Math 107

Exploratory Data Analysis: One Variable
(Sections 2.1, 2.2, 2.3, 2.4)

Other Measures of Location

Percentiles

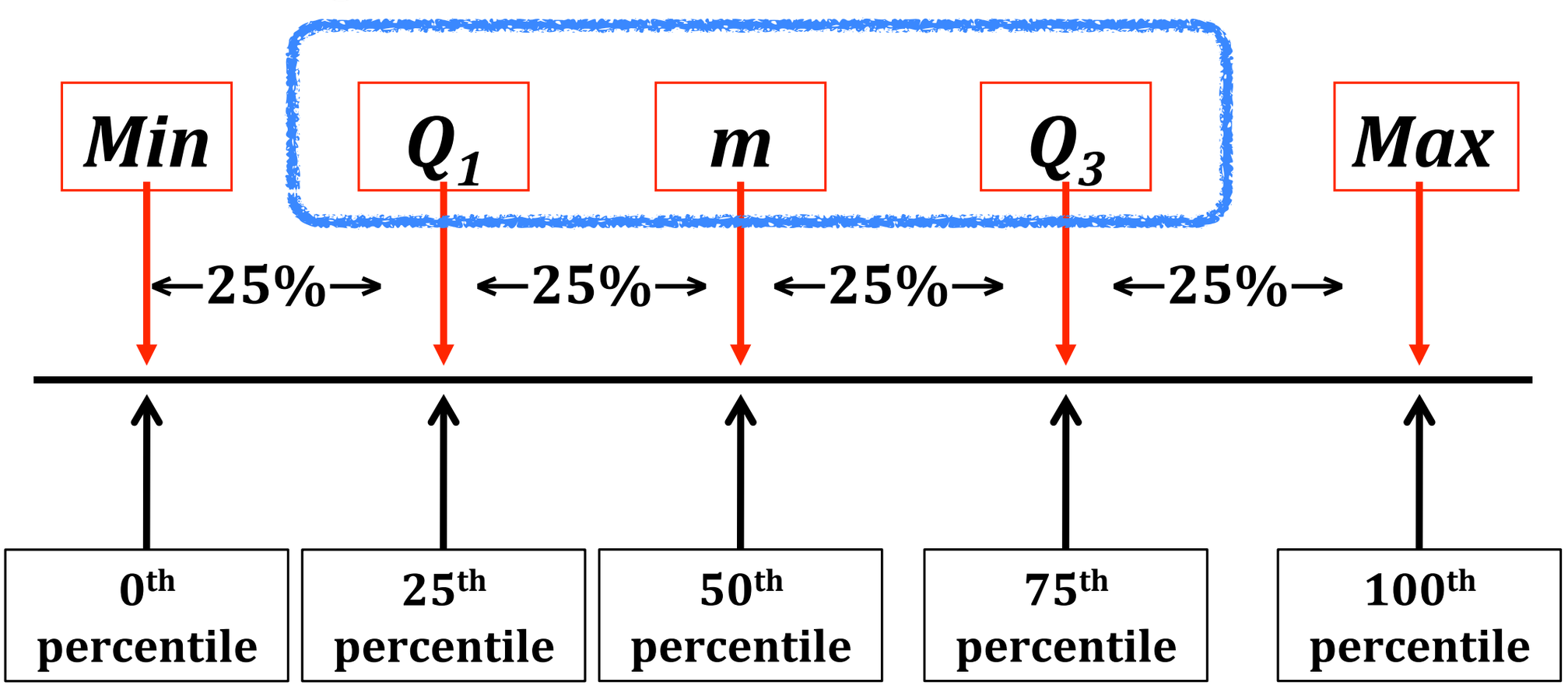
The **P^{th} percentile** if the value which is greater than P% of the data.

The median is the 50^{th} percentile.

