STAT 135, Concepts of Statistics

Helmut Pitters

Summarizing data 2 & nonparametric bootstrap

Department of Statistics
University of California, Berkeley

March 21, 2017

Summarizing data.

In what follows we consider data $x_1, \ldots, x_n$. The data are not necessarily considered as observations of an underlying probabilistic model.

So far encountered

- measures of location
  (trimmed) mean, median, mode
- measures of spread
  percentiles, range, variance
- and some graphical methods.

We turn to some more detailed summaries.

Tukey's 5-number summary.

A 5-*number summary* of a data set (of real numbers) consists of the

1. minimal value, $x_{(1)}$,
2. first quartile (25th percentile), sometimes called the *lower hinge*,
3. median,
4. third quartile (75th percentile), sometimes called the *upper hinge*, and
5. maximal value, $x_{(n)}$.

Example (Heat of sublimation for Iridium and Rhodium)

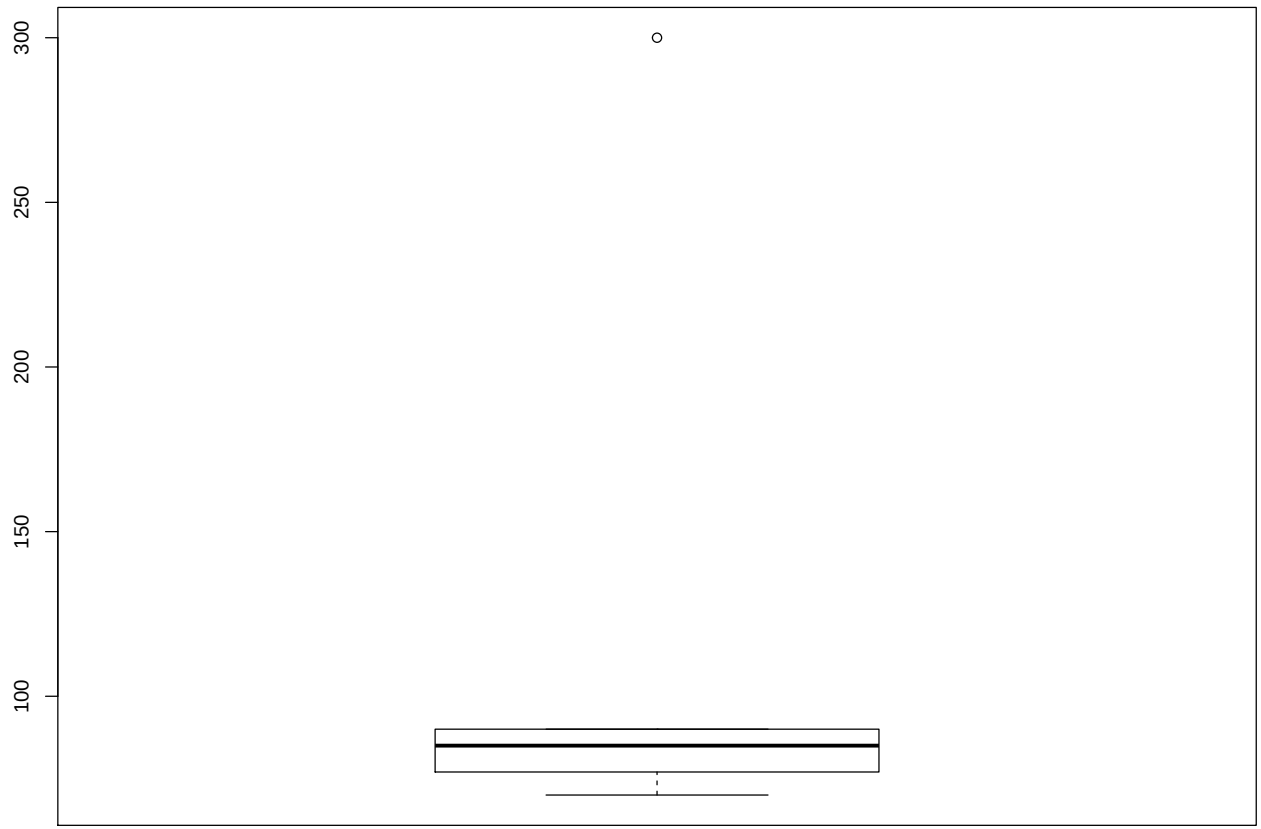
The 5-number summaries for the data sets on heat of sublimation for

1. Iridium: 136.60, 159.50, 159.80, 160.25, 173.90
2. Rhodium: 126.40, 131.45, 132.65, 133.30, 135.70

Box-and-Whisker plot.

