Day 10 - Quantitative data

Introduction

In this activity, we focus on describing quantitative variables, and practice our coding skill. By the end of the activity, you should understand:

- center
- spread
- mean
- distributions
- median
- skewness
- tail
- mode
- robust
- outliers
- variance
- variability
- standard deviation
- dot plot
- histogram
- box plot
- percentiles
- first quartile
- third quartile
- minimum
- inter-quartile range

```
library(tidyverse)
```

Summarizing numeric data

Note

Numeric data should be summarized in terms of center and spread. (this is most important)

Measures of central tendency

Note

The mean, also called the average, is one way to measure the center of a **distribution** of data.

"x-bar" summation

$$\text{Sample Mean } \bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

μ "mu" population mean

The median is another measure of the center of the data. If you sort the data in ascending or descending order, the median is the middle number, such that half of the data are below the median and half are above the median.

```
nums <- c(1, 3, 5, 6, 7, 8)
mean(nums)
```

```
[1] 5
```

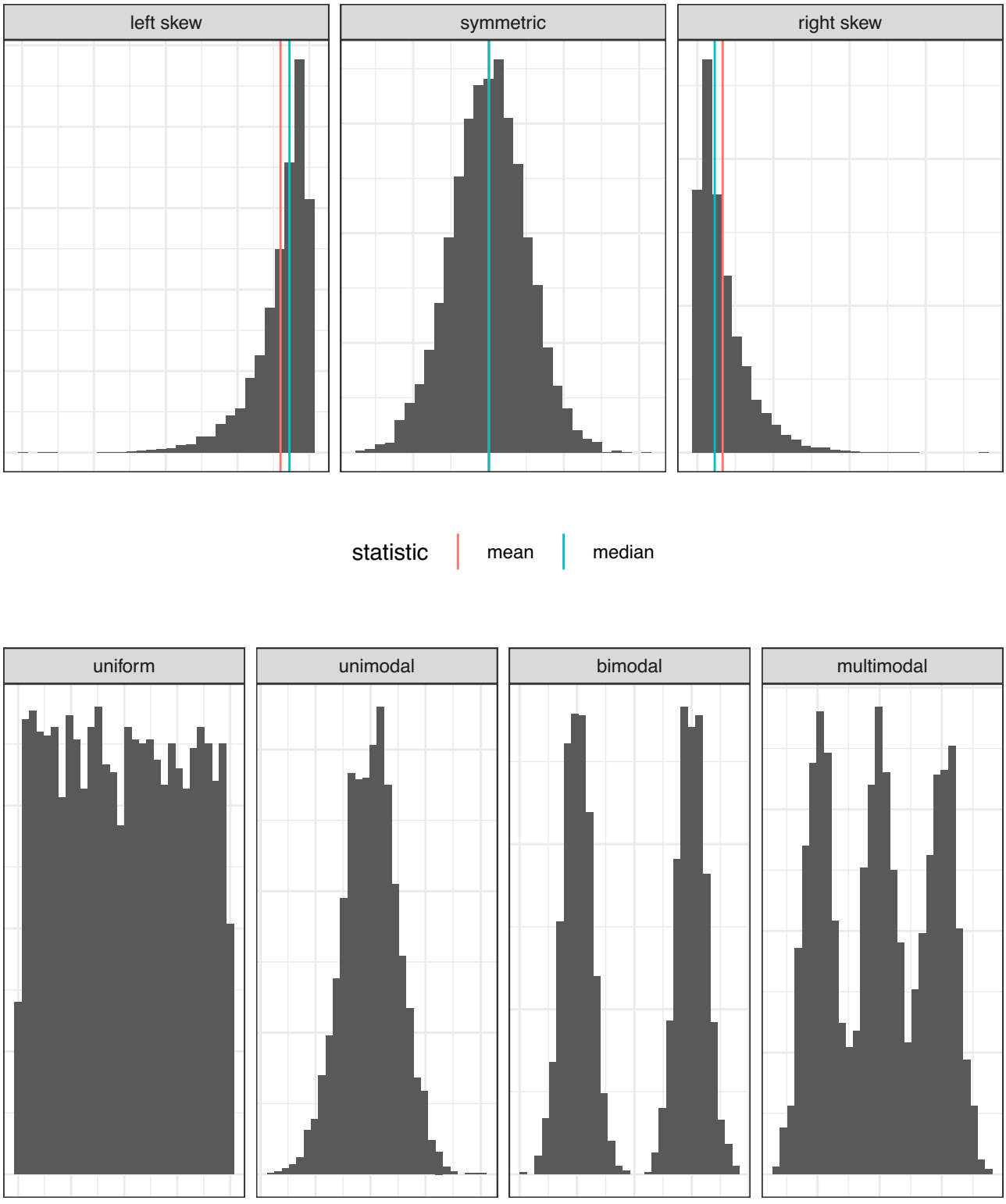
```
median(nums)
```

```
[1] 5.5
```

Skewness and modality

Note

skewness is named based on the direction of the longer **tail**. modality describes the number of prominent peaks in a data set, and one prominent peak is called a **mode**.

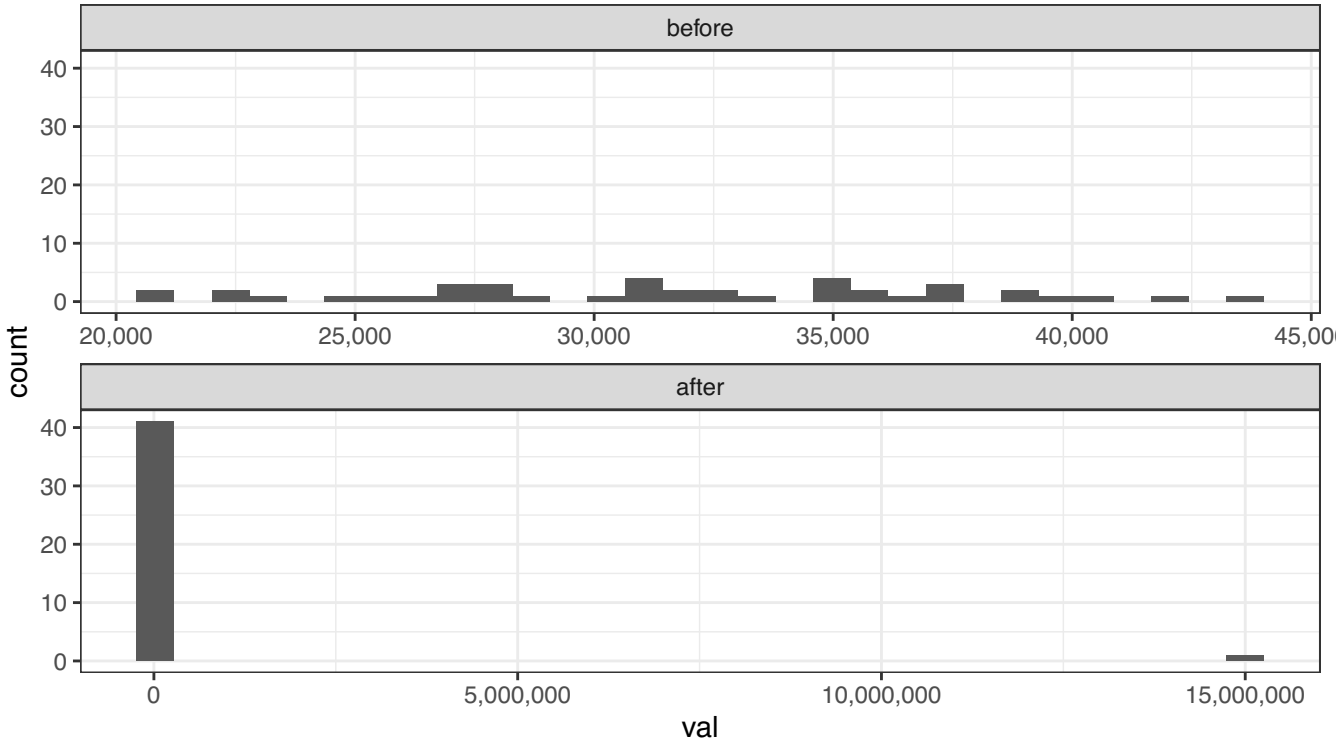

left skew
symmetric
right skew
statistic
mean
median
uniform
unimodal
bimodal
multimodal

