Lecture 04

Variability (dispersion) in data:

- **Consider the following two datasets**
  - **Dataset 1: -2, -1, 0, 1, 2**
  - **Dataset 2: -300, -100, 0, 100, 300**

Sample range R:

- **a numerical descriptive statistic of variability**
- **applicable given univariate data $x_1, \ldots, x_n$**
- $R = x_{(n)} - x_{(1)}$
- **not so widely used anymore**
- **based on only two data values**

Sample variance s^2:

- **a numerical descriptive statistic of variability**
- **applicable given univariate data** $x_1, \ldots, x_n$
- $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \;=\; \frac{(\sum x_i^2) - n\bar{x}^2}{n-1}$
- $s^2 \geq 0$**;** $\quad s^2 = 0$ **corresponds to** $x_1 = \cdots = x_n$
- **large** s^2 **corresponds to widely spread data**
- **note that denominator is** $n - 1$ **instead of** n
- **think about why the difference** $x_i - x_n$ **is squared**
- **distinguish between the two formulae**
- **note that** s^2 **is measured in squared units**
- **the sample standard deviation is given by** s

How do location/scale changes affect $\bar{x}$ and s^2:

- **i.e.** $x_i \rightarrow y_i = a + bx_i$

- **e.g. changing Celsius data to Fahrenheit**

Problem: Can you construct a dataset with $R = 30$ and $s^2 = 100$?

Problem: $n = 5$**,** $x_1 = 10, x_2 = 3, x_3 = 7, x_4 = 8$

(a) If $\bar{x} = 6$**, obtain** x_5**.**

(b) If $s = 5$**, obtain** x_5**.**

Boxplots:

- **a graphical descriptive statistic**
- **applicable given univariate data (in groups)**
- **generated by statistical software**
- **calculations require $\tilde{x}$, lower fourth, upper fourth**
- **interpreting boxplots is our focus**
- **boxplots are not as popular as they should be**

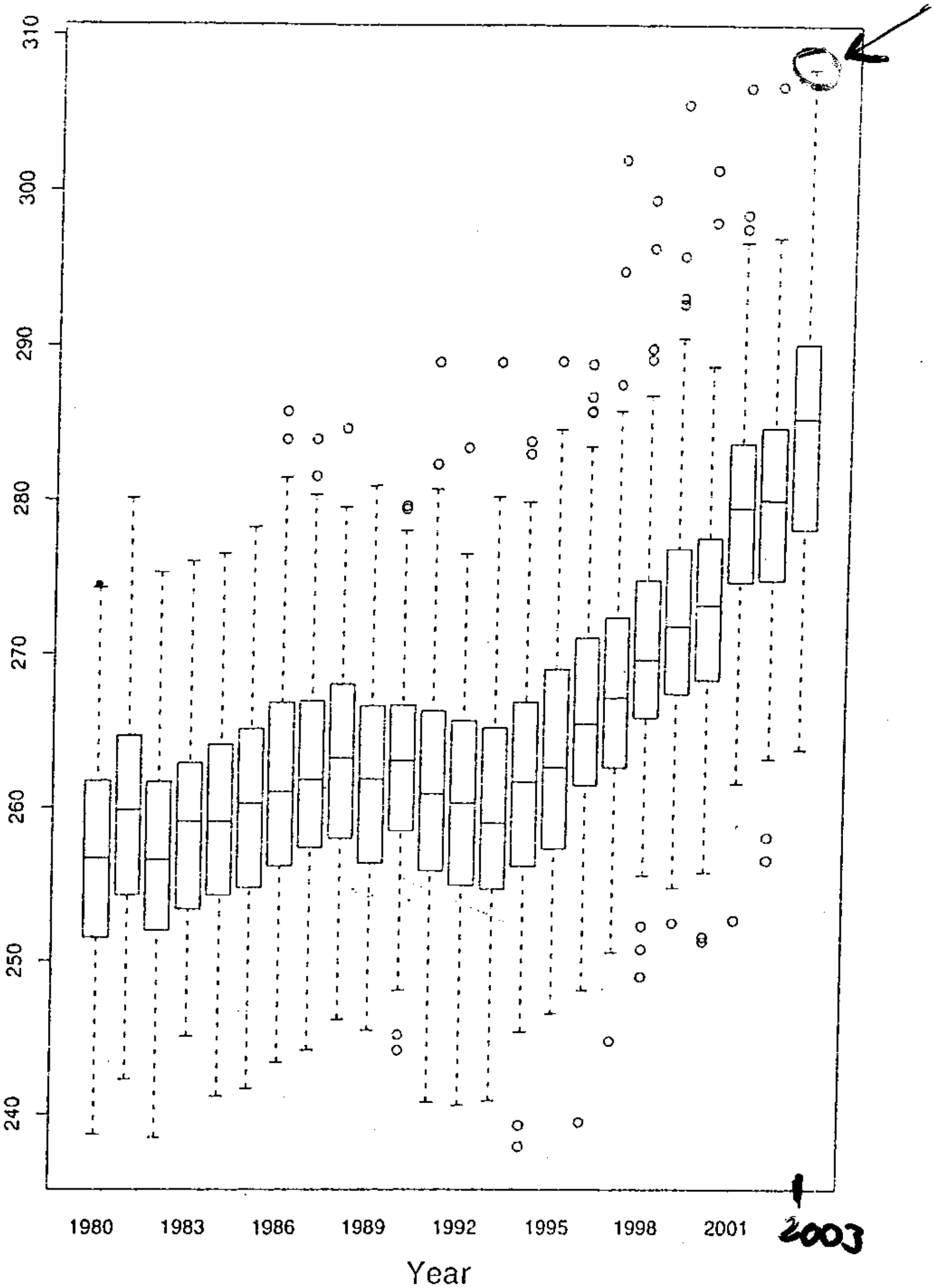

Figure 1. Boxplots for driving distances from 1980 through 2003.