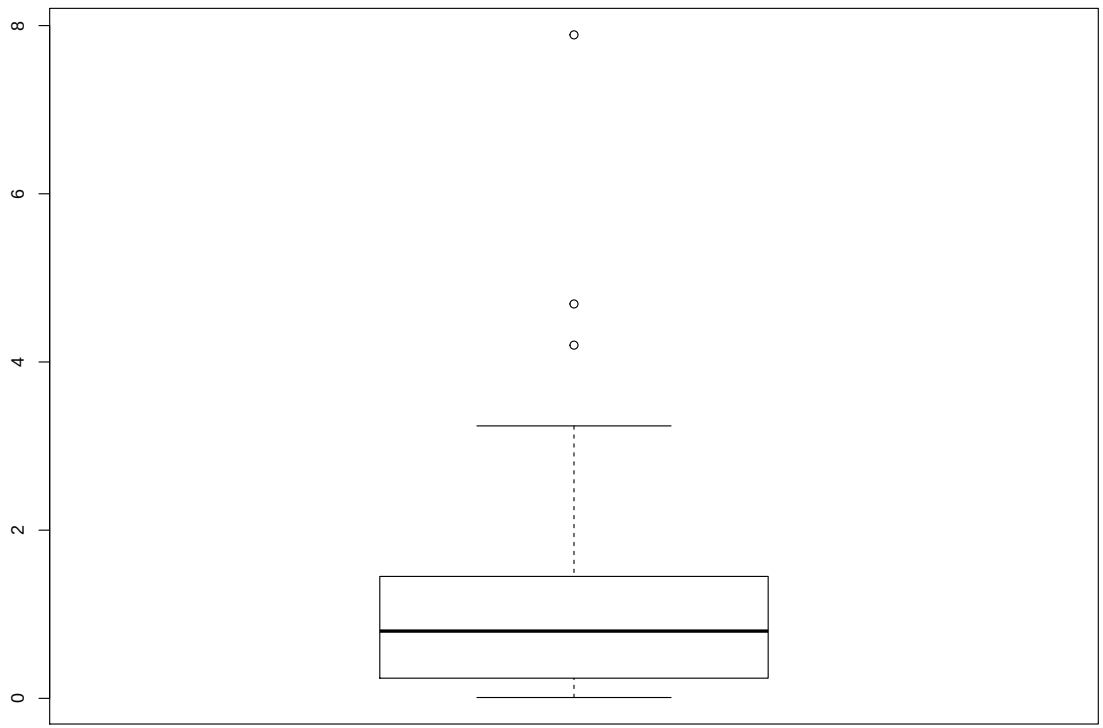
A Box-and-Whisker plot is essentially an easy-to-read graphical display of a 5-number summary.

Below is a boxplot of the data on family income.



