Boxplot: Constructing a Boxplot

Learning Objective: Generate and interpret several different graphical displays of the distribution of a quantitative variable (histogram, stemplot, boxplot).

Now that you understand what each of the five numbers means, you can appreciate how much information about the distribution is packed into the five-number summary. All this information can also be represented visually by using the boxplot.

The Boxplot

The boxplot graphically represents the distribution of a quantitative variable by visually displaying the five-number summary and any observation that was classified as a suspected outlier using the 1.5(IQR) criterion.

There are several ways to plot the whiskers on a boxplot. One convention is to plot whiskers down to the minimum and up to the maximum value. We use the 1.5(IQR criterion), also known as the Tukey method for plotting whiskers. First, calculate the IQR, the difference between the 75th and 25th percentiles (or Q3 – Q1). Multiply the IQR by 1.5. Add this value to the 75th percentile. If the value is greater than (or equal to) the maximum value in the dataset, draw the upper whisker to the maximum value. Otherwise, stop the whisker at the largest value that is less than 75th percentile + 1.5 * IQR. Plot any values that are greater than this as individual points that are outliers. Similarly, subtract 1.5 * IQR from the 25th percentile. If this value is smaller than the minimum value in the dataset, draw the lower whisker to the minimum value. Otherwise, stop the whisker at the lowest value that is greater than 25th percentile – 1.5 * IQR. Plot any values that are smaller than this as individual points that are outliers.

Using the Best Actress dataset, here is how we determine where to draw the whiskers:

Q3 = 42

Q1 = 30.5

IQR: 42 – 30.5 = 11.5

1.5 * IQR = 1.5 * 11.5 = 17.25

Q3 + 1.5 * IQR = 42 + 17.25 = 59.25

The largest observation that is less than or equal to 59.25 is 49 so we draw the upper whisker up to 49. All points above 49 are considered outliers (61, 61, 62, 74, 80).

Q1 – 1.5 * IQR = 30.5 – 17.25 = 13.25

The smallest observation that is greater than or equal to 13.5 is 21 so we draw the lower whisker down to 21, which is also the minimum. There are no outliers.

The following video illustrates how a boxplot is constructed: (this is for the "Best Actress" dataset—to see the dataset, click here).

Constructing a Boxplot

Start of transcript. Skip to the end.

In this walkthrough, we'll construct a boxplot for the Best Actress data set.

The boxplot has several main features: the central box, a line in the middle of the box,

two lines that extend away from the box on either side, and finally,

any outliers that might be present in the data set. Let's take a closer look at each of these.

