RenR 480/711

Statistical Vocabulary Part II

& Descriptive statistics

Statistical Vocabulary Part II

Descriptive statistics

- Numerical/graphical summary of data

Inferential statistics

- Predict or control, draw conclusions
- e.g. Regression: Petal.Length~Sepal.Length
- e.g. Anova: Maximize lentil-yield

Distribution

- Probability associated with each possible value of a variable

Parameter

- Population characteristic (e.g. mean yield), unknown

Statistic

- Estimation of a parameter (e.g. sample mean)

Descriptive Statistics

Measures to describe the distribution of a variable

Three categories:

- Center
- Spread
- Shape

Measures of center

The Arithmetic Mean:

Populations mean

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

(different notation)

Sample mean

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

x_i = observation i

N/n = population/sample size

$\mu/\bar{x}$ = population/sample mean

$\sum_{i=1}^{n}$ = sum all observations from 1 to n

Important: The population mean can never be known, we can only estimate a sample mean

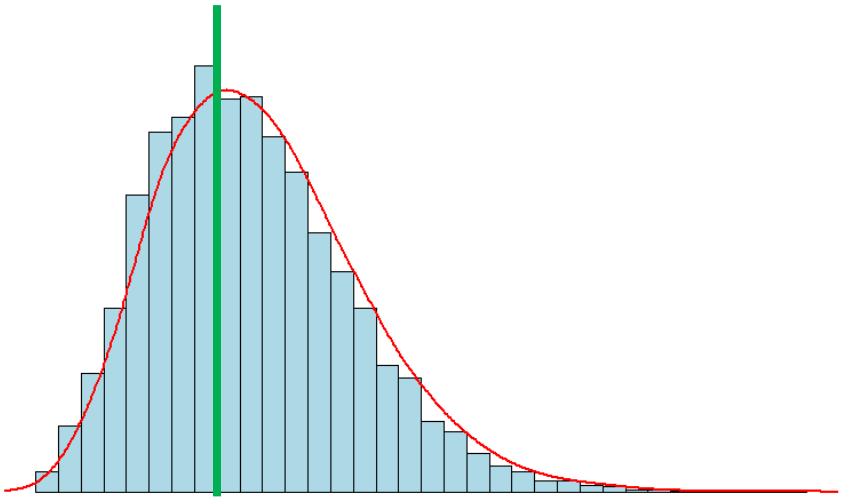
Measures of center

The Median:

50^{th} percentile of observations

Unaffected by outliers and skewed distributions

Example: Median family income

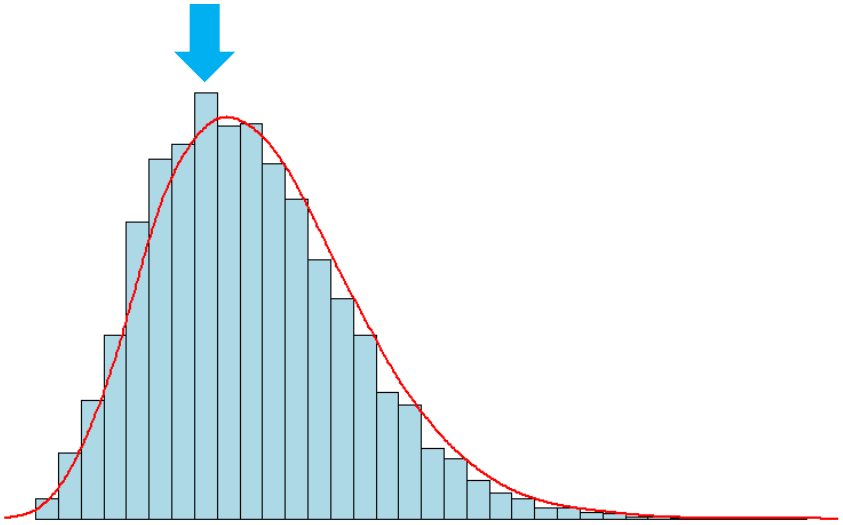


Measures of center

The Mode:

Highest frequency of value or class

Example: Most popular food



Measures of center

The Weighted Mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

x_i = observation i

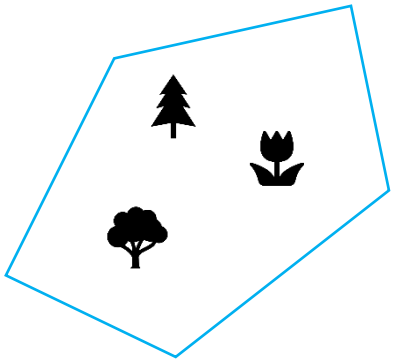
n = sample size

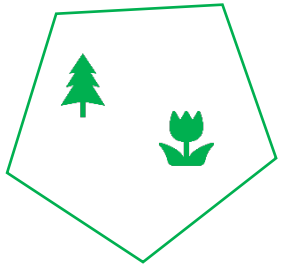
$\bar{x}$ = sample mean

$\sum_{i=1}^{n}$ = sum all observations from 1 to n

w_i = weight of observation i

Example: Correcting geographical bias: 10 samples per ecosystem, but each ecosystem has a different size.