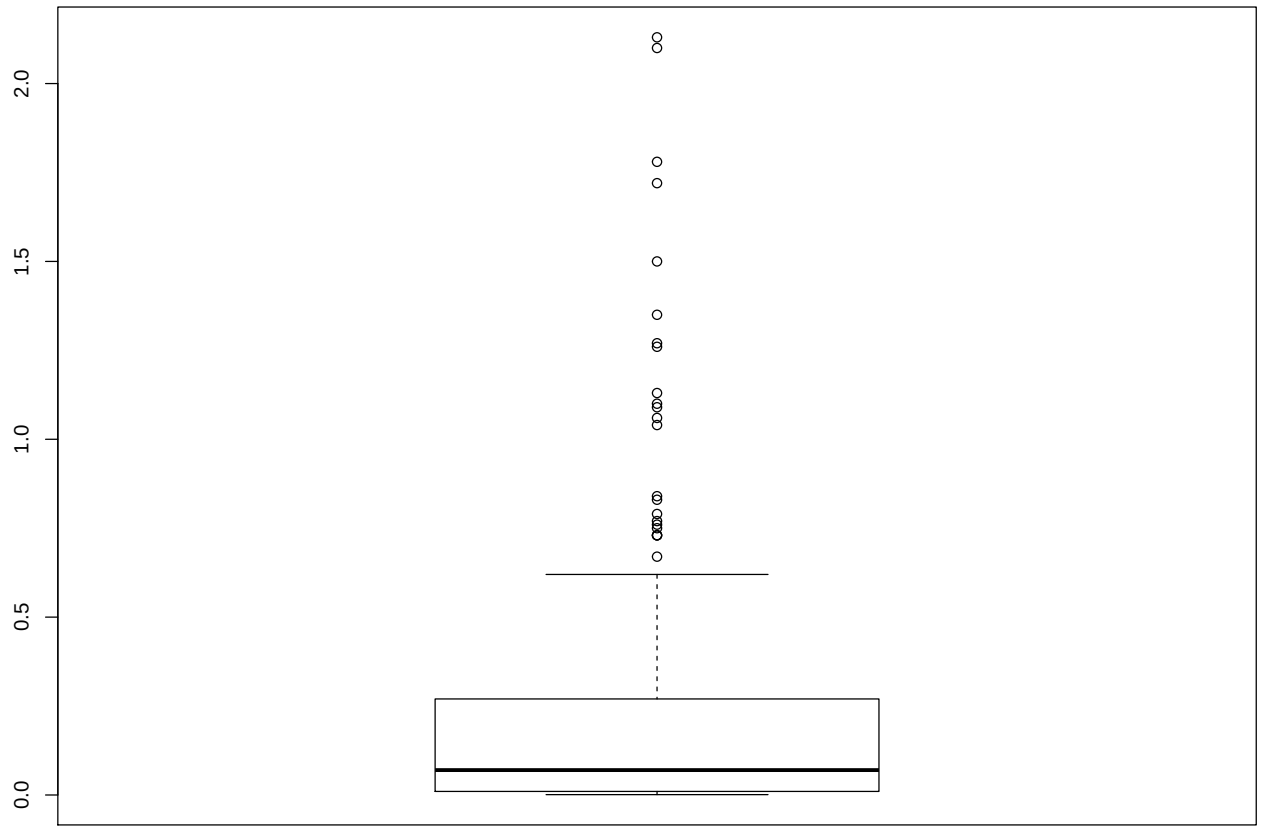
Box-and-Whisker plot.

A boxplot of the times to failure [hours] data of 76 strands of Kevlar material used in space shuttles.
Horizontal bars in the box indicate the lower hinge, median and upper hinge. The two remaining bars indicate the most extreme data points within distance of 1.5 of the upper/lower quartile. Dots indicate outliers.



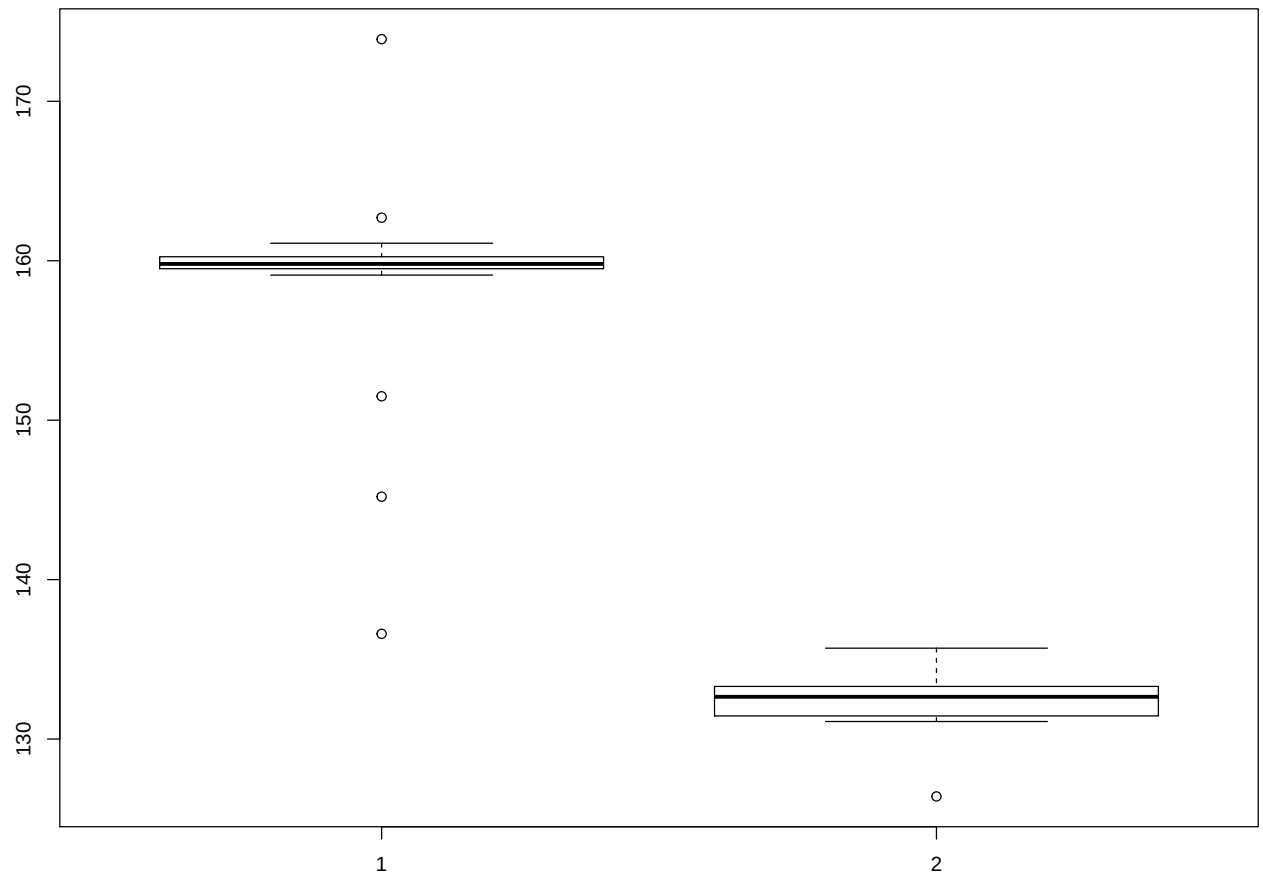
Box-and-Whisker plot.

