STAT 135, Concepts of Statistics

Helmut Pitters

Introduction

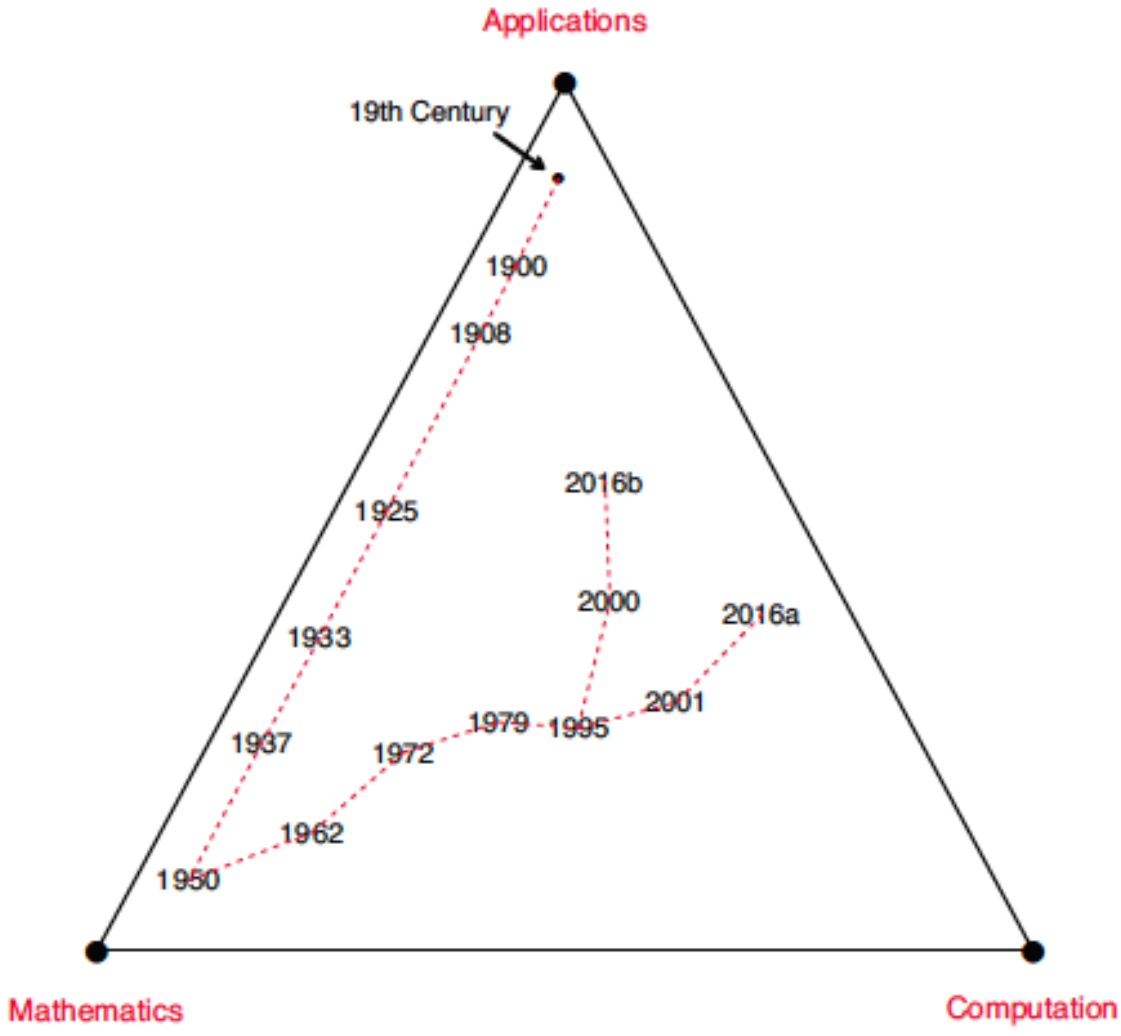
Department of Statistics
University of California, Berkeley

