Missing Values and Anomalies

Detecting missing values

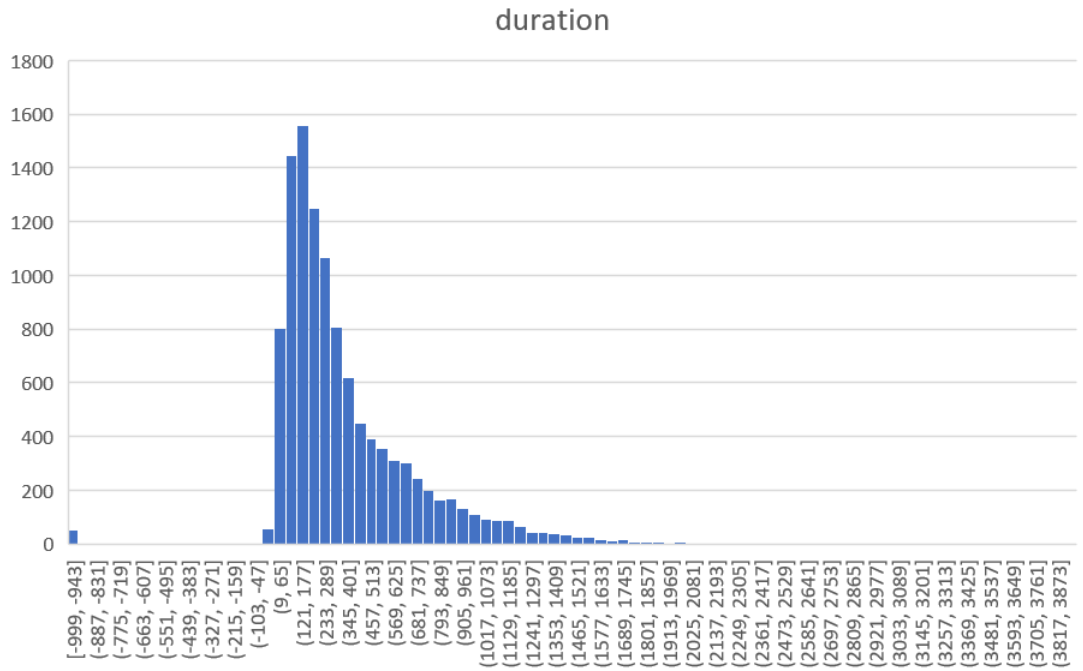
- Missing values come in many forms, e.g. blank, “n/a”, “-99999”, ?

- Missing values of categorical variables can be fairly easily detected, e.g. by means of a frequency table of possible values

marital	Frequency
married	6327
single	3507
divorced	1292
NA	19
	18

Detecting missing values

- Missing values of numerical variables can be detected by a histogram



… or by detecting inliers.

Types of missing values

- Missing completely at random (MCAR): the probability of an instance being missing does not depend on known values nor the missing value itself.