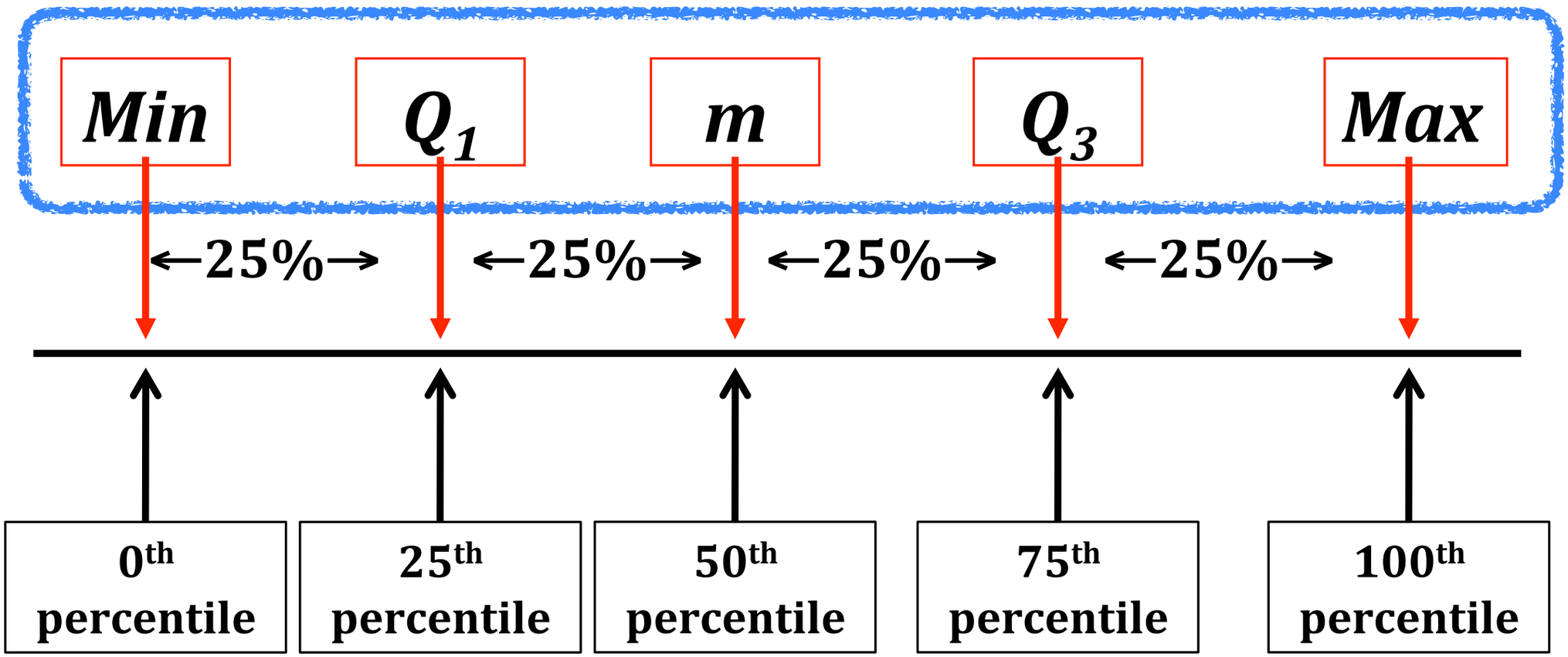
If you scored in the 95^{th} percentile on your ACTs, then you had a better score than 95% of the test takers.

Quartiles

There are three particularly useful percentiles called **Quartiles** that divide the ordered data into four parts, each containing 25% of the data.