January 16, 2017

Introduction

> `Statistics is the science of learning from experience, particularly experience that arrives a little bit at a time: the successes and failures of a new experimental drug, the uncertain measurements of an asteroid's path toward Earth. ...`
>
> Efron, Hastie 2016 - Computer Age Statistical Inference

Introduction



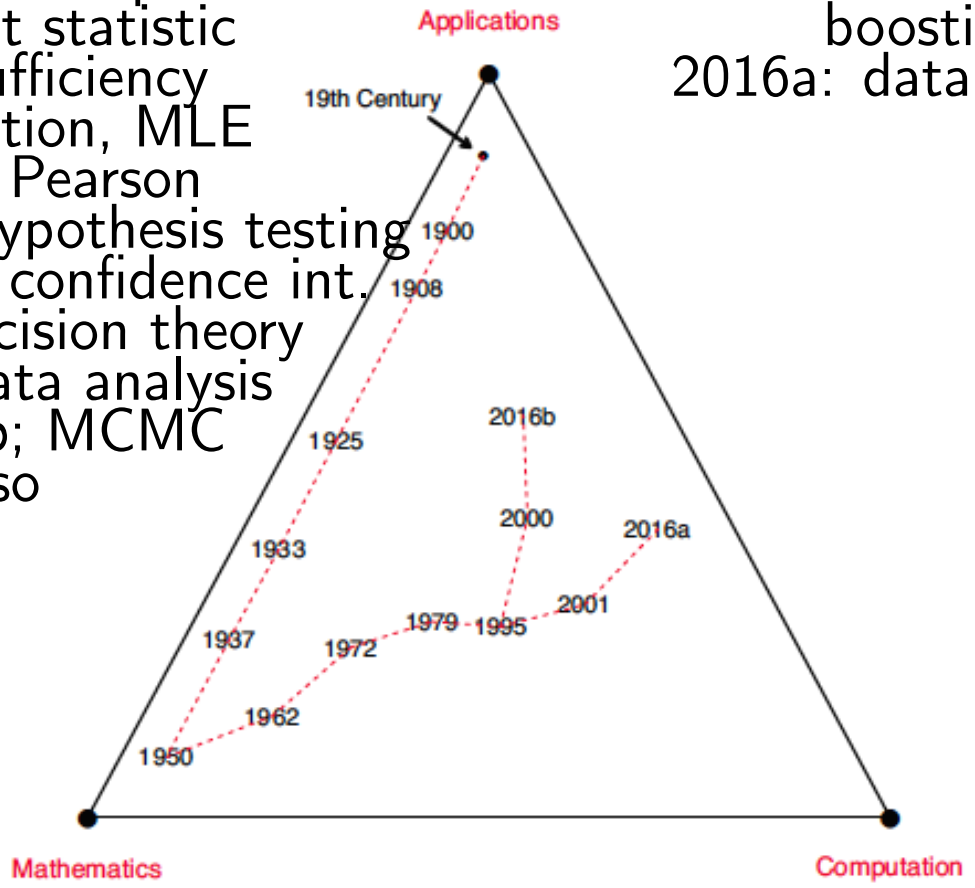
Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

Figure: From Efron, Hastie 2016

Introduction

1900: Pearson; chi square test
1908: Student; t statistic
1925: Fisher; sufficiency
F information, MLE
1933: Neyman, Pearson
optimal hypothesis testing
1937: Neyman; confidence int.
1950: Wald; decision theory
1962: Tukey; data analysis
1979: bootstrap; MCMC
1995: FDR; lasso

2001: random forests
boosting, neural nets
2016a: data science



Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

Figure: From Efron, Hastie 2016

Introduction and setup

We are interested in studying a (usually large) *population* of N individuals, e.g.

- California residents,
- patients worldwide diagnosed with some form of cancer,
- customers of a large mall,
- clients of life insurance company,
- facebook users, etc.

In particular, we are interested in certain characteristics

$$x_1, x_2, \ldots, x_N$$

of the individuals, where

$$x_i = \text{characteristic of } i\text{th individual},$$

e.g.

