Five-Number Summary and Box Plots

Five-Number Summary and Box Plots

- What are the values in the five-number summary?
- What is the purpose of a Box plot?
- What is an outlier?

Median

- The median is the physical middle value for the distribution *when the data values are in numerical order.*

Median

- The median is the physical middle value for the distribution *when the data values are in numerical order.*

- The median is represented using
  - $\tilde{x}$

or

  - Q_2.

Median

- The median is the physical middle value for the distribution *when the data values are in numerical order.*

- The median divides the distribution into two halves.

Median

- The median is the physical middle value for the distribution *when the data values are in numerical order.*

- The median divides the distribution into two halves.
  - Lower half of the distribution
  - Upper half of the distribution

Median

- The median is the physical middle value for the distribution *when the data values are in numerical order.*

- The median separates the
  - Lower 50% of the distribution

  from the
  - Upper 50% of the distribution

Median

- The median is the physical middle value for the distribution *when the data values are in numerical order.*

Caution: The median is the *physical* middle value for the distribution.

Determining the Median

- Arrange the data in numerical order
- Determine the number of data values
- Count off data values from one end to the middle value

Determining the Median

- If there are an odd number of data values, the median is the middle data value

- If there are an even number of data values, the median is the average of the middle two data values.

Determining the Median

- If there are an odd number of data values, the median is the ***middle data value***

- If there are an even number of data values, the median is the ***average of the middle two data values***.

When should the Median be Used?

When should the Median be Used?

- The median should be used when a distribution is
  - Skewed-left
  - Skewed-right
  - Not symmetric around a central value

The Quartiles

- When determining the median, we divide the distribution into halves
  - Dividing each of these halves in half,
    - we determine the first or lower quartile, Q_1,

and

    - the third or upper quartile, Q_3, for the distribution.

The Quartiles

- The first or lower quartile, Q_1, is the median of the lower half of the distribution.
- The third or upper quartile, Q_3, is the median for the upper half of the distribution.

The Quartiles

- The first or lower quartile, Q_1, is the *median of the lower half of the distribution*.
- The third or upper quartile, Q_3, is the *median for the upper half of the distribution*.

The Quartiles

- The median, the first or lower quartile, and the third or upper quartile divide the distribution into quarters.

The Quartiles

- The median, the first or lower quartile, and the third or upper quartile divide the distribution into quarters.
- That is, the median, Q_1, and Q_3 divide the distribution into four pieces (i.e. into *fourths*).

The Quartiles

- The median, the first or lower quartile, and the third or upper quartile divide the distribution into quarters.
- That is, the median, Q_1, and Q_3 divide the distribution into four pieces (i.e. into *fourths*).
- Note: This makes the Q_2 notation for the median make sense.

The Quartiles

- The median, the first or lower quartile, and the third or upper quartile divide the distribution into quarters.
- That is, the median, Q_1, and Q_3 divide the distribution into four pieces (i.e. into *fourths*).
- Note: The median, Q_2, is also known as the *second quartile*.

The Quartiles

- The median, the first or lower quartile, and the third or upper quartile divide the distribution into quarters.
- That is, Q_1, Q_2, and Q_3 divide the distribution into four pieces (i.e. into *fourths*).

The Quartiles

- The median, the first or lower quartile, and the third or upper quartile divide the distribution into quarters.
- That is, Q_1 (first quartile), Q_2 (median or second quartile), and Q_3 (third quartile) divide the distribution into four pieces (i.e. into *fourths*).

The Quartiles

- The median, the first or lower quartile, and the third or upper quartile divide the distribution into quarters.
- That is, Q_1 (lower quartile), Q_2 (median or second quartile), and Q_3 (upper quartile) divide the distribution into four pieces (i.e. into *fourths*).