Video

Download video file

Transcripts

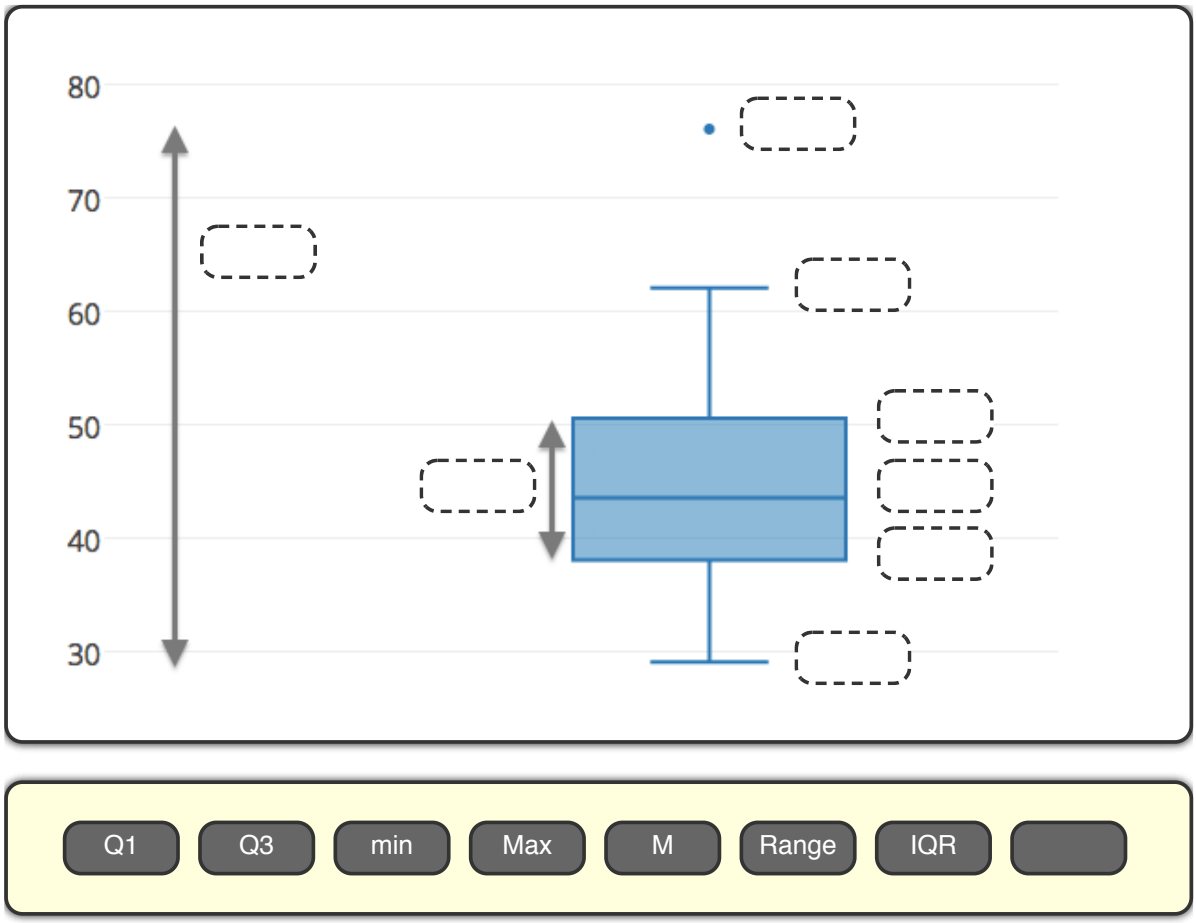
Download SubRip (.srt) file

Download Text (.txt) file

Learn By Doing

The boxplot below displays the ages of Best Actor Oscar winners (1970-2013). (To see the full dataset, click here.) In the last activity, you found numerical measures for this distribution.

Now try this drag-and-drop exercise. Label the different numerical measures as they are depicted in the boxplot.

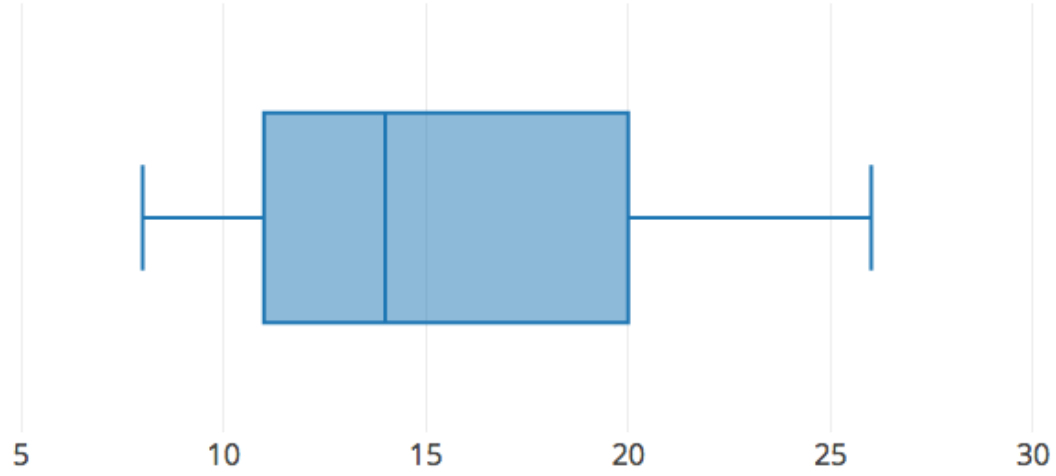


Explore this simulation activity about calculating the median, Q1, Q3, IQR, outliers, and drawing a boxplot. Note that you can edit the data in the chart to see different results.

Did I Get This

1/1 point (graded)

The boxplot below displays ratings for TV shows during sweeps week. Don't worry that it's oriented horizontally instead of vertically - boxplots can be presented either way. You just need to read the data on the x-axis instead of the y-axis.



50% of the shows have a rating greater than which value?

- 11
- 15.5
- 14 ✔
- 17

Answer
Correct:
50% of the shows have a rating greater than the median (M), which, according to the boxplot, is 14.

Submit

Did I Get This

1/1 point (graded)
What is the percentage of shows that have a rating of less than 20?

- 67%

- (●) 75% ✔
- ( ) 84%
- ( ) 25%

Answer
Correct:
The boxplot indicates that 20 is the third quartile (Q3) of the rating distribution, and therefore, by definition, three quarters of the observations, or 75%, of the shows have a rating below it.

Submit

Did I Get This

1/1 point (graded)
Within which interval of rating would you expect to find the largest number of shows?

- ( ) 5-10
- (●) 10-15 ✔
- ( ) 15-20
- ( ) 20-25
- ( ) All are equal

Answer
Correct:
. Note that 10 is below Q1, and 15 is above the M, so we expect more than 25% of the shows to have ratings in that interval. Using a similar reasoning, in all the other intervals we would expect to see LESS than 25% of the shows.

Submit