- age,
- expected lifetime,
- monthly income,
- number of friends/relationships.

Introduction and setup

Example (Speed of light)

Consider 60 of Michelson's measurements of the speed of light (observations=values listed+299000km/s).

Table: Velocity of light. Michelson, 1879.

850	960	880	890	890	740	940	880	810	850
840	900	960	880	810	780	1070	940	860	820
810	930	880	720	800	760	850	800	720	770
810	950	850	620	760	790	980	880	860	740
810	980	900	970	750	820	880	840	950	760
850	1000	830	880	910	870	980	790	910	870

[Example:histogram]

Introduction and setup

Remark

Notice: characteristics $x_1, \ldots, x_N$ are *not random*, but deterministic quantities.
However, usually one does not have access to all of the information

$$x_1, \ldots, x_N$$

about the population, but only to a (randomly chosen) sample

$$x_{j_1}, x_{j_2}, \ldots, x_{j_n} \subseteq \{x_1, \ldots, x_N\}.$$

It is often convenient to summarize important features of the population in a few numbers, referred to as *population parameters*. We now turn to summaries that are often used in statistics.

Summarizing data.

Measures of location

Measures of location. Mean.

Arithmetic mean. The *arithmetic mean*

$$\bar{x} := \frac{1}{N} \sum_{i=1}^{N} x_i$$

or *average value* is probably the most common measure of location.
[Histogram: mean <> center of mass]

Measures of location. Mean.

Example (Speed of light)

Consider 60 of Michelson's measurements of the speed of light (observations=values listed+299000km/s), none of which is completely accurate due to the sensitivity of the measuring apparatus.

Table: Velocity of light. Michelson, 1879.

850	960	880	890	890	740	940	880	810	850
840	900	960	880	810	780	1070	940	860	820
810	930	880	720	800	760	850	800	720	770
810	950	850	620	760	790	980	880	860	740
810	980	900	970	750	820	880	840	950	760
850	1000	830	880	910	870	980	790	910	870

Expect $\bar{x} = 856.33$ to be a more accurate measure than each of the individual observations.

Measures of location. Mean.

Example (Speed of light)

Table: Velocity of light. Michelson, 1879.

850	960	880	890	890	740	940	880	810	850
840	900	960	880	810	780	1070	940	860	820
810	930	880	720	800	760	850	800	720	770
810	950	850	620	760	790	980	880	860	740
810	980	900	970	750	820	880	840	950	760
850	1000	830	880	910	870	980	790	910	870

Q: Consider the histogram of Michelson's observations. Roughly, what kind of shape do you expect to see? Why?

Measures of location. Mean.

A drawback of $\bar{x}$ is its sensitivity to outliers, as the next example shows.

Example (Family incomes)

Incomes of five Berkeley families[1] are

$$\$90k \qquad \$70k \qquad \$77k \qquad \$85k \qquad \$300k.$$

Average

$$\bar{x} = 124.4k$$

of these incomes substantially influenced by single family with highest income. In fact, $\bar{x}$ is greater than the income of any of the other four families.

A data value which is extreme w.r.t. the bulk of other values is called an *outlier*.

[1] http://www.city-data.com/income/income-Berkeley-California.html

Measures of location. Median.

Median. Arrange the items in ascending order

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}.$$

(In particular, $x_{(1)}$ is the smallest value, $x_{(N)}$ is the largest value.)

Definition

The *median* is defined to be the value of the middle item if N is odd, and the average of the values of the two middle items if N is even.

Notice that changing extreme values (e.g. smallest value $x_{(1)}$ or largest value $x_{(N)}$) does not affect the median. For this reason the median is said to be a *robust* measure of location.

Measures of location. Median.

Example (Familiy incomes)

The ordered family incomes are

$$\$70k \qquad \$77k \qquad \$85k \qquad \$90k \qquad \$300k$$

with median

$$\$85k.$$

This number seems to summarize the five family incomes more appropriately.

Measures of location. Trimmed mean.