Interquartile Range

- The interquartile range, denoted IQR, is a measure of spread from the lower quartile to the upper quartile,

$$IQR = Q_3 - Q_1$$

Interquartile Range

- The interquartile range, denoted IQR, is the difference between the upper quartile and the lower quartile,

$$IQR = Q_3 - Q_1$$

Interquartile Range

- The interquartile range, denoted IQR, is the difference between the third quartile and the first quartile,

$$IQR = Q_3 - Q_1$$

Five-Number Summary

- Minimum – The smallest value
- Lower or first quartile, Q_1 – the median of the lower half of the values
- Median – the values that divides the data into halves
- Upper or third quartile, Q_3 – the median of the upper half of the values
- Maximum – the largest value

Five-Number Summary

- Five-number summary is represented as
  - *A* set

  or
  - *A* table with an in-context title

Five-Number Summary

- Five-number summary is represented as
  - A set
    {minimum, Q_1, Q_2, Q_3, maximum}

  or
  - A table with an in-context title

Outlier

- A value is considered to be an outlier if it is *either*
  - more than 1.5 times the interquartile range, IQR, less than the lower quartile, Q_1,

  or
  - more than 1.5 times the interquartile range, IQR, greater than the upper quartile, Q_3.

Outlier

- *More than 1.5 times the interquartile range, IQR, from the nearest quartile* means that
  - An outlier is less than
    $Q_1 - 1.5 \cdot IQR$

  or
  - An outlier is more than
    $Q_3 + 1.5 \cdot IQR$

Outlier

- *To determine the outliers, we calculate the lower fence, L_f, and the upper fence, U_f:*
  - The lower fence is
    $$L_f = Q_1 - 1.5 \cdot IQR$$
  - The upper fence is
    $$U_f = Q_3 + 1.5 \cdot IQR$$

Outlier

- *An outlier is less than the lower fence*
  - $L_f = Q_1 - 1.5 \cdot IQR$

or

- *An outlier is greater than the upper fence*
  - $U_f = Q_3 + 1.5 \cdot IQR$

Outlier

- *An outlier is less than the lower fence*
  - $L_f = Q_1 - 1.5 \cdot IQR$

or

- *An outlier is greater than the upper fence*
  - $U_f = Q_3 + 1.5 \cdot IQR$

Note: The values of L_f and U_f are *never* approximated.

Box Plot
(a.k.a. Box and Whiskers Plot)

- A box plot is a graphical display of the five-number summary for a data set.
  - Minimum
  - Q_1
  - Median
  - Q_3
  - Maximum

Box Plot
(a.k.a. Box and Whiskers Plot)

- To make a Box plot,
  - Create a horizontal axis
  - Mark reasonably, equally-spaced tick marks along the axis
  - Mark the value of the scale on the tick marks

Box Plot
(a.k.a. Box and Whiskers Plot)

- To make a Box plot,
  - Create a horizontal axis
  - Mark reasonably, equally-spaced tick marks along the axis
  - Mark the value of the scale on the tick marks

  Note: The tick marks increase incrementally in value based on the value of the scale

Box Plot
(a.k.a. Box and Whiskers Plot)

- To make a Box plot,
  - Create a horizontal axis
  - Mark reasonably, equally-spaced tick marks along the axis
  - Mark the value of the scale on the tick marks
  - Label the horizontal axis with the variable and its units of measure

Box Plot
(a.k.a. Box and Whiskers Plot)

- To make a Box plot,
  - Create a horizontal axis
  - Mark reasonably, equally-spaced tick marks along the axis
  - Mark the value of the scale on the tick marks
  - Label the horizontal axis with the variable and its units of measure

Note: There is no vertical axis.

Box Plot
(a.k.a. Box and Whiskers Plot)

- To make a Box plot,
  - Make a rectangle parallel to the horizontal axis that starts at Q_1 and ends at Q_3
  - Mark the median Q_2 in the middle of the rectangle parallel to the ends of the rectangle
  - Make "whiskers" that extend from each quartile to the adjacent extreme value

Box Plot
(a.k.a. Box and Whiskers Plot)