Note

The median is robust to the presence of skew and **outliers**, meaning it is less easily impacted than other statistics, such as the mean.

Example: Synthetic data on income at a coffee shop[1]

The first histogram below shows the distribution of the yearly incomes of 40 patrons at Crossroads Cafe in the library. Suppose a new person walks into Crossroads that makes $15,000,000. The second histogram shows the new income distribution. Summary statistics are also provided.



time	n	min	Q1	median	mean	Q3	max	sd	IQR
before	41	20823	26920.0	31554.6	31733.1	35702.2	43618.7	5975.4	8782.2
after	42	20823	27094.9	31642.7	388120.4	36382.7	15000000.0	2309661.3	9287.9

Does the mean or median best represent what we might think of as a typical income for the 41 patrons at the coffee shop? What does this say about the robustness of the two measures?

[1]Adapted from the textbook. These data are not real.

Measures of variability

Note

Variance is roughly interpreted as the average squared deviation from the mean. It is a measure of the variability / spread of the data.

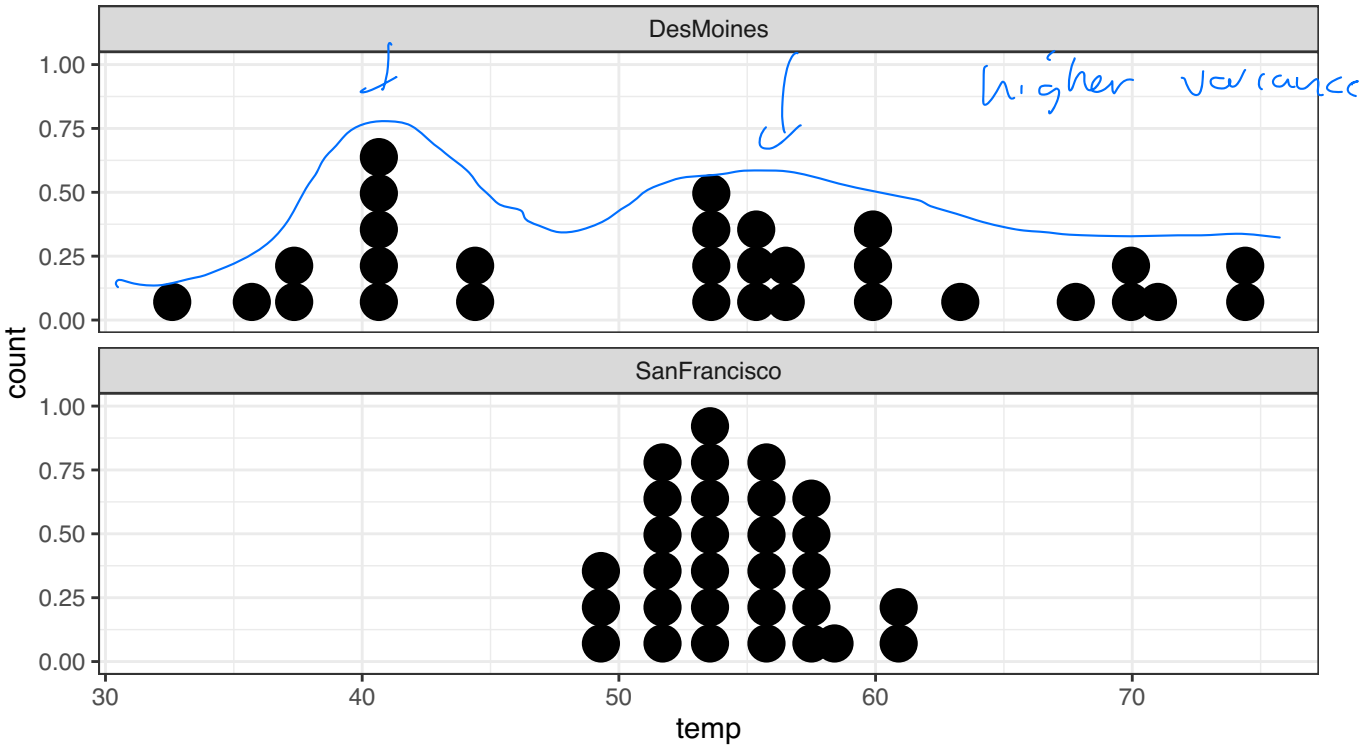
$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$ sample variance