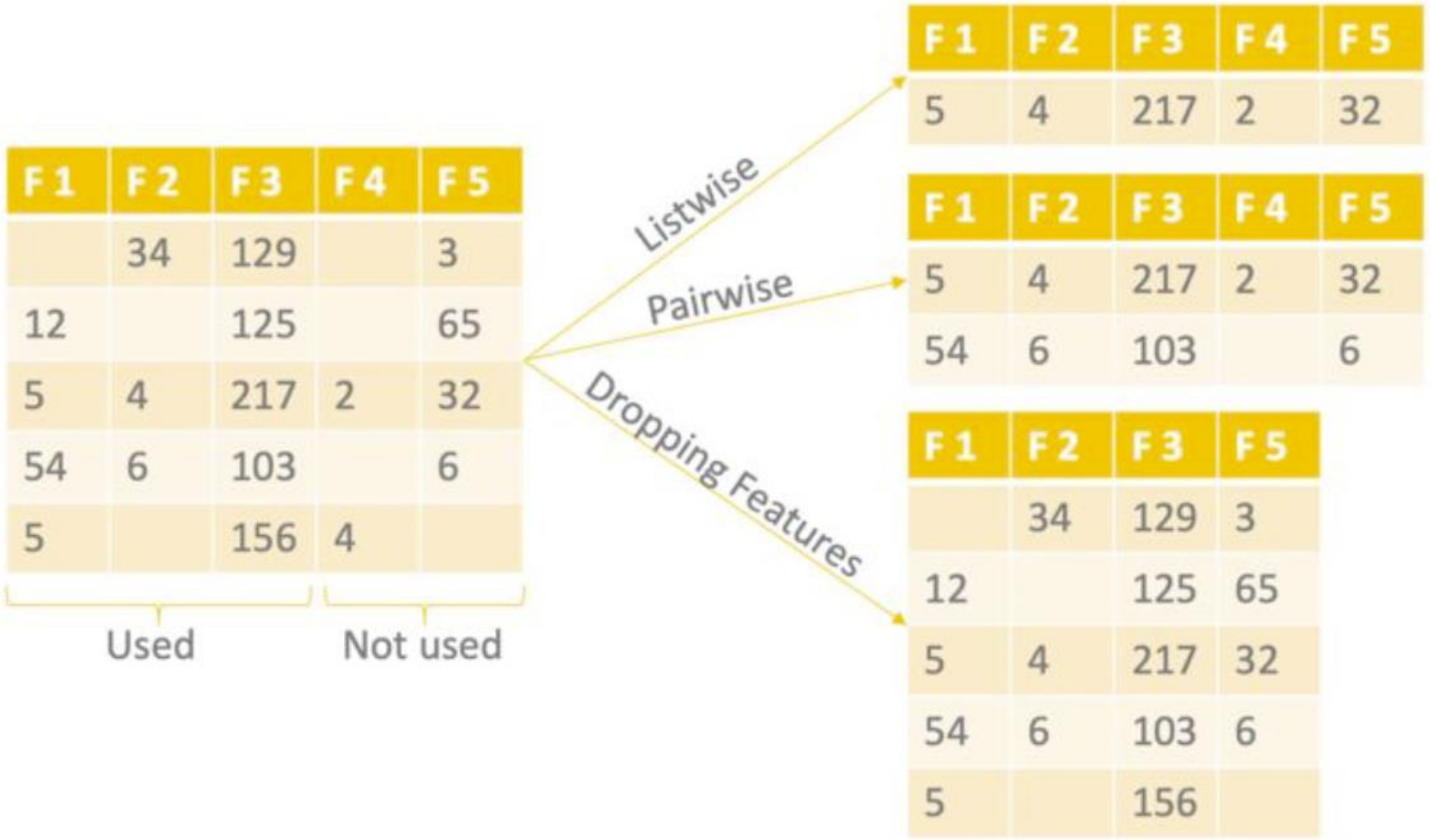
- Missing at random (MAR): The probability of an instance being missing may depend on known values (of other variables), but not on the variable having missing values.

- Missing not at random (MNAR): The probability of an instance being missing depends on other variables which also have missing values, or…

… the probability of missingness depends on the very variable itself.

Imputing missing values

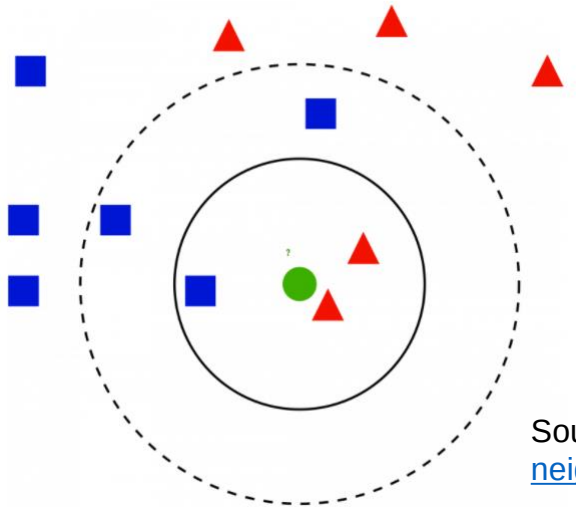
- Deletion methods: listwise, pairwise, and dropping features