A boxplot of the precipitation data for storms in Illinois.



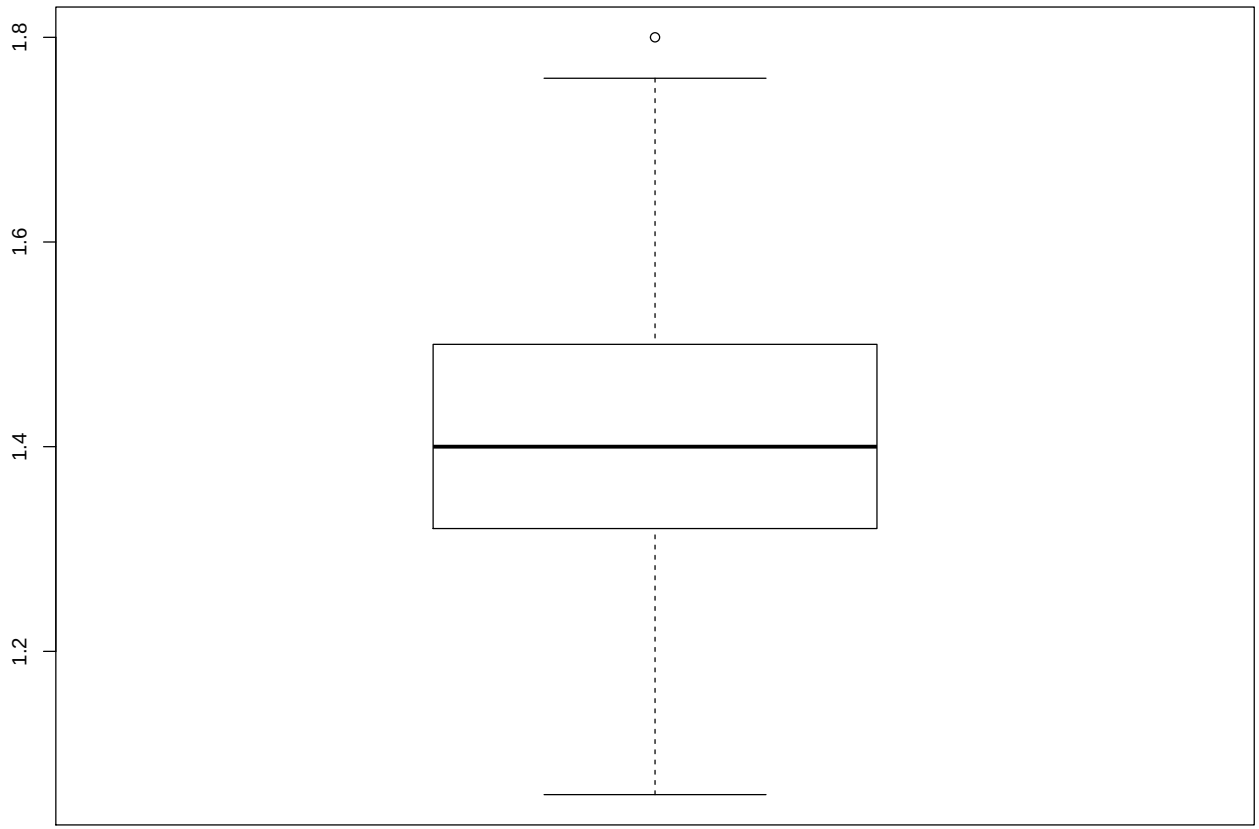
Box-and-Whisker plot.

