

IDENTIFYING OUTLIERS

One Quantitative Variable
Unit 1: Exploratory Data Analysis

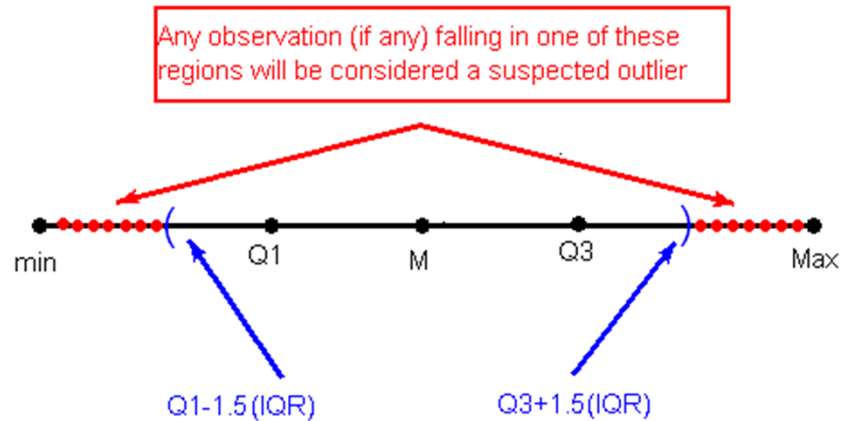


UF UNIVERSITY of
FLORIDA

We will look at two ways to identify outliers. The first of which we will discuss now.

This method, which we will call the IQR method, is based upon the five-number summary, in particular it uses the IQR and the 1st and 3rd quartiles, Q1 and Q3.

Potential or Suspected Outliers



To locate potential or suspected outliers, we need to calculate two values, sometimes called “fences”

These values are not necessarily data points but simply provide a range, where values falling outside the interval are possible outliers.

The two values are calculated by going beyond Q1 and Q3 by 1.5 times the IQR. In other words we take

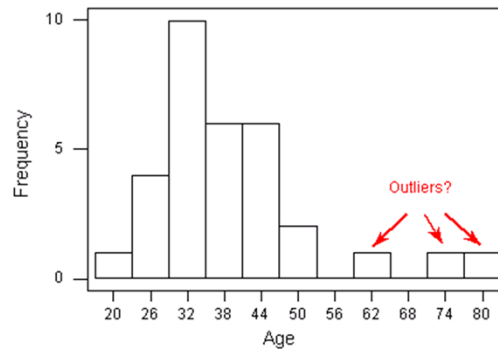
Q1 minus 1.5 times the IQR and

Q3 plus 1.5 times the IQR

Any observation falling outside those values (more toward the extremes) is a potential outlier.

Outliers – IQR Method

- An observation is considered an **EXTREME outlier** if it is:
- below $Q1 - 3(IQR)$ or
- above $Q3 + 3(IQR)$



In addition, any value falling outside 3 times the IQR away from the 1st and 3rd quartiles is considered an EXTREME outlier.

Some software packages will show a different symbol for the extreme outliers, others will not.

For a given sorted dataset where you are also provided Q1 and Q3, you should be able to calculate both sets of values required to determine potential and extreme outliers and then identify the potential and extreme outliers in the data provided.

Examples and practice problems are provided in the materials.

Outliers

- Produced by **essentially the same process** as rest of data?
- Such extremes are expected to **eventually occur again**?
- **Outlier**: important or interesting, **keep** in the data

If an outlier was produced by **essentially the same process** as the rest of data, and if such extremes are expected to **eventually occur again**, then the outlier contains something important and interesting about the process, and it **should be kept** in the data

Outliers

- Produced under **different** conditions from rest of data (or by different process)?
- Outlier: **can be removed** from the data
- **IF your goal** is to investigate only the process that produced the rest of the data

If the outlier was produced under **different** conditions from the rest of the data (or by a different process), the outlier **can be removed** from the data if your goal is to investigate only the process that produced the rest of the data


Outliers

- Does it indicate a **mistake** (typo, measuring error)?
- Outlier:
 - **Correct if possible**
 - **Or else removed** from the data

Before calculating summary statistics or making inferences from the data


Reason for the mistake should be investigated

An outlier which is a **mistake** (typo, measuring error), **should be corrected if possible or else removed** from the data before calculating summary statistics or making inferences from the data (and the reason for the mistake should be investigated)



IDENTIFYING OUTLIERS

One Quantitative Variable
Unit 1: Exploratory Data Analysis



Identifying outliers is an important component of describing the distribution of one quantitative variable.

Next, we will see that this process of identifying outliers is used when creating boxplots.