Source: https://www.kdnuggets.com/2020/09/missing-value-imputation-review.html

Imputing missing values

- Single imputation
  - Fixed value
  - Minimum or maximum value (or most frequent value)
  - Mean or median or moving average (or most frequent value)
  - Previous or next value (only for time sequence or ordered data)
  - K-nearest neighbours
  - Regression



Source: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

Multiple imputation

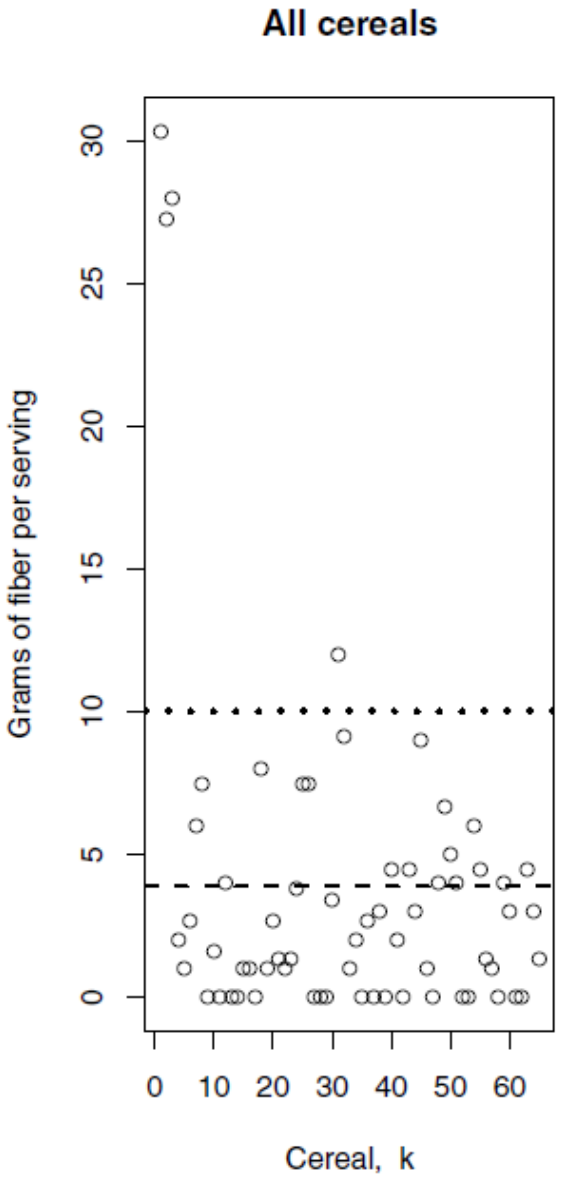
- Creates multiple replacements for each missing value, i.e. multiple versions of the complete dataset.
- Multiple Imputation by Chained Equations
  - Step 1: Make a simple imputation (e.g. mean) for all missing values in the dataset
  - Step 2: Set missing values in a variable ‘A’ back to missing.
  - Step 3: Train a model to predict missing values in ‘A’ using available values of A as dependent and other variables in the dataset as independent.
  - Step 4: Predict missing values in ‘A’ using the trained model in Step 3.
  - Step 5: Repeat Steps 2-4 for all other variables with missing values
  - Step 6: Repeat Steps 2-5 for a number of cycles until convergence (reportedly 10 cycles)