Five Number Summary

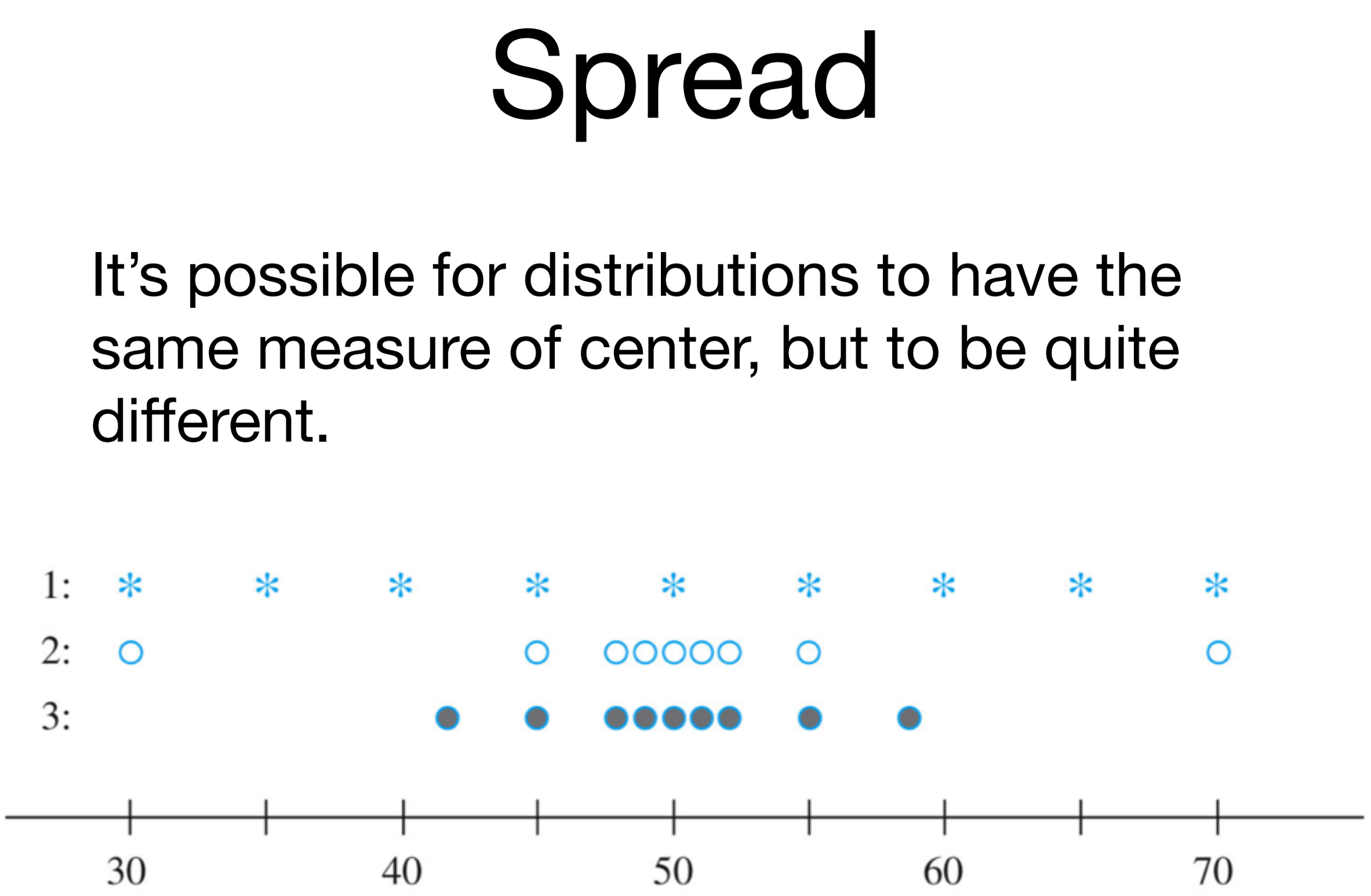


In R: `summary(x)`

Measures of Spread

Spread

It’s possible for distributions to have the same measure of center, but to be quite different.



Statistical Ranges

Range: minimum - maximum

In R: `max(x) - min(x)`

Interquartile range (IQR): range of the central 50% of the data.

In R: `IQR(x)`

Your Turn

Do you think that the range is resistant to outliers? Do you think that the IQR is resistant to outliers?

Turn to the person next to you and discuss for two minutes.

Standard Deviation

The **standard deviation** measures the typical distance of a data value from the mean.

$$\sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}$$

Notation:

Sample standard deviation: s

Population standard deviation: σ

In R: `sd(x)`

Your Turn

Do you think that the standard deviation is resistant to outliers?

The 95% Rule

If a distribution is approximately symmetric and bell shaped, approximately 95% of the data values fall within two standard deviations of the mean.