σ^2 population variance "sigma-squared"

The standard deviation is the square root of the variance, and can be thought of as approximately the average deviation from the mean.

sample s.d. $s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

population s.d. σ "sigma"




```
April14Temps %>%
  pivot_longer(DesMoines:SanFrancisco, names_to = "city", values_to = "temp") %>%
  group_by(city) %>%
  summarize(
    mean = mean(temp),
```

```
    median = median(temp),
    sd = sd(temp)
  ) %>%
  knitr::kable()
```

city	mean	median	sd
DesMoines	53.34333	54.5	12.36936
SanFrancisco	54.49000	54.2	3.10665

Note

Percentiles are values such that some proportion of the data are less than that value. For example, the 37^{th} percentile is the value such that 0.37, or 37%, of the data are less than that value.

Some percentiles are used frequently and receive special names: the 25^{th}, 50^{th}, and 75^{th} percentiles are called the **first quartile (Q1)**, median, and **third quartile (Q2)**, respectively.

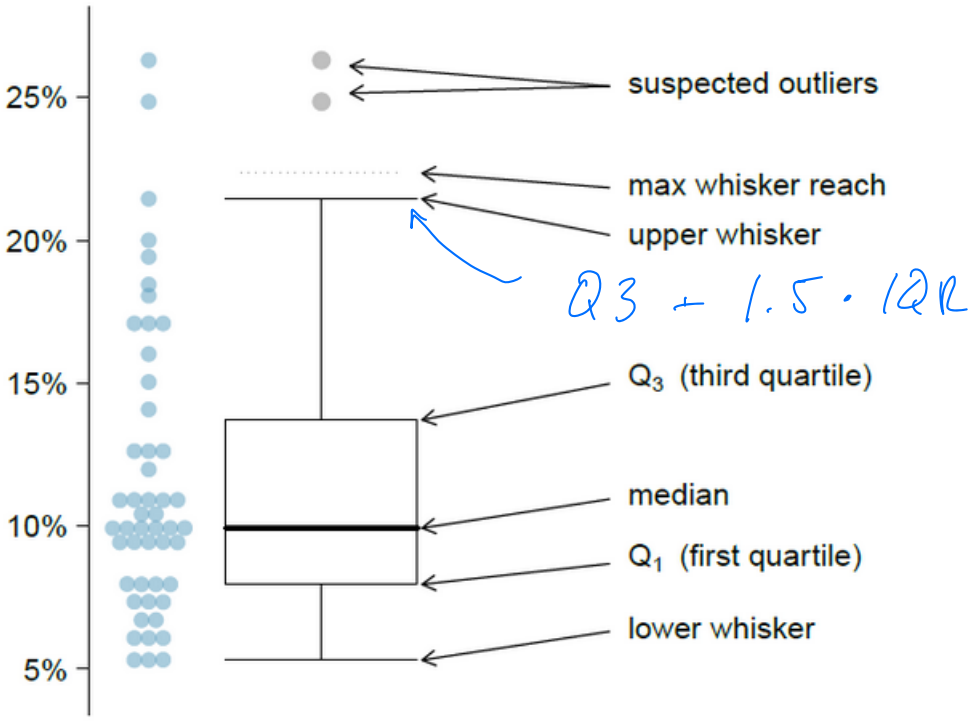
Boxplots

Note

Boxplots are created based on the **five-number summary** of the data.