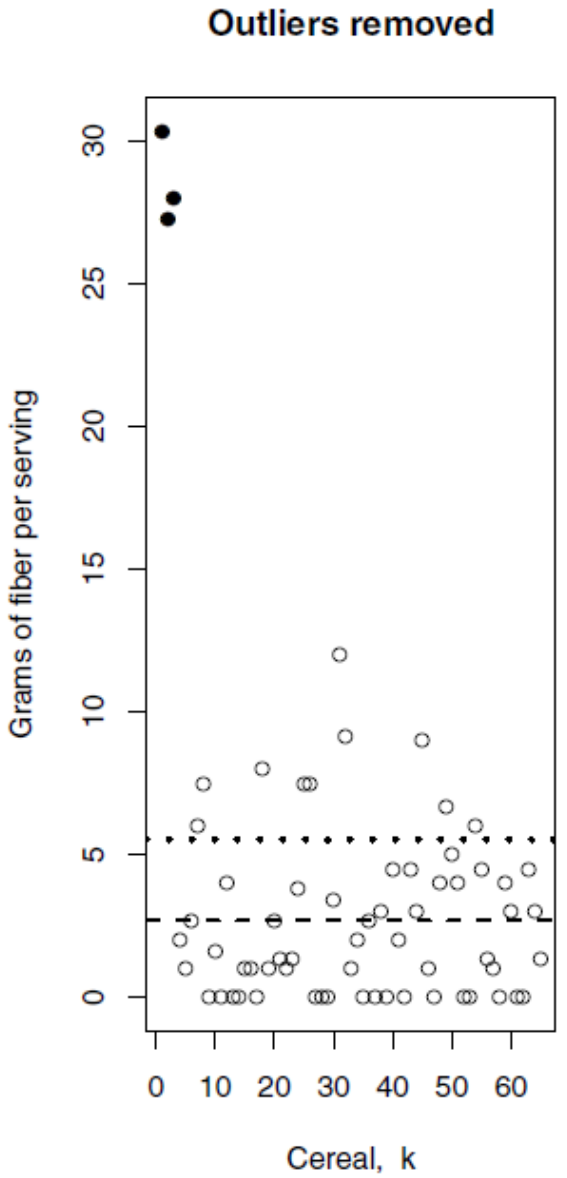
Identifying outliers

- Outlier – “an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.”

(V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, 2nd edition, 1984)

- Outliers significantly change the characteristics of a dataset.
- They can be because of *gross data errors* or from special cases.
- **Example**. Grams of fibre (and potassium - in later slides) in one standard portion of each of 65 cereal brands. Further info here.