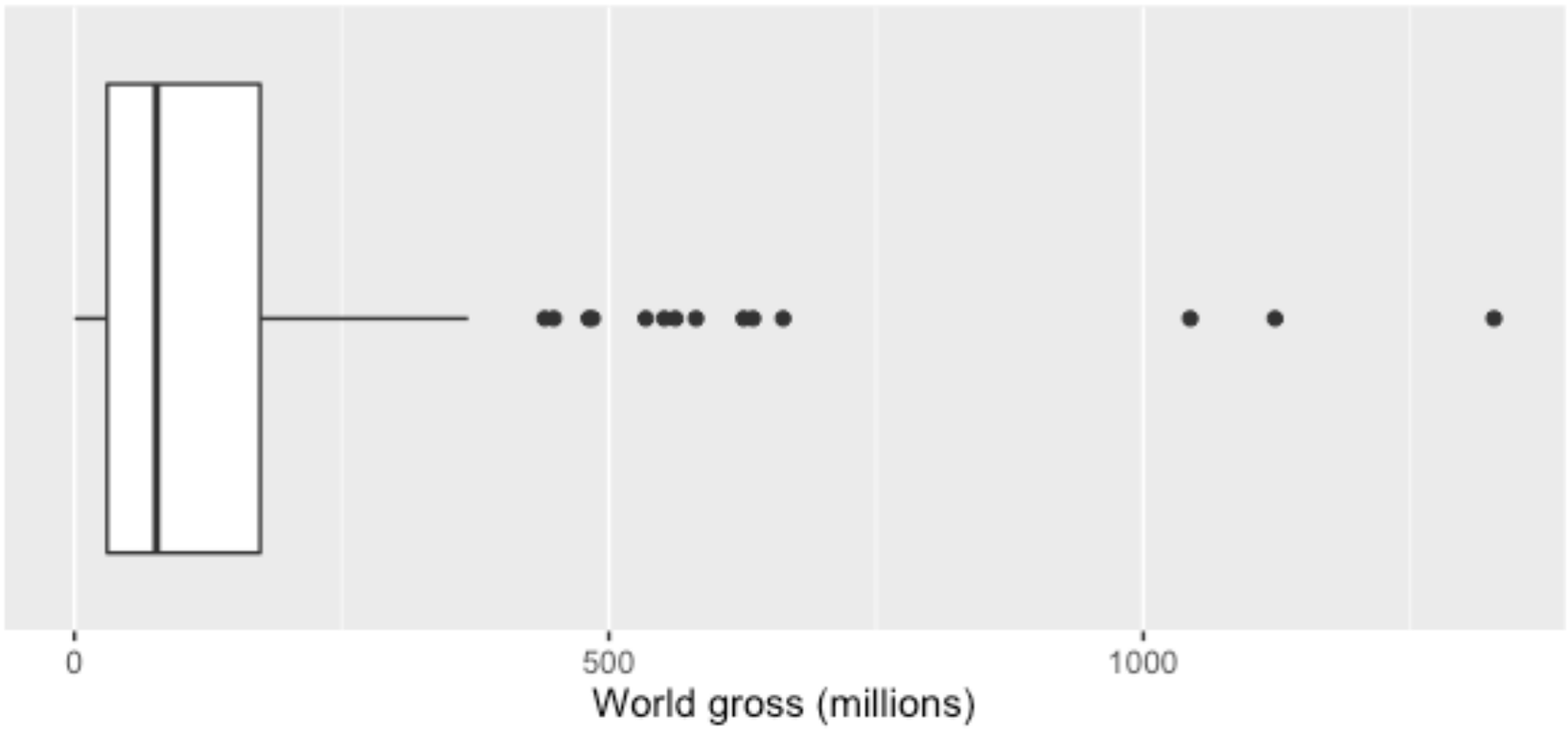
Boxplots

Boxplots

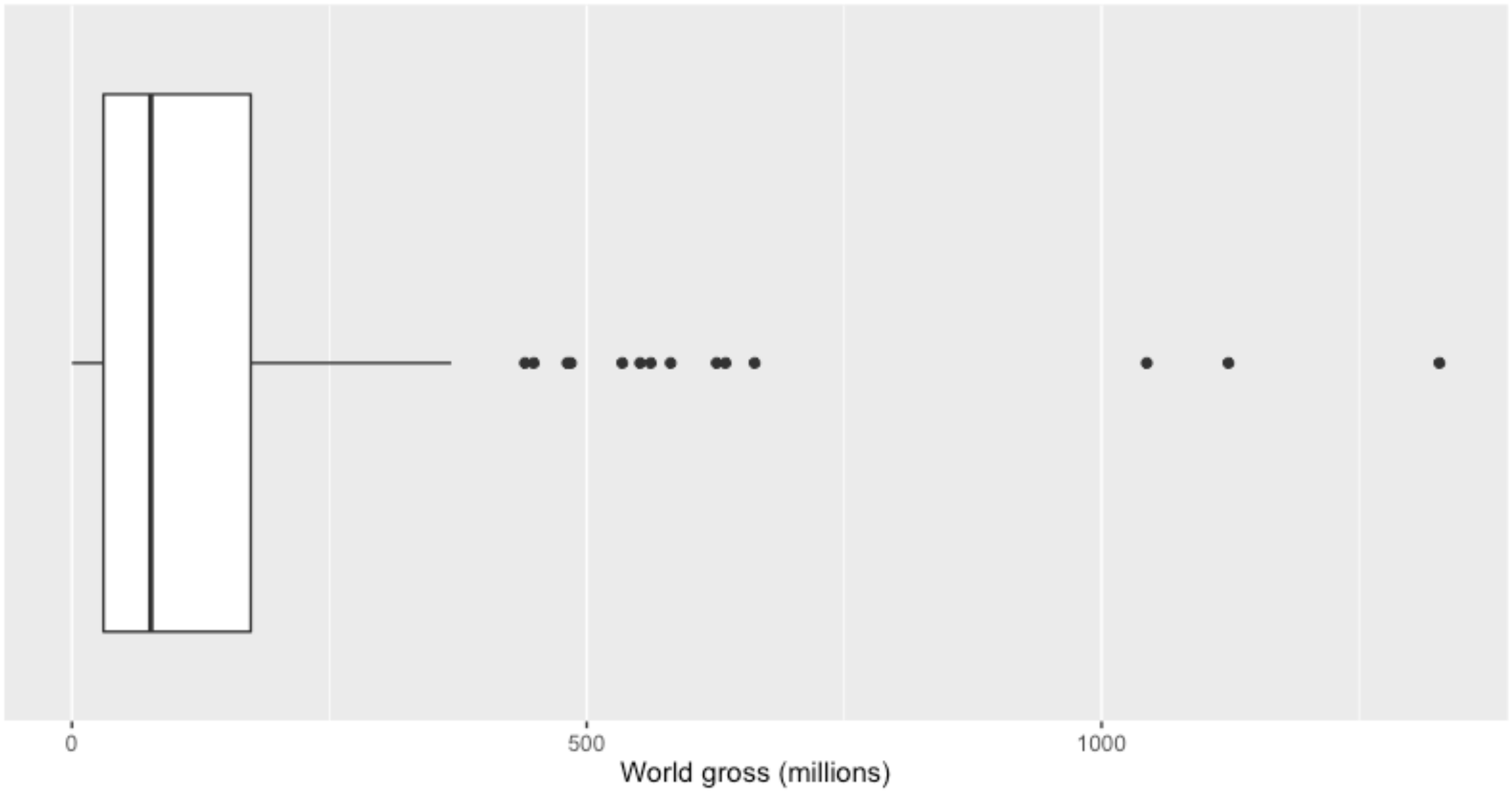
Boxplots visualize the five number summary.

```
summary(HollywoodMovies2011$WorldGross)
```