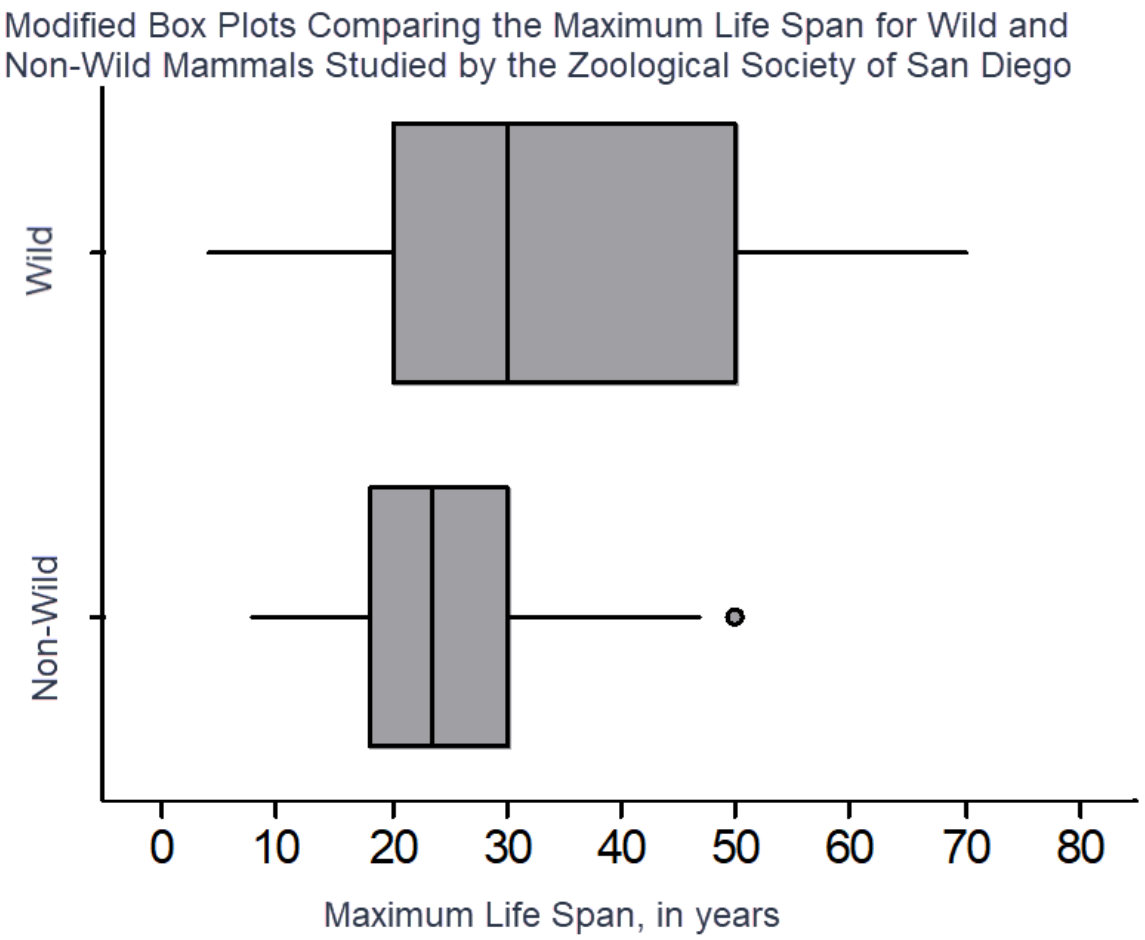
- To make a Box plot,
  - Make a rectangle parallel to the horizontal axis that starts at Q_1 and ends at Q_3
  - Mark the median Q_2 in the middle of the rectangle parallel to the ends of the rectangle
  - Make "whiskers" that extend from Q_1 to the minimum and from Q_3 to the maximum

Modified Box Plot

- Similar to basic box plot except
  - Whiskers extend only as far as the largest and smallest non-outliers for the data
    - Other outliers are marked as individual dots or other symbols
  - Largest and smallest non-outliers are called the *adjacent values*

Modified Box Plot

A modified box plot for the mammal data. Notice that the outlier is marked with a dot.



Modified Box Plots

- Used for quantitative variable
- Does not record individual data values
- Records five-number summary of data *with* outliers

Box Plots
versus
Modified Box Plots???

Box Plots

- Useful when plotting a single quantitative variable
  - Compare shape, center, spread for two or more distributions
  - When distribution has too many values or would require too much space to make a stemplot
  - Do not need to see individual values
  - Do not need more than five-number summary with outliers marked

Modified Box Plots

- Useful when plotting a single quantitative variable
  - Compare shape, center, spread for two or more distributions
  - When distribution has too many values or would require too much space to make a stemplot
  - Do not need to see individual values
  - Need to see the five-number summary with outliers marked