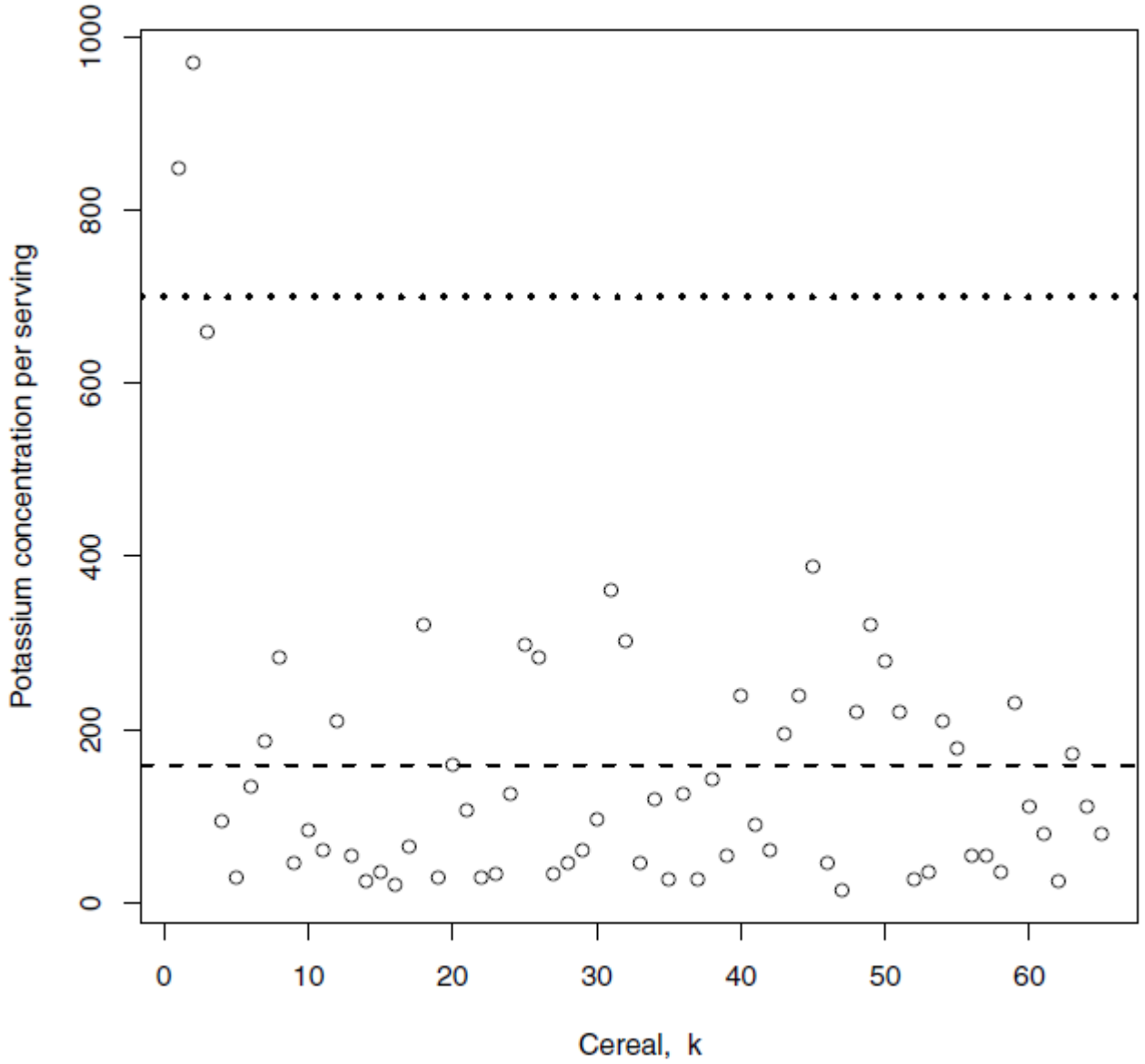




Identifying outliers

- Three-sigma identifier
  - Typical value: mean value $\bar{x}$
  - Data spread: standard deviation σ
  - Bounds: x_k considered outlier if $|x_k - \bar{x}| > 3\sigma$

- Note that σ is *inflated* by outliers.
- Larger outlier values -> larger σ -> larger the bound values -> less effective in identifying unusual values
- We need a different way to measure typical value and the spread so that they are less sensitive to outliers.



Identifying outliers

- The Hampel identifier
  - Typical value: median
  - Data spread: median absolute deviation from the median (MADM)