A boxplot of the heat of sublimation for iridium (left) and rhodium (right).



Box-and-Whisker plot.

A boxplot of the heat of the percentage of manganese in iron.



Empirical cdf.

Suppose data $x = (x_1, \ldots, x_n)$ are given by real numbers.

Definition (Empirical cdf)

The *empirical cumulative distribution function (ecdf)*[1] is defined by

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(x_i) = \frac{1}{n} \#\{i \colon x_i \leq x\}$$

$$= \text{relative frequency of data items with values} \leq x.$$

[1] In R use command `ecdf`(x) to compute the ecdf.

Empirical cdf: Family incomes

Example (Family incomes)

Incomes of five families sampled randomly from Berkeley's population[2] are

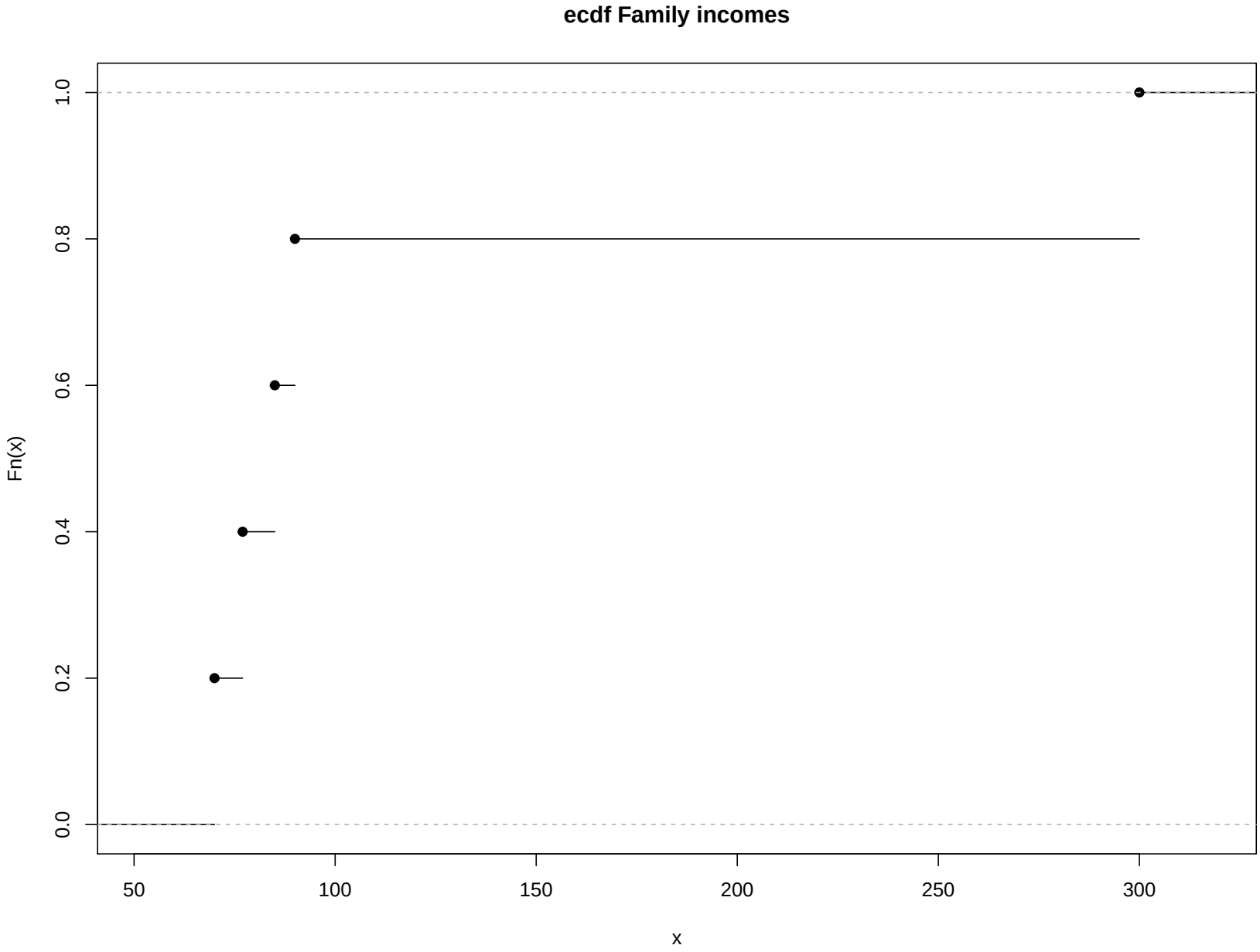
$$\$90k \qquad \$70k \qquad \$77k \qquad \$85k \qquad \$300k.$$