For $\alpha \in [0, 1]$ the $100\alpha\%$ *trimmed mean*, denoted $\bar{x}_\alpha$, is calculated by discarding the

$$\begin{cases} \text{lowest } 100\alpha\%, \text{ and the} \\ \text{highest } 100\alpha\% \end{cases} \tag{1}$$

observations, and computing the mean of the remaining observations.
Commonly the value of α is taken between 0.1 and 0.2.

Measures of location. Mode.

Definition

The *mode* of a set of data (if it exists) is the value that occurs with greatest frequency.

Notice that there may be two or more observations that occur with greatest frequency, in which case the data is said to be *bimodal*, respectively *multimodal*.
In the extreme case that all observed values are different,
e.g. family incomes

$$\$90k \qquad \$70k \qquad \$77k \qquad \$85k \qquad \$300k.$$

the mode is note defined.

Measures of location. Mode.

Example

Organization is to hold a national meeting. 80 members of organization (picked randomly) are asked to indicate their preference of city:

city	frequency
Miami	16
New Orleans	24
New York	12
San Francisco	28

Mode is San Francisco, the most preferred city.

Notice that in this example, and generally for qualitative data, the notion of a mean or median does not make any sense. The mode is therefore particularly suitable for qualitative data.

Measures of location.

There is no single "best" measure of location; instead, the choice of which measure(s) of location to use depends on the problem at hand and the purpose to which the measure is employed.

To summarize data, it is useful to compute different measures of location and to compare with graphical displays (e.g. histogram) of the data.

Measures of dispersion

Measures of dispersion. Percentile.

Measures of dispersion are used to quantify how much "spread-out" data are.

Do you recall (STAT 134) a quantity measuring the "spread" of a random variable X about its mean $\mathbb{E}[X]$?

Measures of dispersion. Percentile.

Percentile: statistical measure locating values in data set that are not necessarily central locations.
Provides information regarding how data are spread over an interval from lowest to highest value.

Definition (Percentile)

A p*th percentile* of a data set is a value such that *at least* $p\%$ of the items take on this value or less and *at least* $(100-p)\%$ of the items taken on this value or more.

Example (Admission test scores)

Admission test scores of universities and colleges often reported in percentiles.
See e.g. `http://www.collegesimply.com/guides/1600-on-the-sat/`.

Measures of dispersion. Percentile.

Example (Family income)

$$\$70k \qquad \$77k \qquad \$85k \qquad \$90k \qquad \$300k.$$

Here a 20th percentile is $\$73.5k$, a 40th percentile is $\$81k$, etc.

Measures of dispersion. Percentile.

Calculating pth percentile of $x_1, \ldots, x_N$.

1. Arrange data values in increasing order: $x_{(1)} \leq \ldots \leq x_{(N)}$.
2. Compute index
$$i := \frac{p}{100} N.$$
3. The pth percentile is
$$\begin{cases} x_{(\lceil i \rceil)} & \text{if } i \text{ is not an integer} \\ \frac{x_{(i)} + x_{(i+1)}}{2} & \text{otherwise.} \end{cases}$$

Here $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

Measures of dispersion. Range.

The range is possibly the simplest measure of dispersion.

Definition (Range)

If the data $x_1, \ldots, x_N$ consist of real numbers, their range is defined to be the difference

$$x_{(N)} - x_{(1)}$$

between the maximal and the minimal value.

Notice however, that this measure is extremely sensitive to outliers.

Example (Family income)

$$\$90k \qquad \$70k \qquad \$77k \qquad \$85k \qquad \$300k.$$

The range here is $\$300k - \$70k = \$230k$.

Measures of dispersion. Variance.

The *variance*

$$s^2 := \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

is the most common measure of dispersion together with the *standard deviation* s.

Measures of dispersion. Sample variance.

