Lecture 3 Numerical Measures

BIO210 Biostatistics

Xi Chen

Spring, 2023

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Summarise data using numbers

- Central tendency (mean, median, mode)
- Extremes (smallest, largest)
- Range
- Interquartile range (IQR)
- Dispersion (variance, standard deviation)

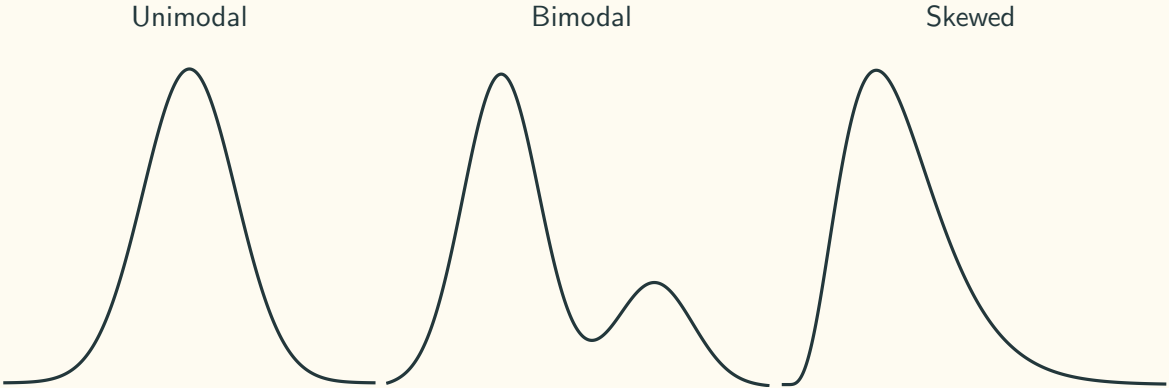
Central tendency

- Arithmetic mean: $\bar{x} = \dfrac{x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$
- Median: the 50th percentile
- Mode: the most frequent values

Central tendency

	Mean	Median	Mode
1, 2, 3, 4, 5, 6, 7, 8, 9	5	5	X
2, 3, 5, 7, 7, 7, 7, 11, 13	6.89	7	7
2, 3, 5, 7, 7, 7, 7, 11, 13, 100	16.2	7	7
2, 2, 2, 5, 6, 100, 100, 100, 103, 104	52.4	53	2, 100
2, 3, 4, 5, 6, 100, 101, 102, 103, 104	53	53	X

Central tendency

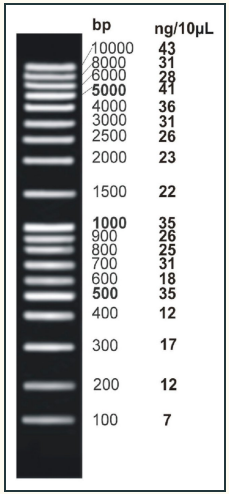


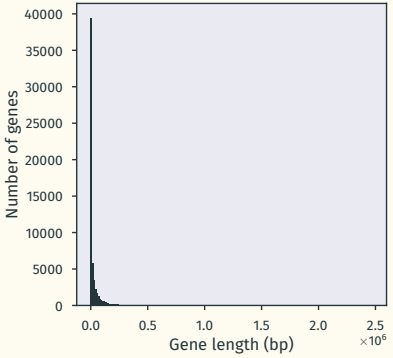
Central tendency

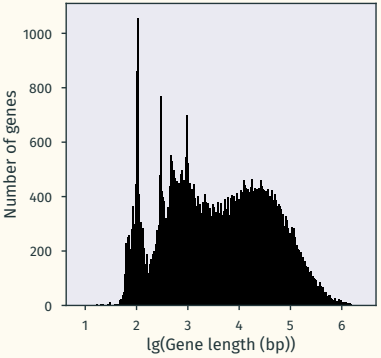
Geometric mean: $GM = \sqrt[n]{x_1 x_2 x_3 \cdots x_{n-1} x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$

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

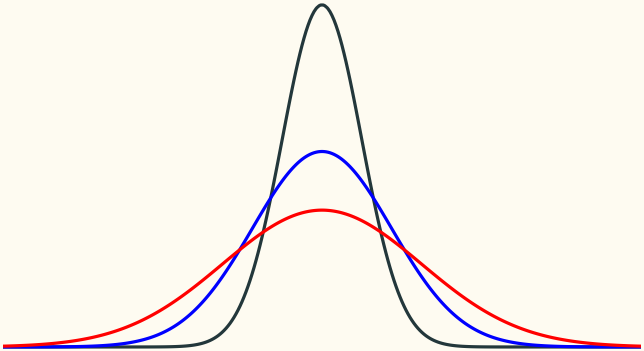
Log scale in biology







Dispersion



Dispersion

- Variance: $\mathbb{V}\mathrm{ar} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ or $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- Standard deviation: $\sqrt{\mathbb{V}\mathrm{ar}}$

Grouped data

Transcription factor binding motif length (bp):

[11, 13, 13, 11, 17, 11, 11, 13, 11, 13, 13, 11, 13, 13, 16, 11, 11, 11, 16, 11]

Mean:

$$\bar{x} = \frac{11 + 13 + 13 + 11 + 17 + 11 + 11 + 13 + 11 + 13 + 13 + 11 + 13 + 13 + 16 + 11 + 11 + 11 + 16 + 11}{20}$$

Grouped data

Motif length (bp)	Absolute frequency	Relative frequency
11	10	0.5
13	7	0.35
16	2	0.1
17	1	0.05
Total	20	1

Weighted average

Value	Absolute frequency
x_1	f_1
x_2	f_2
x_3	f_2
$\vdots$	$\vdots$
$\vdots$	$\vdots$
x_{n-1}	f_{n-1}
x_n	f_n

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_{n-1} f_{n-1} + x_n f_n}{f_1 + f_2 + \cdots + f_{n-1} + f_n}$$

$$= \frac{x_1 f_1 + x_2 f_2 + \cdots + x_{n-1} f_{n-1} + x_n f_n}{\sum_{i=1}^{n} f_i}$$

$$= \frac{f_1}{\sum_{i=1}^{n} f_i} x_1 + \frac{f_2}{\sum_{i=1}^{n} f_i} x_2 + \cdots + \frac{f_n}{\sum_{i=1}^{n} f_i} x_n$$

$$= w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \sum_{i=1}^{n} w_i x_i$$

Weighted average

Grading system:

Attendance	Assignments	Mid-term exam	Final exam
10%	15%	30%	45%

Pixel luminance:

$0.2126 \times \mathrm{R} + 0.7152 \times \mathrm{G} + 0.0722 \times \mathrm{B}$

$0.299 \times \mathrm{R} + 0.587 \times \mathrm{G} + 0.114 \times \mathrm{B}$

Weighted average

A badly manufactured dodecahedron die

Values	Relative frequency
1	8%
2	5%
3	5%
4	5%
5	5%
6	10%
7	10%
8	10%
9	10%
10	12%
11	10%
12	10%

We roll the die repeadly for a large number of times. What will be the average number?

Summary of descriptive statistics

- Types of data
  - Categorical
    - Nominal data
    - Ordinal data
  - Quantitative
    - Discrete data
    - Continuous data

- Data presentation
  - Tables (absolute/relative frequency table)
  - Graphs (Bar chart, pie chart, histogram, box plot, line graph, scatter plot)
  - Numerical measures
    - Central tendency (arithmetic/geometric mean, median, mode)
    - Dispersion (variance, standard deviation)
    - Others (range, IQR)

New way of calculating mean: weighted average $\bar{x} = \sum_{i=1}^{n} w_i x_i$