Chapters 2. Graphics and Simple Numerical Techniques (B)

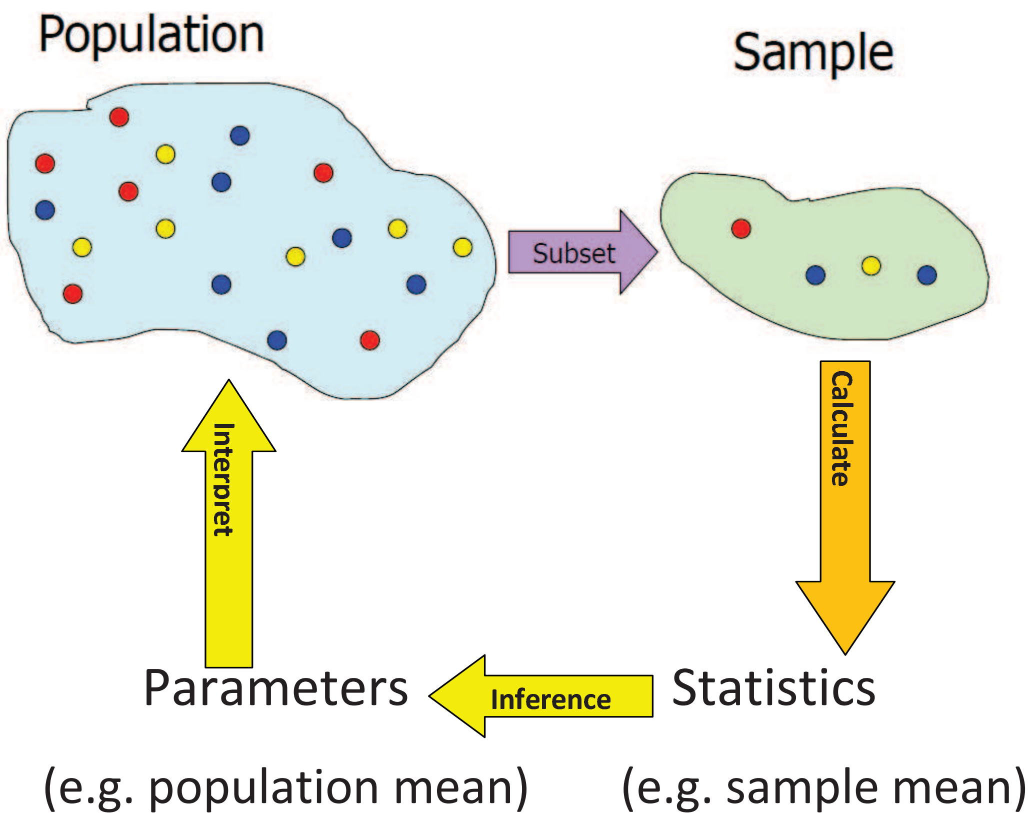
January 19, 2011

1 Numerical Techniques

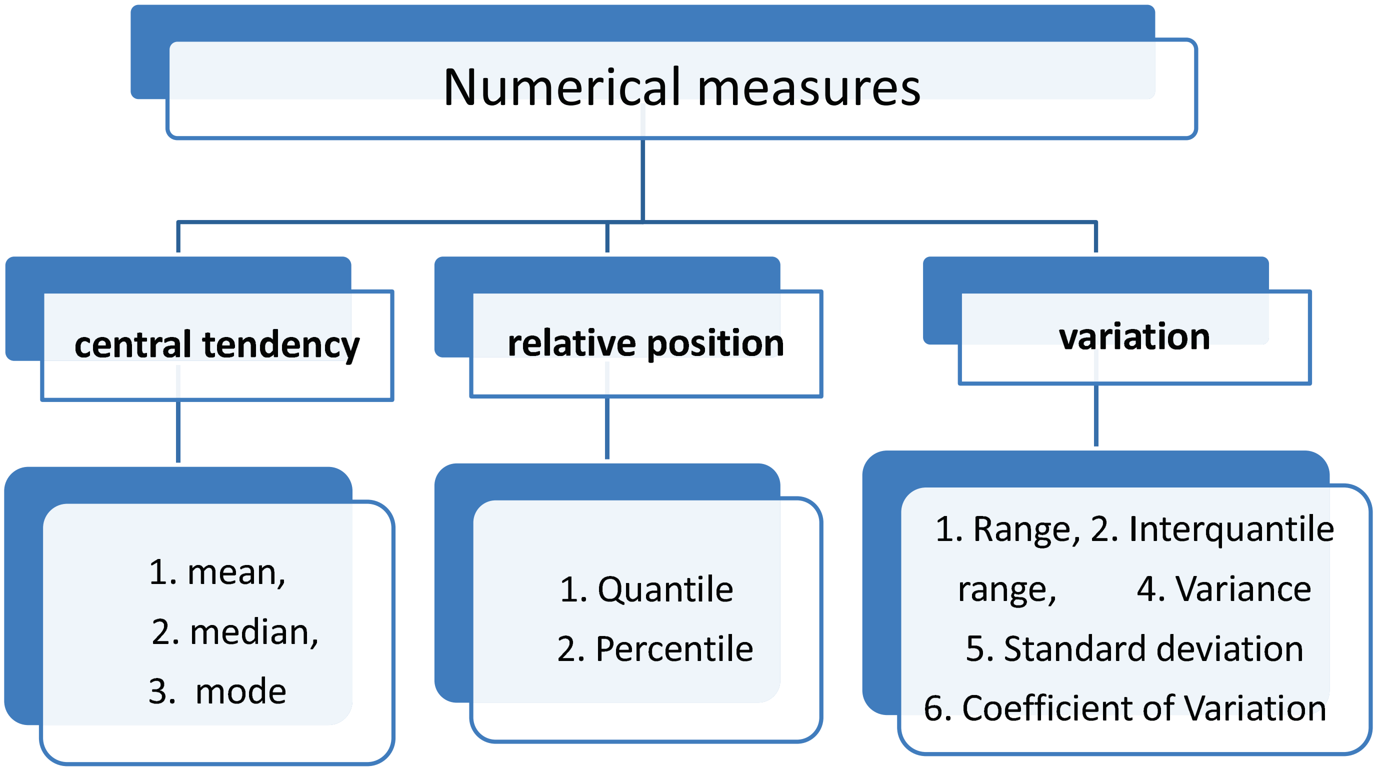
Parameter—a numerical measure that describes a characteristic of a population. E.g. mean/expectation, variance, ...

Statistic —a numerical measure that describes a characteristic of a sample. E.g. sample mean, sample variance, ...

A full procedure for statistics



“inference” includes estimating, judging ....


Numerical measures
central tendency
relative position
variation
1. mean,
2. median,
3. mode
1. Quantile
2. Percentile
1. Range, 2. Interquantile range,
4. Variance
5. Standard deviation
6. Coefficient of Variation

1.1 Central tendency

- Mean

  - Population mean, denoted by μ For a finite size population $\{x_1, ..., x_N\}$, the population mean is defined

$$\mu = \frac{1}{N}(x_1 + ... + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

e.g. in Example A above, we have

$$\mu = \frac{1}{80}(76.4 + 76.1 + ... + 75.9) = 75.965kg$$

(for population with infinite number of individuals, the calculation of mean will be discussed later)

– sample mean, denoted by $\bar{x}, \bar{y}, ...$ For sample $\{x_1, ..., x_n\}$, its sample mean

