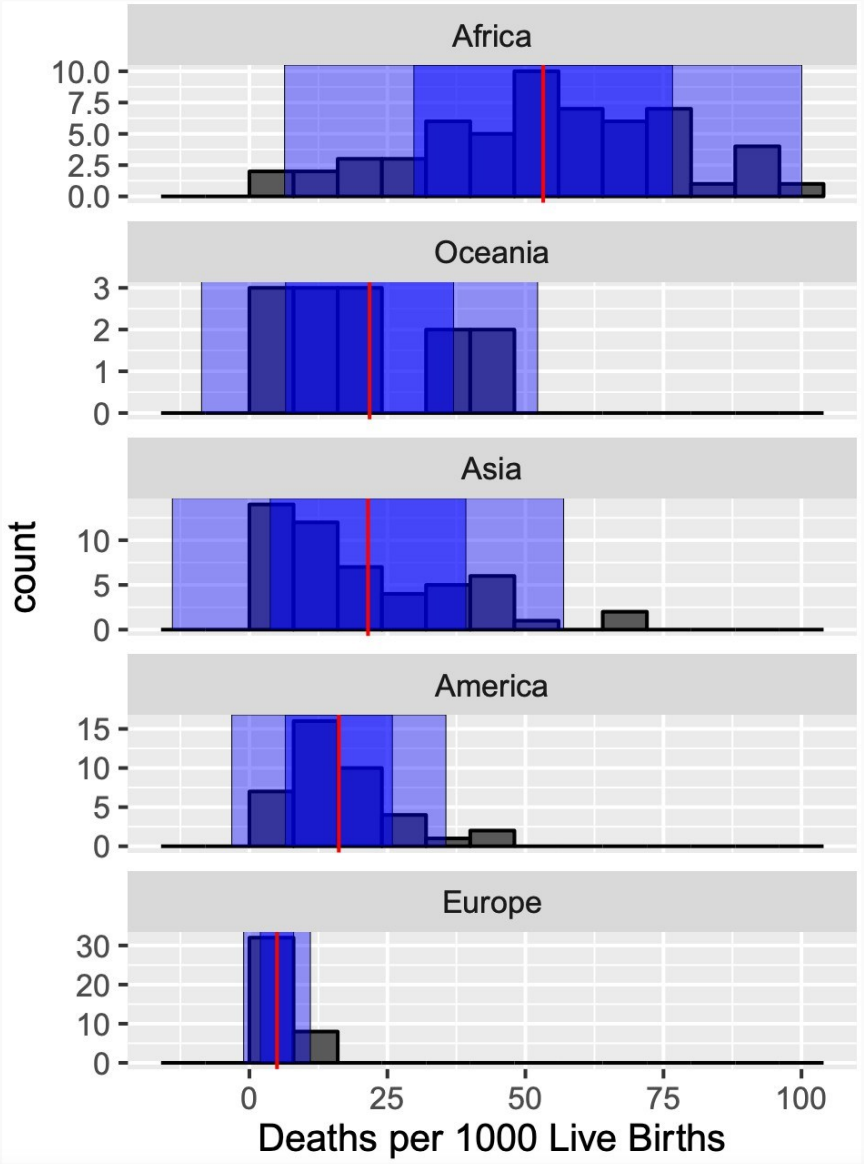
| Continent | proportion within 1 SD from mean |
|---|---|
| Africa | $39/57 \approx 68\%$ |
| Oceania | $8/13 \approx 62\%$ |
| Asia | $35/51 \approx 69\%$ |
| America | $29/40 \approx 72\%$ |
| Europe | $31/40 \approx 78\%$ |



# The 68% and 95% Rule

| 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 | $39/40 = 97.5\%$ |



# Cont. SD

**Understanding Data Spread**:
•Helps to know how much the data varies.
•If you have test scores, it tells you if most scores are close to the average or spread out.

**Identifying Outliers**:
•Data points far from the mean can be considered outliers. Useful in detecting anomalies in data, like fraud detection.

**Comparing Variability**:
•Compare how spread-out different datasets are.
•Useful in comparing different experiments or groups.

# 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 do the observations look like?

- 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 of spread:
  - Range: max − min
  - Standard deviation (SD)
  - Interquartile range (IQR)

# Recap: Common Numerical Summaries of Numerical Variables

- 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!