- **Minimum**: minimum observed value in the data.
- **Q1**: first quartile. The value such that 25% of observations are less than it.
- **Median**: the median. The value such that 50% of observations are less than it.
- **Q3**: the third quartile. The value such that 75% of observations are less than it.
- **Maximum**: maximum observed value in the data.

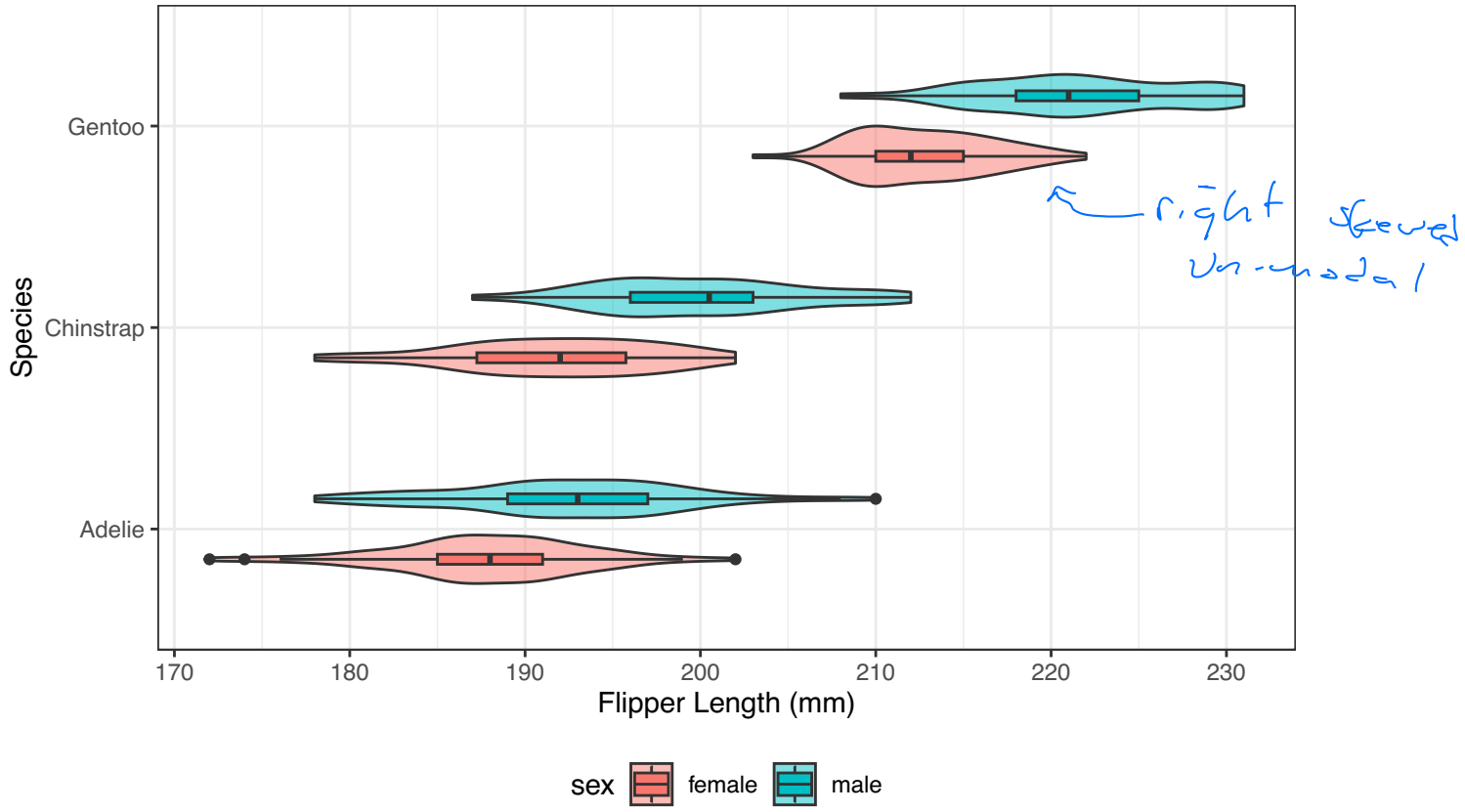
The interquartile range (IQR) is a measure of spread, and represents the length of the box (Q3 - Q1).