$$\bar{x} = \frac{1}{n}(x_1 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

e.g. in the above sample of Example A, we have

$$\begin{aligned} \bar{x} &= \frac{1}{9}(75.6 + 75.7 + 75.8 + 75.9 + 76 + 76.1 + 76.1 + 76.2 + 76.3) \\ &= 75.967kg \end{aligned}$$

– sample/population mean for grouped/repeated data. Suppose the different values are $x_1, x_2, ..., x_k$ with corresponding frequencies n_i and relative frequencies p_i

$$\bar{x}(or\ \mu) = \frac{\sum_{i=1}^{k} n_i x_i}{n(or\ N)} = \sum_{i=1}^{k} p_i x_i$$

Fact. If data $\{x_1, x_2, ..., x_n\}$ has mean $\bar{x}$, the differences $x_1 - \bar{x}, x_2 - \bar{x}, ..., x_n - \bar{x}$ are called the deviations (from the mean), then

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

For grouped/repeated data

$$\sum_{i=1}^{k} f_i(x_i - \bar{x}) = 0, \quad \sum_{i=1}^{k} p_i(x_i - \bar{x}) = 0$$

where k is the number of different values

Example A(continued). for the population

$$\begin{aligned}\mu &= 75.6 * 0.075 + 75.7 * 0.100 + 75.8 * 0.100 \\ &\quad +75.9 * 0.175 + 76.0 * 0.225 + 76.1 * 0.1375 \\ &\quad +76.2 * 0.100 + 76.3 * 0.0625 + 76.4 * 0.025 \\ &= 75.965\end{aligned}$$