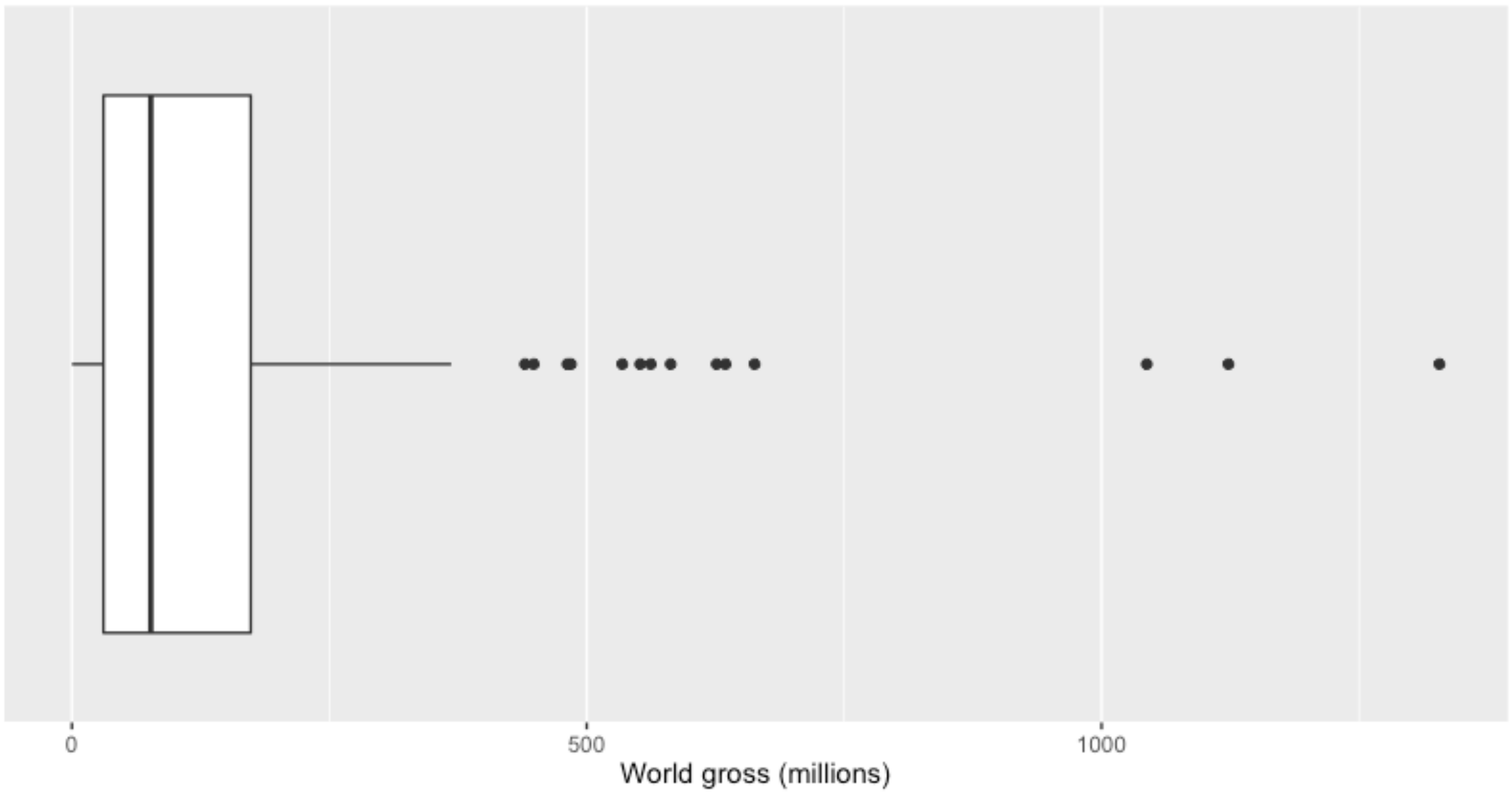
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
   0.025   30.710   76.660  150.700  173.700 1328.000        2
```



Boxplots

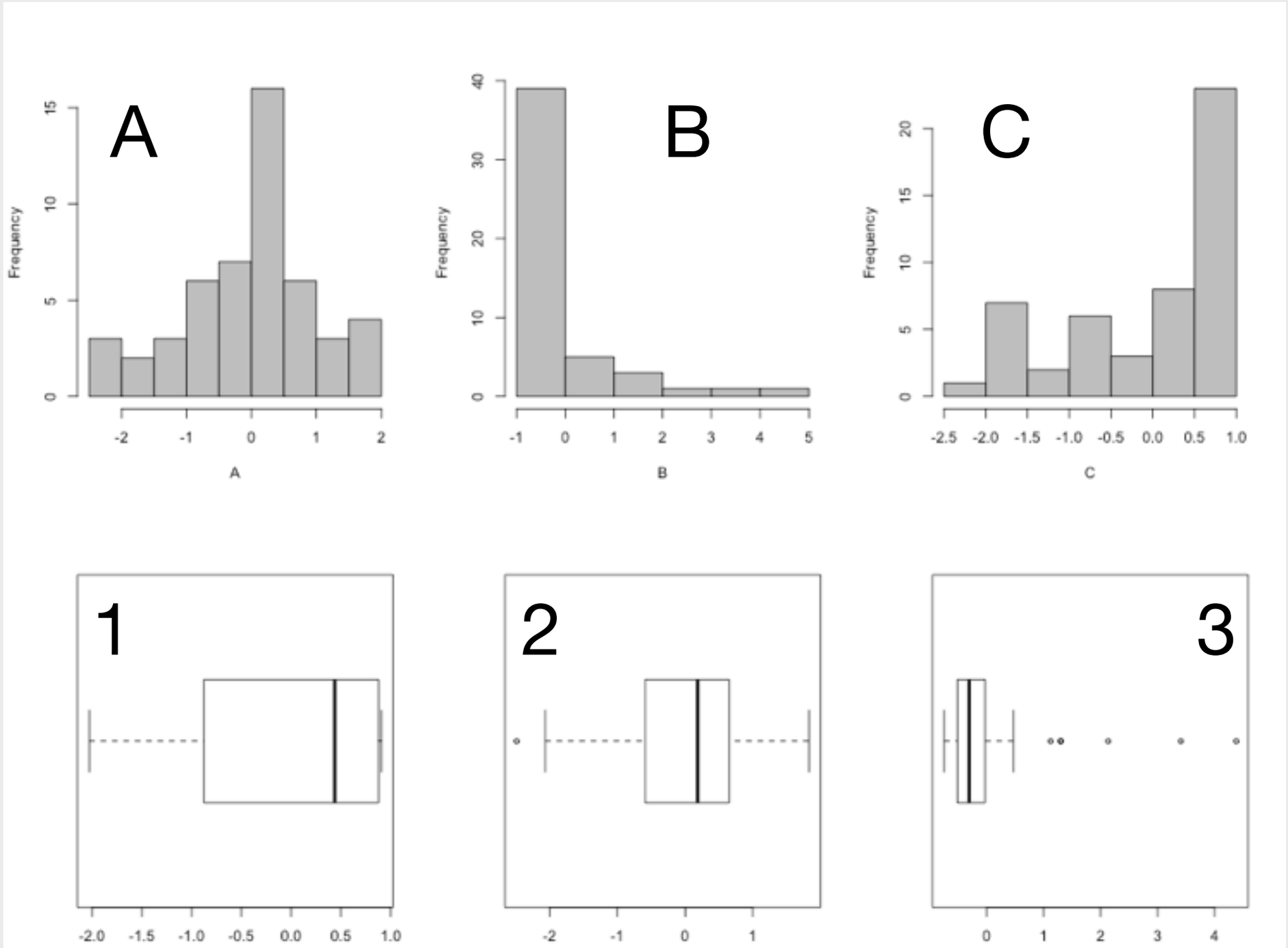


Boxplots



Your Turn

Match the boxplots and histograms.



Standardizing

Your Turn

The average score on the ACT English exam is 21.0 with a standard deviation of 4.0. The average score on the SAT Verbal exam is 520 with a standard deviation of 100.

If Ann scores a 27 on the ACT English exam and Denise scores a 770 on the SAT Verbal exam, who has the better score?

Standardizing

The exams are on different scales, so we must first put the two scores on a common scale.

A **z-score** is the number of standard deviations a data value falls from the mean.

$$\mathbf{z = \frac{x - \overline{x}}{s}}$$

Notation:

- The above is the notation for samples, but we can get the notation for populations by simple substitution.

z-scores

- z-scores have no units
- All observations are on the same scale
  - mean 0
  - standard deviation 1
- Standardizing does not change shape of the distribution.
  - *shifts* location (by subtracting off mean)
  - *rescales* distribution (by dividing by the standard deviation)
- $z < 0 \rightarrow$ data value is *below* the mean
- $z > 0 \rightarrow$ data value is *above* the mean
- The larger the z-score, the more unusual the data value.

Your Turn

The average score on the ACT English exam is 21.0 with a standard deviation of 4.0. The average score on the SAT Verbal exam is 520 with a standard deviation of 100.

If Ann scores a 27 on the ACT English exam and Denise scores a 770 on the SAT Verbal exam, who has the better score?

Your Turn

The average score on the ACT Math exam is 20.7 with a standard deviation of 4.1. The average score on the SAT Math exam is 510 with a standard deviation of 100.

If Jim scores a 15 on the ACT Math exam and Dwight scores a 340 on the SAT Math exam, who has the better score?