25%
20%
15%
10%
5%
suspected outliers
max whisker reach
upper whisker
$Q_3 + 1.5 \cdot IQR$
Q_3 (third quartile)
median
Q_1 (first quartile)
lower whisker

Alternatives to boxplots

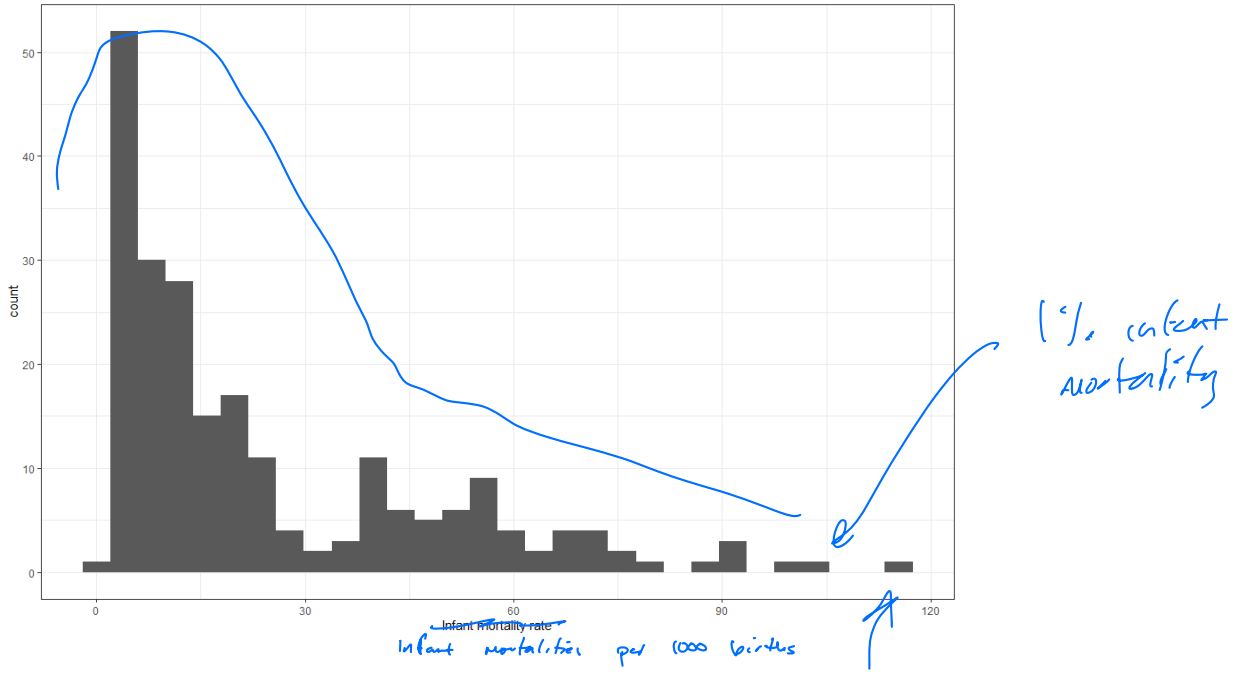
> **Note**
>
> Violin plots are similar to boxplots, but also visualize the density.

```
library(palmerpenguins)
data(penguins)
penguins %>%
  na.omit() %>%
  ggplot(aes(x = species, y = flipper_length_mm, fill = sex)) +
  geom_violin(width = .6, alpha = .5) +
  geom_boxplot(width = .1, position = position_dodge(.6)) +
  theme_bw() +
  coord_flip()+
  theme(legend.position="bottom") +
  labs(
    y = "Flipper Length (mm)", x = "Species"
  )
```