R commands,

```
> mydata = c(76.4, 76.1, 76.2, 75.9, 75.9, 75.9, 76.1, 75.8, 76.0, 76.0, 75.6, 76.1, 76.0, 75.9,
76.3, 75.8, 75.6, 76.2, 76.0, 75.9, 75.8, 76.0, 76.0, 75.7, 76.2, 75.9, 76.0, 76.0, 75.9, 76.0, 75.7, 75.8,
76.0, 75.7, 76.2, 76.0, 75.6, 75.7, 76.2, 76.0, 75.9, 76.4, 76.1, 75.9, 75.6, 75.7, 76.2, 76.1, 75.9, 76.3,
75.8, 75.9, 76.1, 76.1, 76.0, 76.1, 75.7, 75.7, 76.0, 75.8, 76.2, 75.9, 76.1, 76.2, 75.6, 75.6, 76.1, 76.0,
76.0, 75.9, 76.1, 76.3, 75.8, 76.3, 76.0, 76.3, 75.7, 75.8, 76.0, 75.9)

> mu = mean(mydata)

> mu

> mysample = c(75.6,75.7, 75.8, 75.9, 76, 76.1, 76.1, 76.2,
76.3)
```

```
> xbar = mean(mysample)

> xbar
```

For grouped/repeated data, we can use

```
> mygroup = c(75.6, 75.7, 75.8, 75.9, 76, 76.1, 76.2, 76.3,
76.4)

> myfreq = c(6, 8, 8, 14, 18, 11, 8, 5, 2)

> weighted.mean(mygroup, w=myfreq)
```

- Median For n (or finite N) values $x_1, x_2, ..., x_n$, the median is the "middlemost" value once the data are arranged according to the value, called ordered values (or order statistics). Specifically if n is odd,

$$\text{median} = \text{the ordered value at position } \frac{n+1}{2}$$

if n is even,

$$\text{median} = \text{average of 2 values at } \frac{n}{2} \text{ and } \frac{n+2}{2}$$

Medians for sample and for finite population are computed the same way.

Example A(continued). After rearranging the population, 75.6(smallest), 75.6(second smallest), 75.6, 75.6, 75.6, ..., 76.0, 76.0, 76.0(40'th smallest), 76.0(41'th smallest), 76.0, ..., 76.3, 76.4, 76.4(largest), then the population median

$$\text{median} = \frac{76.0 + 76.0}{2} = 76.0(kg)$$

For the sample, after arranging, 75.6(smallest), 75.7(second smallest), 75.8(3rd smallest), 75.9(4th smallest), 76(5th smallest), 76.1(6th smallest), 76.1(7th smallest), 76.2(8th smallest), 76.3(largest), the sample median is

$$\text{median} = 76.0(kg)$$

R commands,

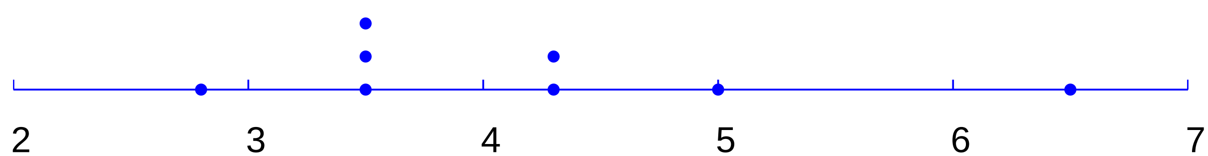
```
> mydata = c(76.4, 76.1, 76.2, 75.9, 75.9, 75.9, 76.1, 75.8, 76.0, 76.0, 75.6, 76.1, 76.0, 75.9,
76.3, 75.8, 75.6, 76.2, 76.0, 75.9, 75.8, 76.0, 76.0, 75.7, 76.2, 75.9, 76.0, 76.0, 75.9, 76.0, 75.7, 75.8,
76.0, 75.7, 76.2, 76.0, 75.6, 75.7, 76.2, 76.0, 75.9, 76.4, 76.1, 75.9, 75.6, 75.7, 76.2, 76.1, 75.9, 76.3,
75.8, 75.9, 76.1, 76.1, 76.0, 76.1, 75.7, 75.7, 76.0, 75.8, 76.2, 75.9, 76.1, 76.2, 75.6, 75.6, 76.1, 76.0,
76.0, 75.9, 76.1, 76.3, 75.8, 76.3, 76.0, 76.3, 75.7, 75.8, 76.0, 75.9)
> median(mydata)
```

- Mode — the value that occurs most often in the data.

