Data Analysis, Statistics, Machine Learning

Leland Wilkinson
Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai
leland.wilkinson@gmail.com
Summarizing

We summarize to remove irrelevant detail
  We summarize batches of data in a few numbers
  We summarize variables through their distributions
The best summaries preserve important information
All summaries sacrifice information (lossy)

Summaries
  Location
    Popular: mean, median, mode
    Others: weighted mean, trimmed mean, ...
  Spread
    Popular: sd, range
    Others: Interquartile Range, Median Absolute Deviation, ...
  Shape
    Skewness
    Kurtosis
Summarizing

Location (Mean or Average)

\[
\text{mean} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

Mean is the value whose sum of squared deviations to \(x_i\) is smallest

The mean is a good location summary if the batch is symmetrically distributed

The mean **balances** the batch

Because squared deviations have a lot of leverage in the overall result, the mean is not a good summary when there are outliers or severe skewness
Summarizing

Location (Median)

Median = middle value of ordered list of $x_i$ \( (i = 1, \ldots, n) \)

If \( n \) is an even integer, any value between the two middle values is a median (we usually average the two middle values)

Median is the value whose sum of absolute deviations is smallest

The median \textbf{splits} the batch

Median is robust against outliers

---

Copyright © 2016 Leland Wilkinson
Summarizing

Location (Mode)

Mode = frequent($x_i$)
The mode **votes** the batch
Some batches have no mode
Some have several (multimodal)
Kernel estimate of mode
Summarizing

Location (Midrange)

Midrange = \((\max(x) - \min(x)) / 2\)

Not efficient for most distributions (subject to sampling variation)

Not robust against outliers

But can be trimmed
Summarizing

Location (Trimmed mean)

\[
\text{TrimmedMean} = \text{mean}(\text{trim}_\text{percent}(x_i))
\]

Delete top and bottom percent and compute mean

The trimmed mean is robust against outliers

The arrow below is for 25% trimmed mean (half the values used)

Used for voting schemes to remove biased or extreme judges
Summarizing

Location (Winsorized mean)

\[ \text{WinsorizedMean} = \text{mean}(\text{winsor}_{\text{percent}}(x_{ij})) \]

The `winsor()` function changes extreme (outer) values to nearest inner value

An inner value is \textit{inside} selected lower and upper percentile

An outer value is \textit{outside} selected lower and upper percentile

The Winsorized mean is robust against outliers

The arrow below is for 25% Winsorized mean (half the values used)
Summarizing

Location (Biweight location estimate)

\[
M = \text{median}(x) \\
S = \text{MAD}(x, M) \\
u_i = \frac{x_i - M}{cS + \epsilon} \\
w_i = (1 - u_i^2)^2 \text{ if } |u_i| \leq 1, \text{ else } w_i = 0 \\
\bar{x} = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i
\]
Summarizing

Spread (standard deviation)

\[ sd = \text{root}(\text{mean}(\text{square}(\text{deviation}(\text{mean})))) \]

Forget the formulas in stat books
They are inaccurate on a computer
They tell you nothing about what \( sd \) means

Designed for symmetric distributions (especially Normal)

The sample standard deviation is not robust
Squaring large deviations leads to roundoff error
Even worse behavior with desk-calculator algorithm
Summarizing

**Spread (Interquartile Range)**

\[ IR = (q_{.75}(x) - q_{.25}(x)) \]

Designed for symmetric and asymmetric distributions

does not depend on location estimator

Not as efficient as standard deviation on Normal distribution

More robust than the sample standard deviation

no squaring values

ignores outliers

Similar rationale to trimmed mean

Used in box plots
Summarizing

Spread (Median Absolute Deviation)

\[ \text{MAD} = \text{median}(\text{list_of_deviations}(\text{median}(x))) \]

Designed for symmetric contaminated distributions

Not as efficient as standard deviation on Normal distribution

More robust than the sample standard deviation

No squaring large deviations
Summarizing Spread (Rousseeuw and Croux, JASA 1993)

\[ S = \text{median}_i(\text{median}_j(\left| x_i - x_j \right|)) \]

for each \( i, i=1, \ldots, n \), compute median of absolute differences to all other values and compute the median of the resulting list of size \( n \).

Computationally complex: \( O(n \log n) \)

Designed for symmetric and asymmetric contaminated distributions

does not depend on location estimator

Almost as efficient as standard deviation on Normal distribution

Much more robust than the sample standard deviation
Summarizing

Shape (Skewness)

\[ S = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{sd(x)} \right)^3 \]

Measures positive or negative asymmetry
- think of long tail as arrow – if it points right, it’s positively skewed

Really, really not robust
- we’re cubing things!

\[ S = 0 \quad S = 2 \]
Summarizing

Shape (Kurtosis)

\[
K = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{sd(x)} \right)^4
\]

Measures peaked or flat shape
Really, really, really not robust
we’re fourth-powering things!

\[
K = 3 \quad K = 9
\]
Summarizing

Shape (L-moments)

\[ \ell_1 = \left( \frac{n}{1} \right)^{-1} \sum_{i=1}^{n} x(i) \]

\[ \ell_2 = \frac{1}{2} \left( \frac{n}{2} \right)^{-1} \sum_{i=1}^{n} \left\{ \binom{i-1}{1} - \binom{n-i}{1} \right\} x(i) \]

\[ \ell_3 = \frac{1}{3} \left( \frac{n}{3} \right)^{-1} \sum_{i=1}^{n} \left\{ \binom{i-1}{2} - 2 \binom{i-1}{1} \binom{n-i}{1} + \binom{n-i}{2} \right\} x(i) \]

\[ \ell_4 = \frac{1}{4} \left( \frac{n}{4} \right)^{-1} \sum_{i=1}^{n} \left\{ \binom{i-1}{3} - 3 \binom{i-1}{2} \binom{n-i}{1} + 3 \binom{i-1}{1} \binom{n-i}{2} - \binom{n-i}{3} \right\} x(i) \]

Very robust
Summarizing

No estimate of location, spread, or shape is as good as a graphic