Practice

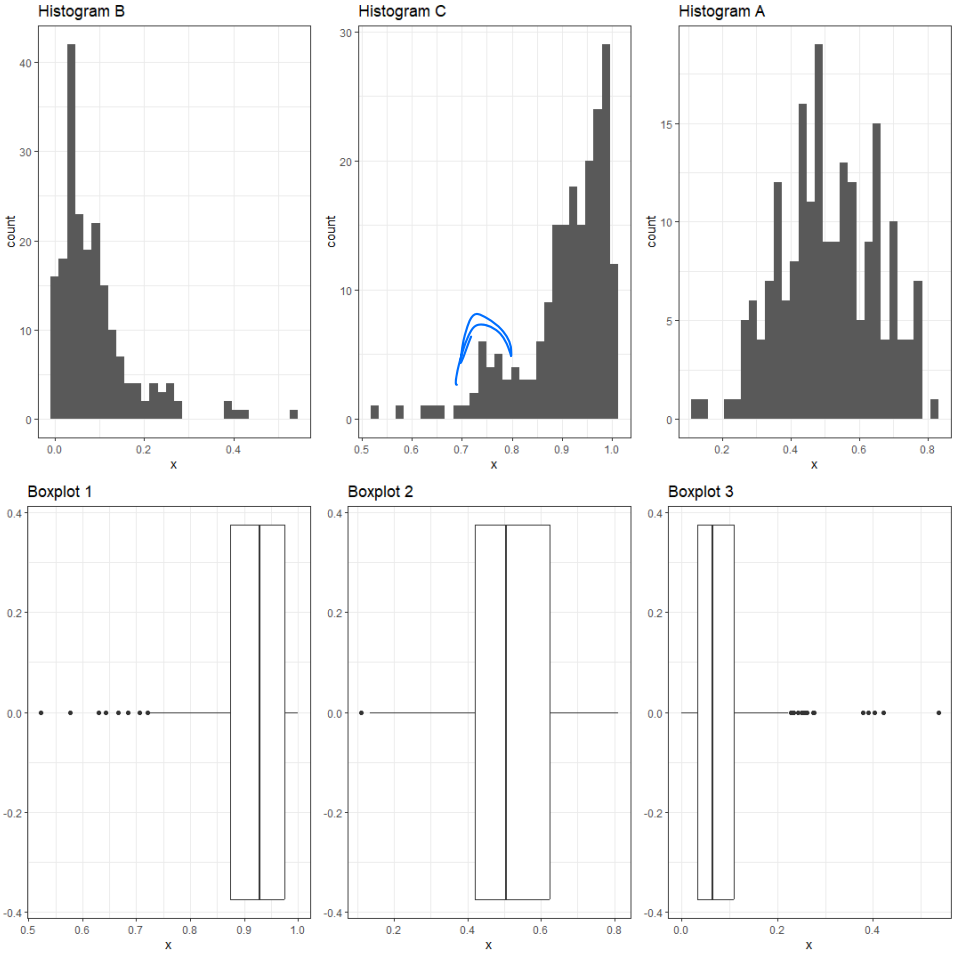
1. The histogram below shows the infant morality rate for for 259 countries in 2014, courtesy of the CIA factbook. The infant mortality rate, defined as the number of infant deaths per 1,000 live births, is often used as an indicator of the level of health in a county.



a) *Estimate* the first quartile, median, and third quartile from the histogram. **Note:** You do not need any code to do this, just give me your best guess based on the plot.

b) Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.

2. For each histogram below, describe the distribution in words, then match the histogram with its corresponding box plot.



- Histogram A:
- Histogram B:
- Histogram C:

3. The histogram and boxplot below were created from the same data source. What characteristics of the distribution are apparent in the histogram but not in the box plot?