  Example Monthly salaries for its staff members in a small company, 2.8K$, 3.5K$, 3.5K$, 3.5K$, 4.3K$, 4.3K$, 5K$, 6.5K$. Then

$$mode = 3.5(K\$)$$

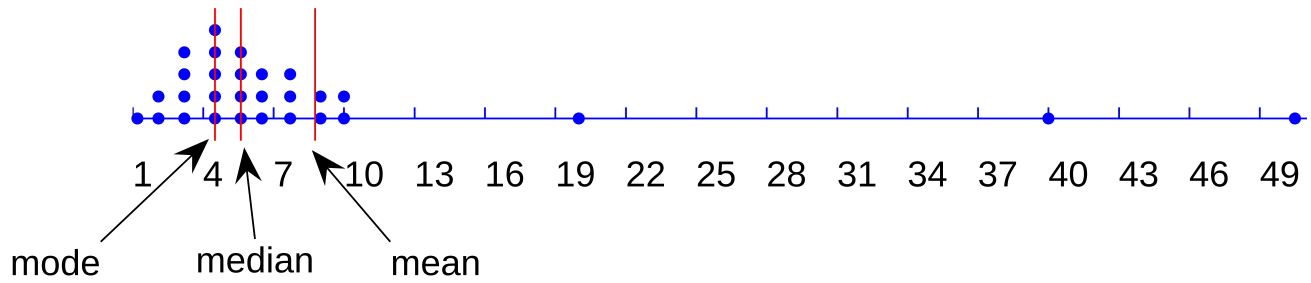
  It is more clear if we represent the data as follows



Comparison of mean, median and mode All reflect the central tendency, and are used to summarize the central tendency of the data by one value.

- If a distribution is symmetrical, the mean, median and mode may coincide
- The median and the mode are not affected by to the extreme values (i.e. very large or very small compared with the rest) at the ends.
- Except for some special case, mean is still more commonly used than median and mode.

Example. For people's incomes, the distribution usually looks as below, where median or mode is better than the mean in the sense of representing the majority!



Dont be disappointed when you find your income is below the average!

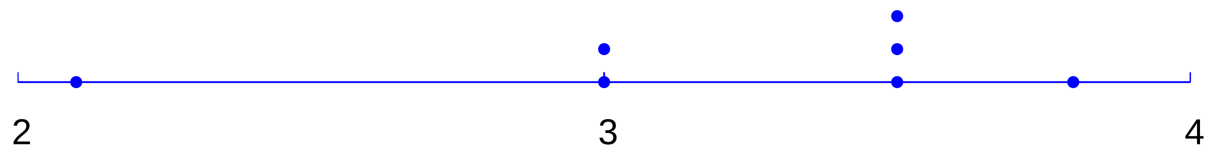
1.2 Dispersion

A measure of statistical dispersion is a nonnegative real number that is zero if all the data are identical, and increases as the data becomes more diverse.

- Range—maximum value (denoted by $X_{max}, Y_{max}, ...$) minus smallest value (denoted by $X_{min}, Y_{min}, ...$) in a data, i.e. $Range = X_{max} - X_{min}$

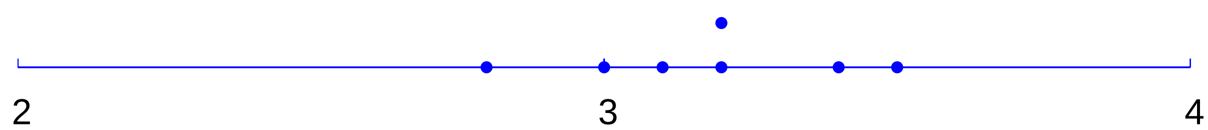
  For data A: 3cm, 3cm, 2.1cm, 3.5cm, 3.5cm, 3.5cm, 3.8cm, $X_{max} = 3.8cm, \quad X_{min} = 2.1cm$. So

$$Range = 3.8 - 2.1 = 1.7(cm)$$



For data B, 2.8cm,3cm, 3.1cm, 3.2cm, 3.2cm, 3.4cm, 3.5cm. we have

$$Range = 3.5 - 2.8 = 0.7(cm)$$



Conclusion: Data A has bigger dispersion than Data B.