$$MADM = 1.4826 * median(x_k - median(x))$$

  - Bounds: x_k considered outlier if $|x_k - median| > 3MADM$

x
15
20
25
25
26.32
…
388.06
660
848.48
969.70

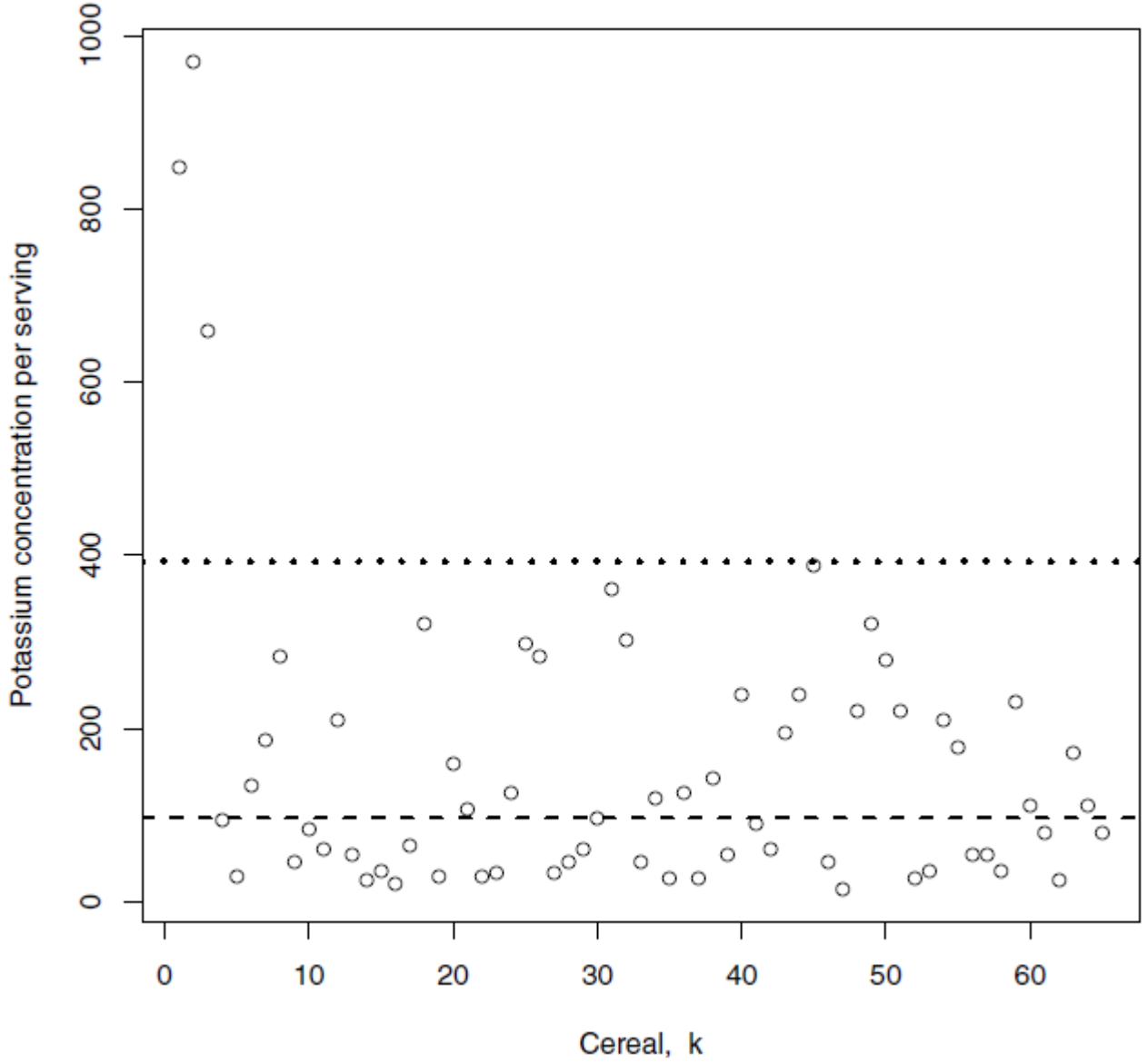
→ $median(x) = 96.59$ →

y = x-median(x)
81.59
76.59
71.59
71.59
70.28
…
291.47
563.41
751.89
873.11

→ $MADM = 1.4826 * median(y) = 98.73$

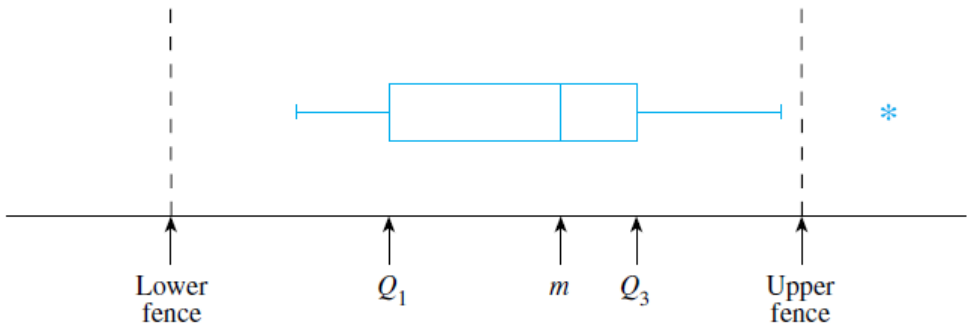
Identifying outliers

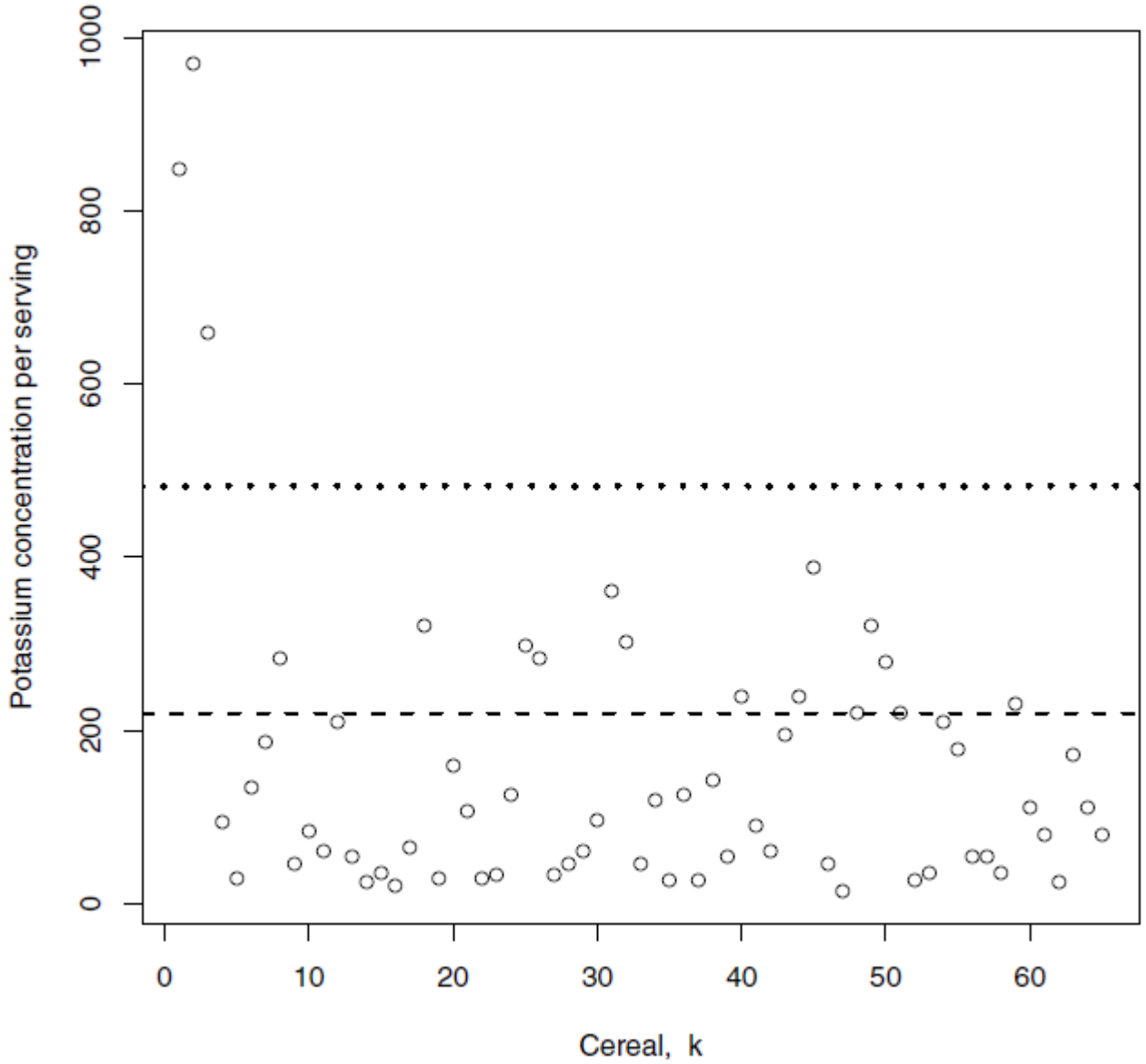
- The Hampel identifier



Identifying outliers

- The boxplot identifier
  - A graphical tool “expressly designed” for isolating outliers from a sample.
  - Bounds: x_k considered outlier if $x_k > Q_3 + 1.5IQR$ or $x_k < Q_1 - 1.5IQR$





Identifying outliers

- The three procedures described above may identify different sets of outliers.
- A suggested strategy:
  - Apply all three procedures and compare (i) the number and the value of outliers identified by each procedure, and (ii) the range of the data values not declared as outliers.
  - Apply application-specific assessments, i.e. does the nominal range (excluded outliers) make sense? Do outliers seem extreme enough to be excluded?
  - Visualise the data either with different colours for nominal values and for outliers, or with indication of outlier detection thresholds.