Average

$$\bar{x} = 124.4$$

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

Empirical cdf: Family incomes



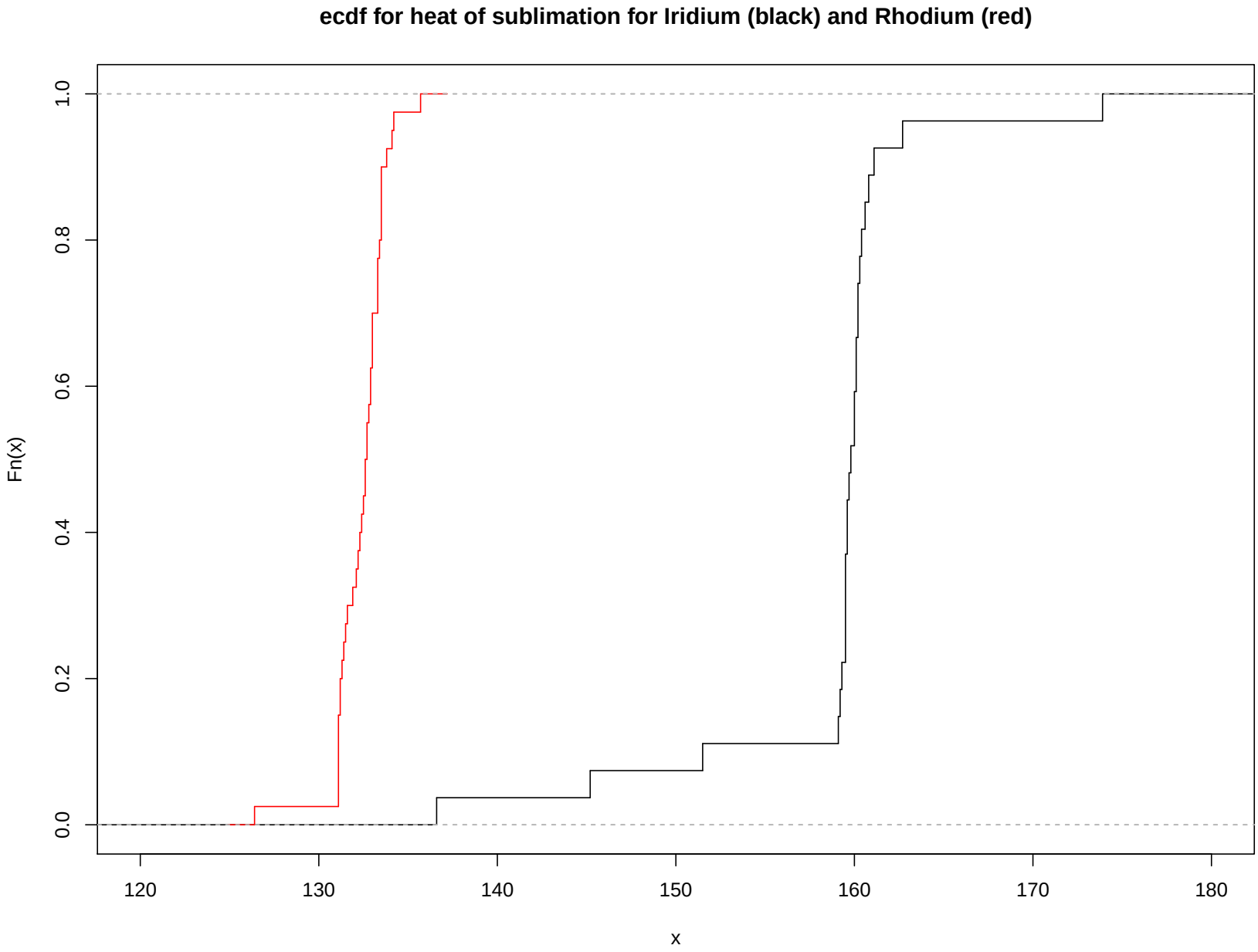
Empirical cdf.