Disadvantage of Range:

– Ignore the distribution between the maximum and the minimum;

– very sensitive to the extreme values

E.g. consider data M: 1.5, 2.4,2.4,2.4, 2.4, 3.2 and data N: 1.5, 1.6,2.0,2.2, 2.8, 3.2? The 2 data should have different variation, but their ranges are the same!

- Variance —average of the squared deviations from the mean ($\bar{x}$ or μ).Population variance is denoted by σ^2, sample variance s^2

  Notation comparison between population and sample

	Population	**Sample**
Size	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	s^2

population mean

The variance of a finite size **population** is:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

population size

sample mean

The variance of a **sample** is:

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

Note! the denominator is sample size (n) minus one !

For grouped/repeated data of finite population

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i(x_i - \mu)^2}{N} = \sum_{i=1}^{k} p_i(x_i - \mu)^2,$$

where $n_1 + ... + n_k = N$. For grouped/repeated data of sample

$$s^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{n - 1} = \frac{n}{n-1}\sum_{i=1}^{k} p_i(x_i - \bar{x})^2,$$

where $n_1 + ... + n_k = n$.

Example C. The sampled delay times (handling, setting, and positioning the tools) for cutting 6 parts on an engine lathe are 0.6, 1.2, 0.9, 1.0, 0.6, and 0.8 minutes. calculate s^2

$$\bar{x} = \frac{0.6 + 1.2 + 0.9 + 1.0 + 0.6 + 0.8}{6} = 8.5$$

observations	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	0.6	-0.25	0.0625
2	1.2	0.35	0.1225
3	0.9	0.05	0.0025
4	1.0	0.15	0.0225
5	0.6	-0.25	0.0625
6	0.8	-0.05	0.0025
sum	5.1	0.00	0.2750

$s^2 = \frac{0.2750}{6-1} = 0.055$(minute^2)

Alternative formulation for the variance

- $$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

- population variance with finite size

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

- Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2.$$

- Grouped/repeated data

$$\sigma^2 = \sum_{i=1}^{k} p_i x_i^2 - \mu^2, \qquad s^2 = \frac{n}{n-1}\sum_{i=1}^{k} p_i x_i^2 - \frac{n}{n-1}\bar{x}^2,$$

- Standard deviation. Square-root of variance

$$s = \sqrt{s^2}, \qquad \sigma = \sqrt{\sigma^2}$$

s and σ have the same unit as the original data.

E.g. in industry, 3-σ rule is applied. A product cannot departure too far away from the central. Otherwise, it could be a defect.

R commands: `var(data), sd(data) --- only for sample variance and sample standard deviation`

```
var(c(0.6,1.2,0.9, 1.0, 0.6,0.8))

sd(c(0.6,1.2,0.9, 1.0, 0.6,0.8))
```

Example (continued) For data A and Data B above, if we treat them as population

$$\mu_A = 3.2000, \qquad \sigma_A^2 = 0.2743, \qquad \sigma_A = 0.5237(cm)$$

$$\mu_B = 3.1714, \qquad \sigma_B^2 = 0.0478, \qquad \sigma_B = 0.2185(cm)$$

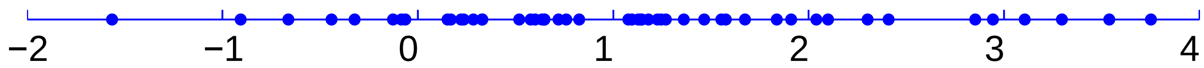
if we treat them as sample

$$\bar{x}_A = 3.2000, \qquad s_A^2 = 0.3200(cm^2), \qquad s_A = 0.5657(cm)$$

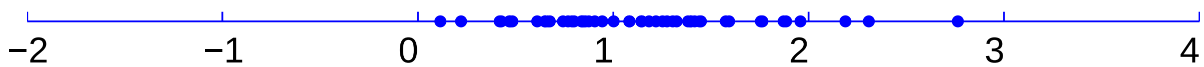
$$\bar{x}_B = 3.1714, \qquad s_B^2 = 0.0557(cm^2), \qquad s_B = 0.2360(cm)$$

Another look at the standard deviation/variance

Data E:

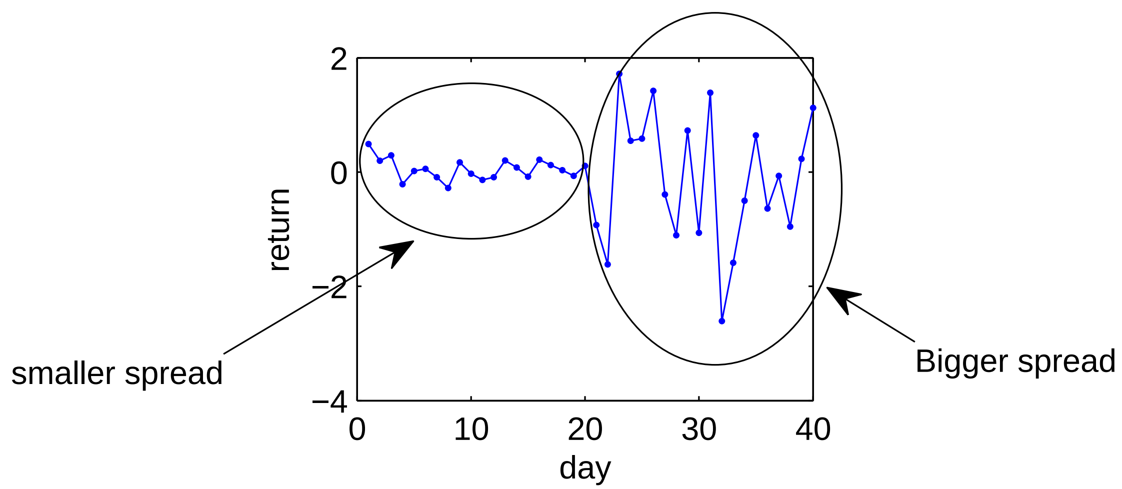


Data F:



Data E has bigger standard deviation/variance than data F

Example 1. (application in financial study) Suppose the following data are observed daily in 40 days 0.4893, 0.1997, 0.2898, -0.2113, 0.0174, 0.0582, -0.0917, -0.2806, 0.1707, -0.0288, -0.1359, -0.0902, 0.2009, 0.0776, -0.0830, 0.2157, 0.1236, 0.0316, -0.0688, 0.1059, -0.9277, -1.6168, 1.7187, 0.5464, 0.5871, 1.4227, -0.3915, -1.1076, 0.7268, -1.0649, 1.3864, -2.6094, -1.5886, -0.4994, 0.6422, -0.6385, -0.0676, -0.9566, 0.2317, 1.1240. We plot them against the date. What can you observe?