Example (Family income)

$$\$70k \qquad \$77k \qquad \$85k \qquad \$90k \qquad \$300k.$$

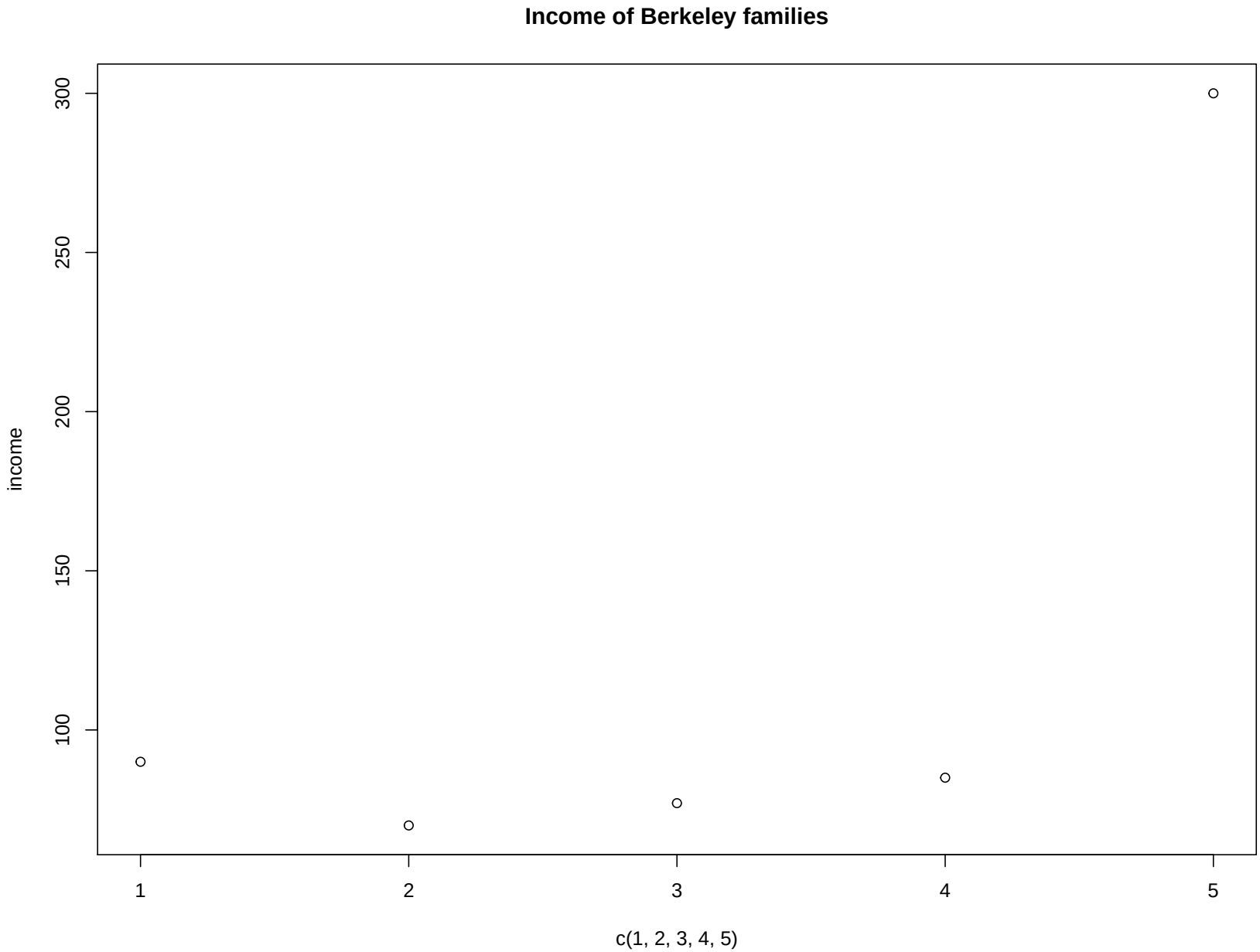
Here the standard deviation is

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \$98.46k.$$

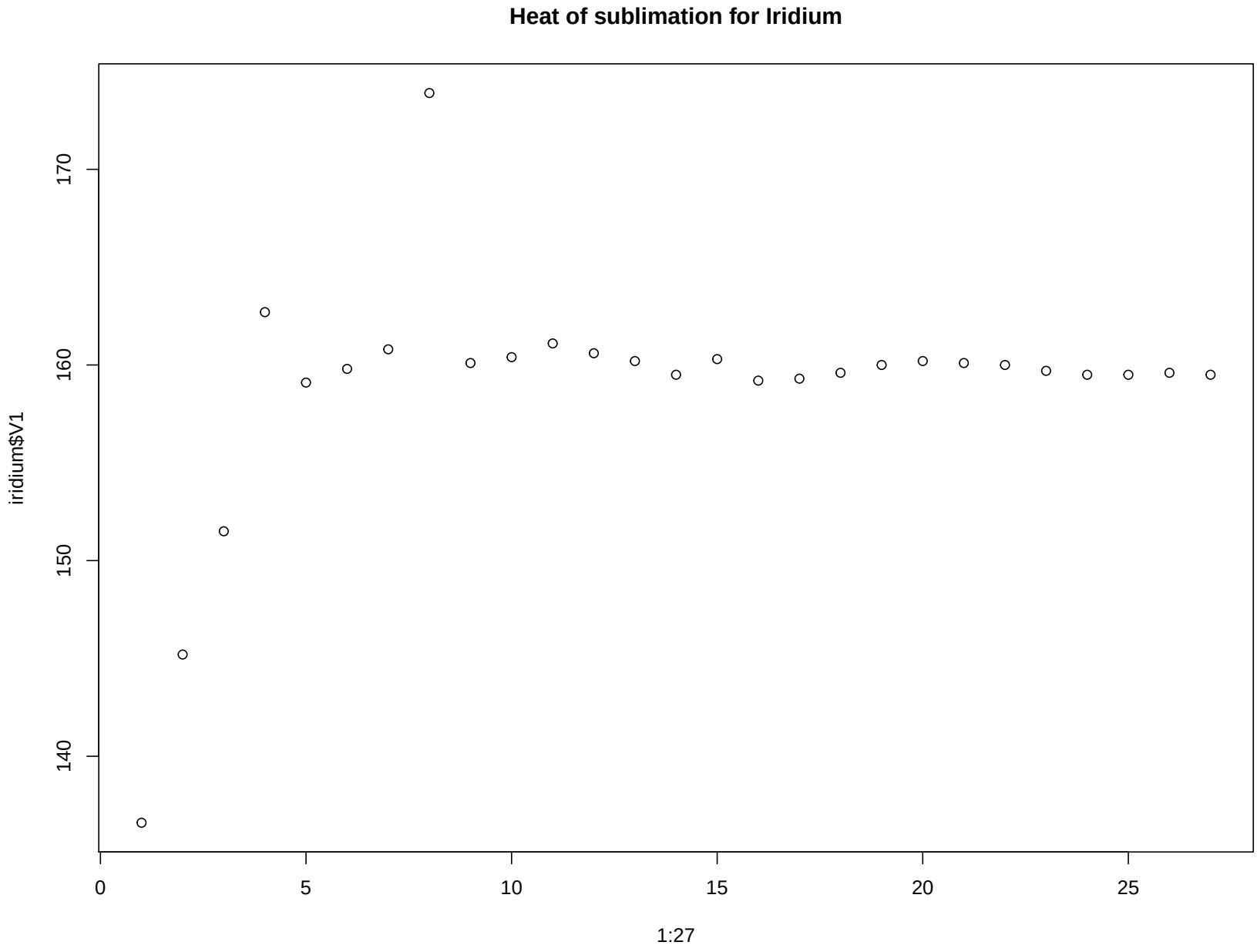
The standard deviation is rather sensitive to outliers, and we will encounter more robust measures of dispersion (e.g. interquartile range) later.

Graphical methods

It can often be useful to plot the data $x_1, \ldots, x_n$ in sequential order.



We can immediately find the outliers in the data set on heat of sublimation of Iridium.



And we immediately see that the sublimation points for Rhodium are more dispersed.