Notice that the ecdf $F_n(x) = \frac{1}{n}\#\{i \colon x_i \leq x\}$

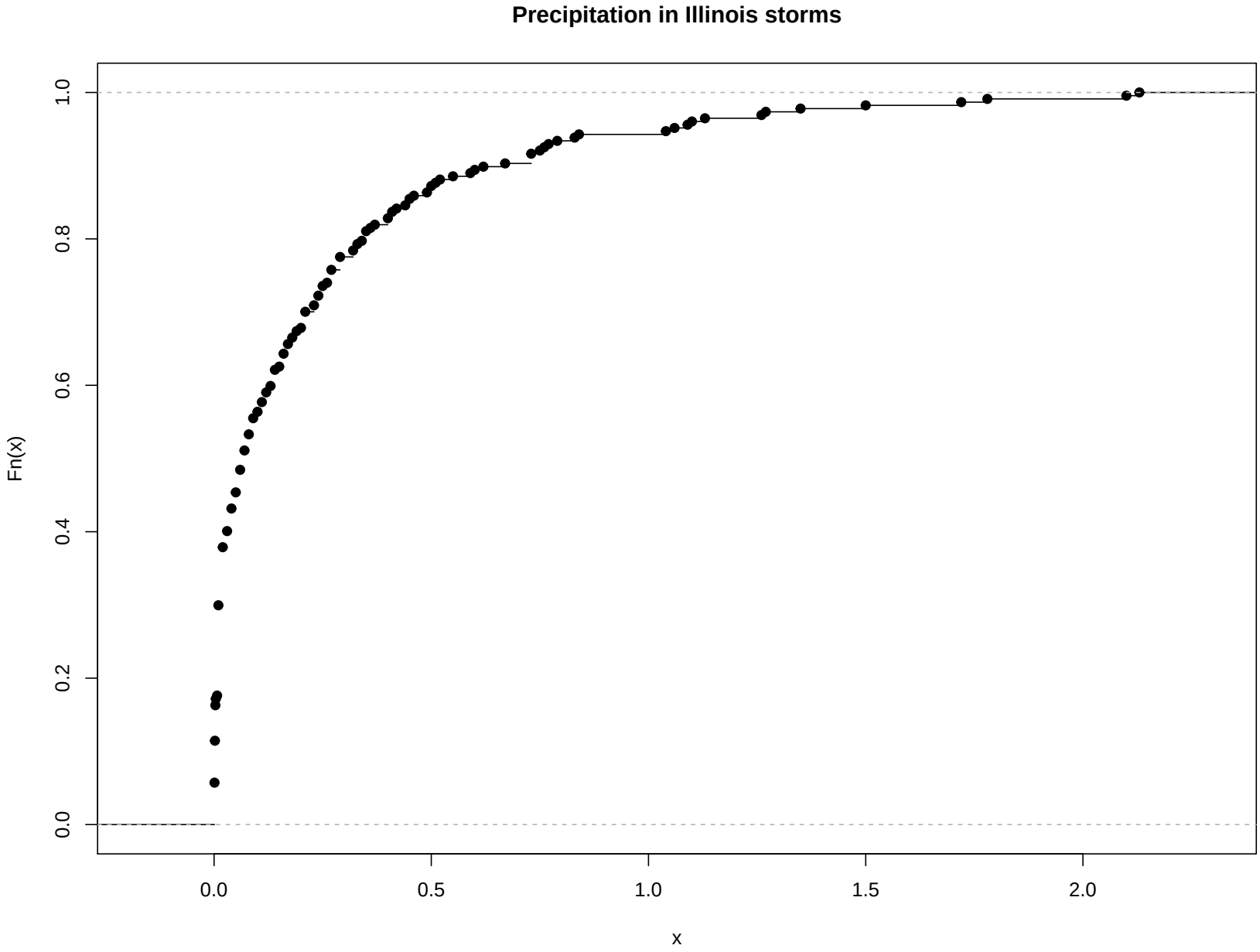
$$\begin{cases} \text{has a jump at } x \text{ if there is an observation with value } x, \text{ and} \\ \text{the jump is of size } \frac{i}{n} \text{ if there are } i \text{ observations with value } x \end{cases}$$