Ecosystem 1
10 Species
20 km²

Ecosystem 2
5 Species
8 km²

$$\bar{x} = \frac{\left(\frac{20}{28}\right)*10+\left(\frac{8}{28}\right)*5}{\left(\frac{20}{28}\right)+\left(\frac{8}{28}\right)} = 8.57$$

Measures of center

The harmonic mean:

Average of ratios

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Example:
You drive 20 km in your car. The first 10 km you go 40 km/h and the second 10 km you go 60 km/h.

What is your average speed?

$$\bar{x}_h = \frac{2}{\left(\frac{1}{40}\right) + \left(\frac{1}{60}\right)}$$

$$\bar{x}_h = 48 \text{ km/h}$$

Measures of center

The Geometric mean:

Average of products

$$\bar{x}_g = \sqrt[n]{x_1 * x_2 * \cdots * x_n}$$

Example:
Your invest in the stock market and make 10% in the first, 25% in the second and 15% in the third year.

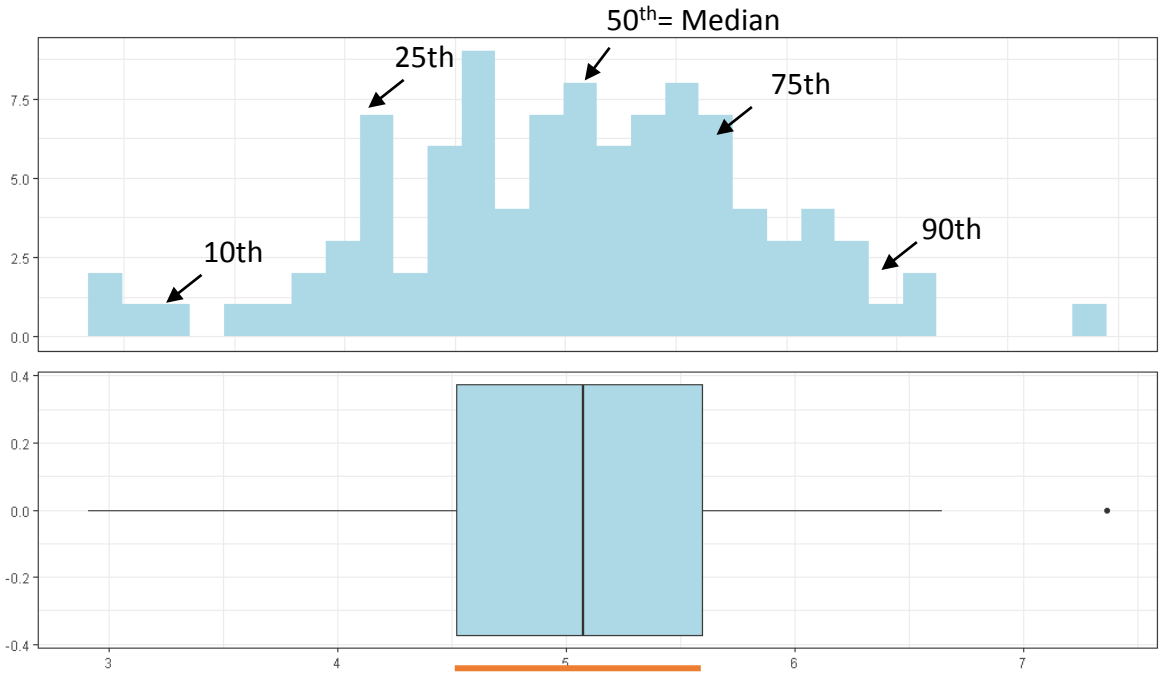
What is your average return?

$$\bar{x}_g = \sqrt[3]{1.10 * 1.25 * 1.15}$$

$$\bar{x}_g = 1.165$$

Measures of spread

Percentiles: % of observations below



Interquartile Range (IQR)

Range minus outliers

more than 1.5*IQR = moderate outlier

more than 3*IQR = far outlier outlier

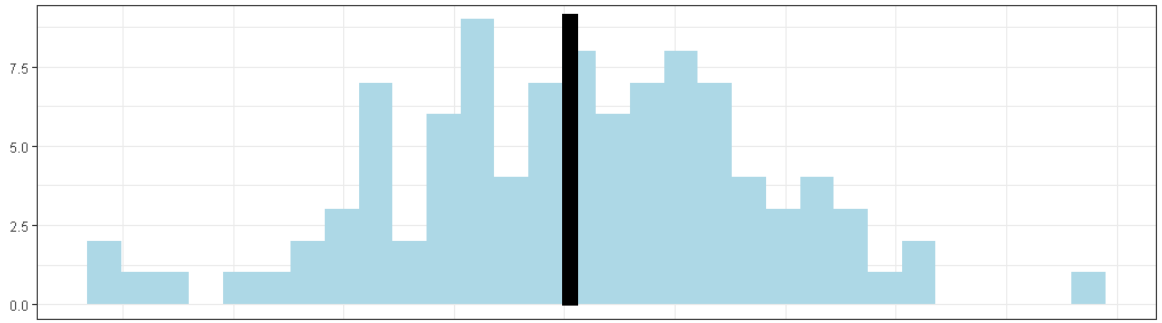
Measures of spread

Variance and standard deviation:

Based on deviations from the mean:

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

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



x_i = observation i

$\mu/\bar{x}$ = population/sample mean

N/n = population/sample size

$\sum_{i=1}^{n}$ = sum all observations from 1 to n

σ^2/s^2 = population/sample variance

σ/s = population/sample standard deviation

Measures of spread

Variance and standard deviation:

Calculated from deviations:

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

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

ID	FARM	VARIETY	YIELD	Deviations $x_i - \bar{x}$	Squared deviation $(x_i - \bar{x})^2$
1	Farm1	A	720	-7.5	56.25
2	Farm1	A	740	12.5	156.25
3	Farm1	A	690	-37.5	1406.25
4	Farm1	A	760	32.5	1056.25
		$\bar{x}$=	727.5	0	s^2= 891.67
					s = 29.86

Measures of spread

Coefficient of variation:

Standard deviation as a percentage of the mean

$$\mathrm{CV} = \frac{s}{\bar{x}}$$

$\bar{x}$ = sample mean

s = standard deviation

For example:
Comparing variability in height and diameter of trees:

Diameter in cm

$\bar{x} = 30\mathrm{cm}$

$\mathrm{s} = 4\mathrm{cm}$

$\mathrm{CV} = 0.13$ or 13%

Height in m

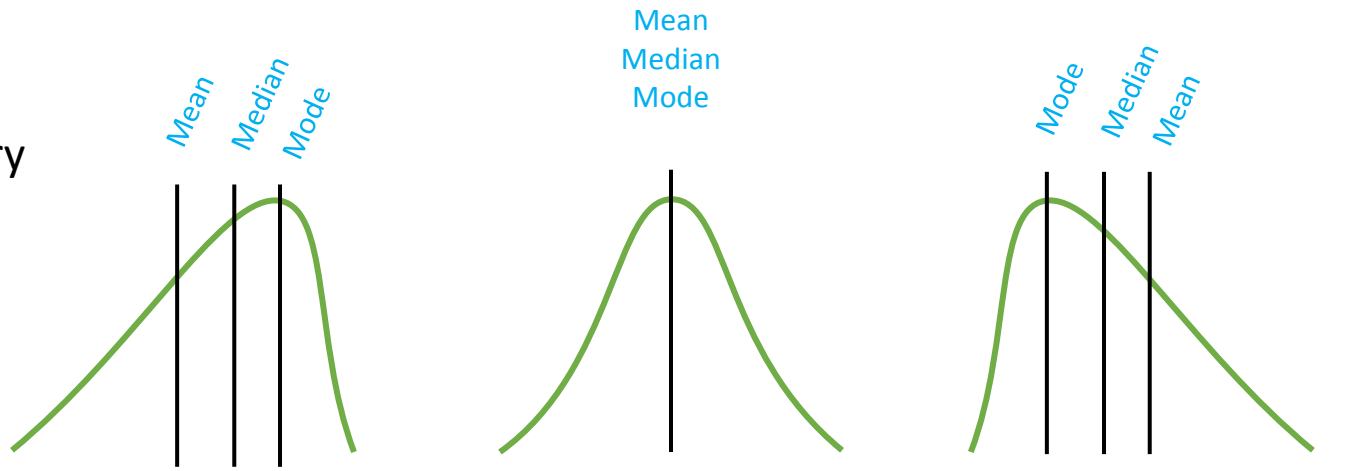
$\bar{x} = 25m$

$\mathrm{s} = 3\mathrm{m}$

$\mathrm{CV} = 0.12$ or 12%

Measures of shape

Symmetry



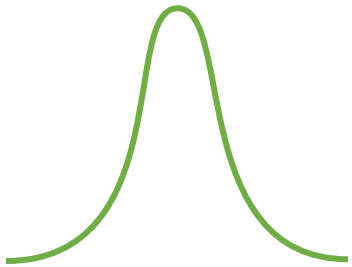
Left-skewed
Negatively skewed

Normal-curve
Perfectly symmetrical

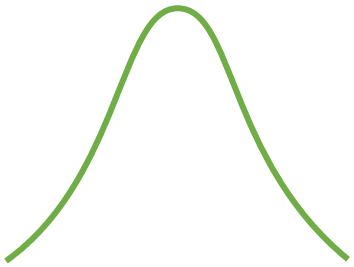
Right-skewed
Positively skewed

Measures of shape

Kurtosis



Leptokurtic
Tall and skinny, wide tails
Positive excess Kurtosis



Mesokurtic
Normal
Zero excess Kurtosis

Platykurtic
Flat – thin tails
Negative excess Kurtosis