For the first 20 values, their sample standard derivation is

$$s_{_I} = 0.1833$$

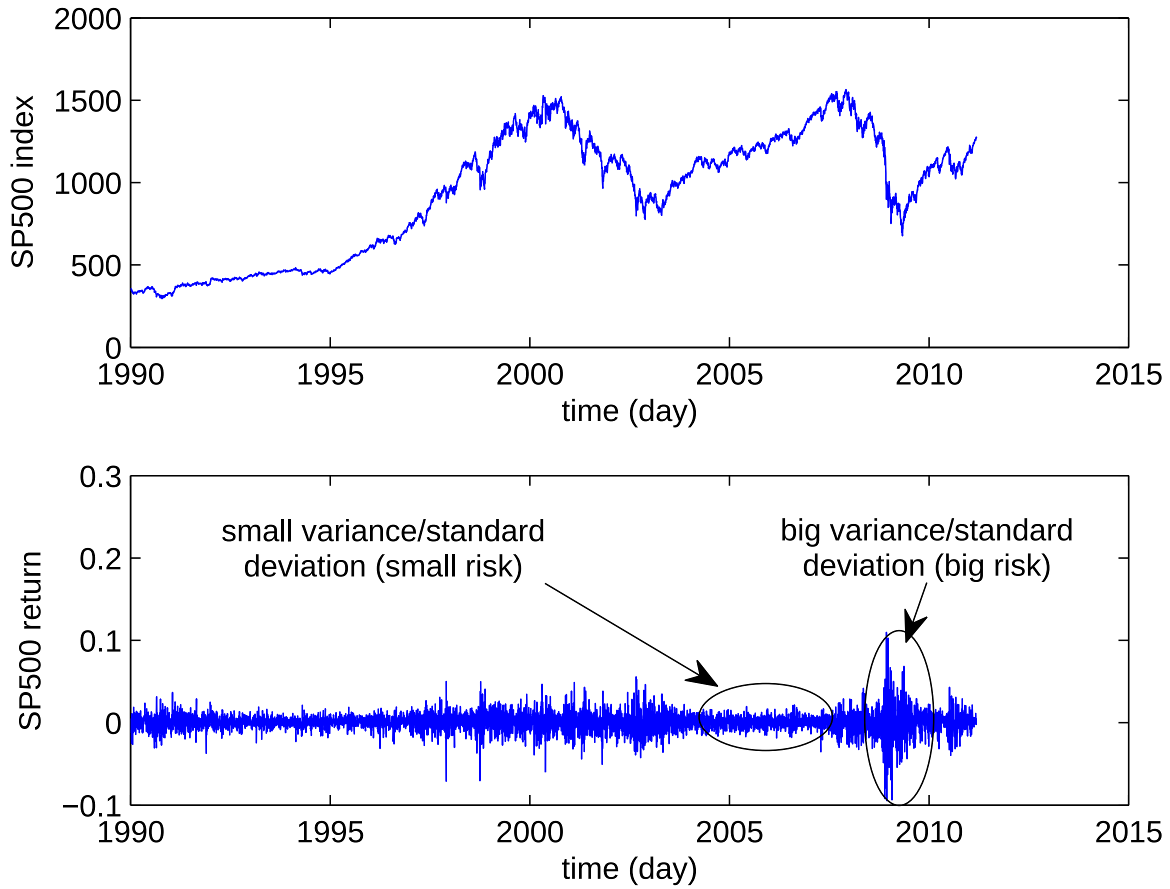
and for the second

$$s_{_{II}} = 1.1748$$

indicating that the second half has much wider spread than the first.

Now consider the SP500 daily index from 1990 to yesterday. The return of a day is defined as

$$r_t = \log(\text{index of the day: } t) - \log(\text{index of the previous day: } t-1)$$



In finance, the standard deviation can measure the risk of investment[1]

[1] Modeling of the risk, called GARCH model, is the work of the 2003 Nobel Memorial Prize in Economic Sciences.

- Coefficient of variation. To compare the variation in several sets of data (with substantial difference in their means), it is generally desirable to use a measure of relative variation. The coefficient of variation gives the standard deviation as a percentage of the mean.

$$V = \frac{s}{\bar{x}} \qquad \text{or} \qquad V = \frac{\sigma}{\mu}$$

or

$$V = 100 \times \frac{s}{\bar{x}}\% \qquad \text{or} \qquad V = 100 \times \frac{\sigma}{\mu}\%$$

Example D. The average incomes in regions A and B are respectively

$$\mu_A = 10000\$, \qquad \mu_B = 1000\$$$

and their standard derivations are respectively

$$\sigma_A = 200\$, \qquad \sigma_B = 100\$$$

Then their coefficients of variation are

$$V_A = 0.02, \qquad V_B = 0.1.$$

indicating that the gap between rich and poor in B is bigger than that in A.

1.3 Quartiles and Percentiles

The 100p'th percentile is a value such that at least $100p\%$ of the observations are at or below this value, and at least $100(1-p)\%$ are at or above this value.

first quartile	$Q_1 = 25'$th percentile
second quartile	$Q_2 = 50'$th percentile (i.e. median)
third quartile	$Q_3 = 75'$th percentile

Calculating the sample (finite population) 100pth percentile:

- Order the n observations from smallest to largest.
- Determine the product np.

  If np is not an integer, round it up to the next integer and find the corresponding ordered value.

  If np is an integer, say k, calculate the mean of the k'th and $(k+1)$'st ordered observations.

Example E. The ordered heights of the nanopillars are

221	234	245	253	265	266	271	272	274	276
276	276	278	284	289	290	290	292	292	296
297	298	300	303	304	305	305	308	308	309
310	311	312	314	315	315	323	330	333	336
337	338	343	346	355	364	366	373	390	391

find $Q_1, Q_2, Q_{0.93}$?

$$np = 50 * 0.25 = 12.5$$

the rounded position is 13

$$Q_1 = 278(nm).$$

Since $p = 0.5$ for the second quartile, or median,

$$np = 50 * 0.5 = 25$$

which is an integer. Therefore, we average the 25th and 26th ordered values

304 + 305

$$Q_2 = 304.5(nm)$$

Note that

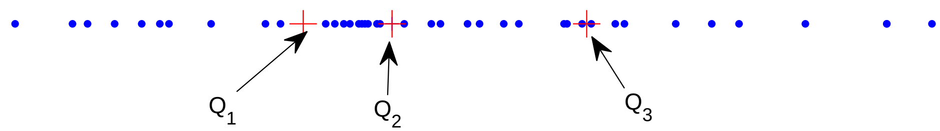
$$np = 50 * 0.93 = 46.5,$$

which we round up to 47. Counting to the 47th position, we obtain

$$P_{0.93} = 366(nm).$$

Interquartile range is also a measure of the dispersion

$$\text{Interquartile range} = \text{third quartile - first quartile} = Q_3 - Q_1$$