Can conventiently read of percentiles from the ecdf.
The median is close to the value x such that $F_n(x) = \frac{1}{2}$.
The pth percentile is close to the value x such that $F_n(x) = \frac{p}{100}$.

Empirical cdf. Heat of sublimation for Iridium and Rhodium.



Empirical cdf. Precipitation in Illinois storms.



Nonparametric bootstrap.

We first recall the **parametric bootstrap** method.
Suppose we are interested in a measure of location (e.g. mean, median, trimmed mean) of the data $x_1, \ldots, x_n$. Data is assumed to be independent random sample from some (unknown) distribution with cdf F_θ.
(E.g. we assumed F_θ to be a gamma(α, λ) cdf in the precipitation data in storms in Illinois.)

Nonparametric bootstrap.

An estimator, $\hat{\theta}$ say, for the measure of location θ is usually straightforward to compute from $x_1, \ldots, x_n$. In particular, $\hat{\theta}$ is a function of $X_1, \ldots, X_n$ and therefore random. In the parametric bootstrap approximated the distribution of $\hat{\theta}$ by bootstrap simulations as follows.

1. Use $F_{\hat{\theta}}$ as an approximation for F_θ.
   (Since we don't know the 'true' cdf F_θ from which the data is sampled.)
2. Draw n independent samples $x_1^\star, \ldots, x_n^\star$ from $F_{\hat{\theta}}$
   Compute $\theta = \theta(x_1^\star, \ldots, x_n^\star)$ and denote this value by $\theta_1^\star$.
   Repeat B times to obtain $\theta_1^\star, \theta_2^\star, \ldots, \theta_B^\star$
   take empirical distribution of $\theta_1^\star, \theta_2^\star, \ldots, \theta_B^\star$ as a good approximation to distribution of $\hat{\theta}$.

The empirical distribution of $\theta_1^\star, \theta_2^\star, \ldots, \theta_B^\star$ approximates distribution of $\hat{\theta}$ better and better as we increase B.

Nonparametric bootstrap.

Nonparametric bootstrap. How does the nonparametric bootstrap differ from the parametric bootstrap?
In the nonparametric bootstrap we do *not* assume that the true cdf F from which the data $x_1, \ldots, x_n$ are sampled is contained in a parametric model F_θ for some θ.
Instead, in the nonparametric bootstrap one uses the empirical cdf

$$F_n(x) = \frac{1}{n}\#\{i \colon x_i \leq x\}$$

as an approximation for F.

Nonparametric bootstrap.

Consequently, the distribution of some estimator $\hat{\theta}$ is approximated in the nonparametric bootstrap as follows.

1. Use F_n as an approximation for F.
   (Since we don't know the 'true' cdf F from which the data is sampled.)
2. Draw n independent samples $x_1^\star, \ldots, x_n^\star$ from F_n.
   Compute $\theta = \theta(x_1^\star, \ldots, x_n^\star)$ and denote this value by $\theta_1^\star$.
   Repeat B times to obtain $\theta_1^\star, \theta_2^\star, \ldots, \theta_B^\star$.
   Take empirical distribution of $\theta_1^\star, \theta_2^\star, \ldots, \theta_B^\star$ as a good approximation to the distribution of $\hat{\theta}$.

Again, the aproximation of the distribution of $\hat{\theta}$ by empirical distribution of $\theta_1^\star, \theta_2^\star, \ldots, \theta_B^\star$ becomes better as we increase B.

Nonparametric bootstrap.

Example. Heat of sublimation for Iridium/Rhodium.
Bootstrapping mean and median.

- Which of the two estimators, for mean or median, do you expect to have a larger spread/standard error?
- Is the sampling distribution for the mean more spread-out for the Iridium or the Rhodium data?

R: bootstrap, histograms of sampling distributions.