- *Identifying* outliers can be a mathematical procedure – *interpreting* the outliers is NOT.
- Outliers are not necessarily bad data that should be removed/rejected – they simply need further investigation.

Identifying inliers

- “A data value that lies in the interior of a statistical distribution and is in error”

(D. DesJardins. Paper 169: Outliers, inliers and just plain liars – new eda+ techniques for understanding data. In *Proceedings SAS User’s Group International Conference*, SUG126. Cary, NC, USA, 2001)

- Inliers often represent in the form of *similar* values *repeating unusually frequently*.
- **Example**. Dataset “Chile” in package “car” available in R (more info here).

  We wish to find a way to conclude that values such as -1.29617, which appears 201 times, as *inliers*.

  In other words, we wish to conclude that 201 is an outlier among the values in Frequency.

	Chile$statusquo	Frequency
1	-1.80301	1
2	-1.74401	1
…	…	…
19	-1.29617	201
20	-1.29293	2
21	-1.28924	1
22	-1.28897	1
23	-1.27876	3
24	-1.27556	1
25	-1.2727	5
…	…	…
2092	2.04859	1
2093	NA	17

Identifying inliers

Because the majority of numerical values in Chile$statusquo appears only *once*,

- the majority of values in Frequency is 1, median of Frequency is 1, MADM of Frequency is 0 => we cannot use Hampel identifier to detect inliers.