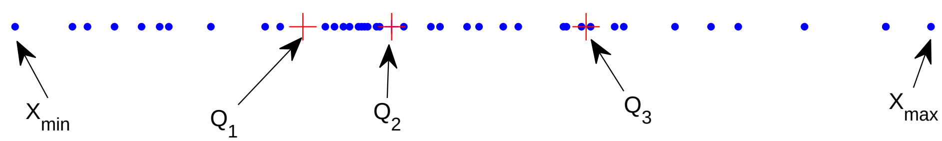


Advantage over the Range

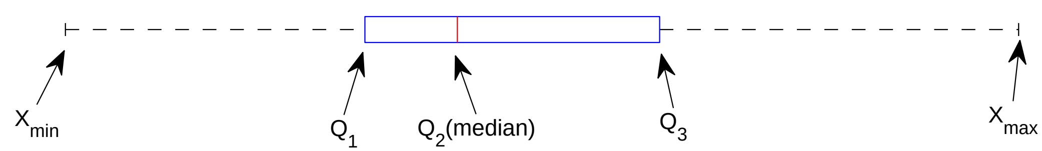
- Eliminate high and low valued observations and calculate the range from the remaining values
- Eliminate problems caused by 'extreme values'

1.4 Boxplot and shape of distribution

Five-number $(X_{min}, Q_1, Q_2, Q_3, X_{max})$ summary



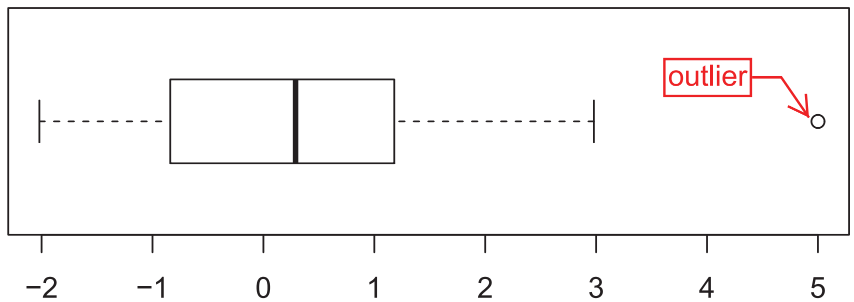
Boxplot: Graphical distribution of the Five-number summary



We can apply Boxplot to

- Identify the possible outliers if either
  - observations fall more than $1.5 \times (Q_3 - Q_1)$ below Q_1
  - observations fall more than $1.5 \times (Q_3 - Q_1)$ above Q_3

E.g. -0.17, -2.02, -1.18, -1.56, -0.28, 1.40, -0.98, 0.29, 0.42, 1.18, 1.07, 0.22, 2.98, 0.47, -0.75, -1.75, 1.72, 0.99, -0.84, 1.76, 5.00. Value 5.00 is possibly an outlier.

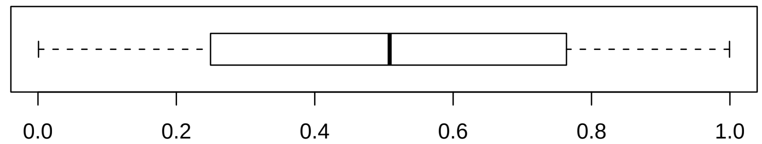


R code:

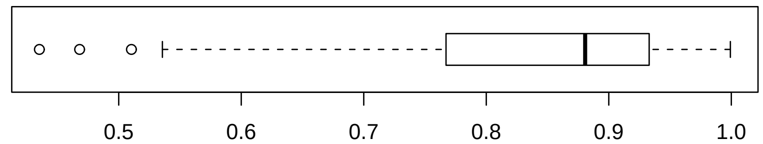
```
x = c(-0.17, -2.02, -1.18, -1.56, -0.28, 1.40, -0.98, 0.29, 0.42, 1.18, 1.07, 0.22, 2.98, 0.47, -0.75, -1.75, 1.72, 0.99, -0.84, 1.76, 5.00)
boxplot(x, horizontal = TRUE)      # or boxplot(x, horizontal = FALSE)
```

(since there is one outlier, it is excluded and the boxplot is drawn without it)

- Tell the shape of distribution
  - symmetrical



  - skewed to the left



  - skewed to the right