- Quartiles of Frequency are as below

0%	25%	50%	75%	100%
1	1	1	1	201

- Both Hampel and boxplot procedures would declare that all data points in Frequency are outliers!

	Chile$statusquo	Frequency
1	-1.80301	1
2	-1.74401	1
...	...	...
19	-1.29617	201
20	-1.29293	2
21	-1.28924	1
22	-1.28897	1
23	-1.27876	3
24	-1.27556	1
25	-1.2727	5
...	...	...
2092	2.04859	1
2093	NA	17

Frequency													
1	2	3	4	5	6	8	9	13	17	18	21	61	201
1955	72	22	19	8	5	4	1	1	2	1	1	1	1

Identifying inliers

- Applying the three-sigma procedure to identify outliers in Frequency.
  - Mean $\bar{x} = 1.29$
  - Standard deviation $\sigma = 4.67$
  - A value x_k in Frequency is considered outlier if $|x_k - \bar{x}| > 3\sigma$ or $x_k > 15.3$

	Chile$statusquo	Frequency
19	-1.29617	201
39	-1.25795	21
61	-1.21834	18
137	-1.14049	17
2074	1.5877	61
2093	NA	17

- Similar to outliers, inliers are not necessarily bad data and need to be rejected/removed – they simply need further investigation.

References and further readings

- Missing data imputation
- Tutorial: Introduction to Missing Data Imputation
- Review: A gentle introduction to imputation of missing values
- Missing value imputation – a review
- Multiple imputation by chained equations: what is it and how does it work?