IEDA 2540 Statistics for Engineers
Descriptive Statistics

Wei YOU

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Spring, 2025

Descriptive Statistics

Descriptive statistics consist of methods for organizing and summarizing information about the sample.

- The construction of graphs, charts, and tables.
- The calculation of descriptive measures such as averages, measures of variation, and percentiles.

Samples and Variables

- Samples are made up of individuals.
- Each individual has many attributes/characteristics.
  Example: color, gender, price, durability, weight, etc.
- Since individuals differ in one or more attributes, these attributes are called variables.

Variable

A variable is an attribute of a research subject or participant that can take on different values.

Example: Variables to Consider When Buying a Second-Hand Bicycle

- **Brand:** (Giant, Specialized, Cannondale, etc.)
- **Model Number:** (entry-level, high-end, etc.)
- **Type:** (road racer, off-road, tourer, etc.)
- **Condition:** (Excellent, Acceptable, Poor, etc.)
- **Frame Size**
- **Number of Gears**
- **Price**

Types of variables: Qualitative - Nominal

Categorical/nominal variable: Discrete and unordered.

- Numbers are often used to represent the categories for ease of use.
- Neither order nor magnitude of the numbers matter.
- Arithmetic operations cannot be performed on such variables.

Example:

- Gender: F – 0, M – 1.
- Blood type: A – 0, B – 1, O – 2, AB – 3.
- Color: R – 1, G – 2, B – 3.

Types of Variables: Qualitative - Ordinal

Ordinal variable: Discrete but ordered.

- Numbers include the information about order.
- The magnitude of the numbers does not matter.
- Arithmetic operations cannot be performed with such variables.

Example:

- Stages of cancer: I, II, III, IV.
- Condition of second-handed goods: Excellent – 1, Acceptable – 2, Poor – 3.

Types of Variables: Quantitative - Discrete

Discrete variable: Discrete but ordered, where magnitude matters.

- Numbers represent measurable quantities.
- Both ordering and magnitude matter.
- Arithmetic operations can be performed with such variables.
- Discrete data are restricted to specified values that differ by fixed amounts (no intermediate values are possible; e.g., counting).

Example:

- Number of children in a family.
- Number of cars on a road.

Types of Variables: Quantitative - Continuous

Continuous variable: Continuous and ordered, where magnitude matters.

- Continuous data represent measurable quantities but are not restricted to taking on specific values (measuring).
- The difference between any two possible data values can be arbitrarily small.
- One limiting factor is the degree of accuracy with which measurements can be made.
- Even though we usually observe a finite set of values, the variable itself is continuous.

Example:

- Height, weight, humidity.

Example: Variables for Second-Handed Bicycle

- Brand of the bicycle (Nominal)
- Model number (Nominal or Ordinal)
- Type of the bicycle (Nominal)
- Condition (Ordinal)
- Size of the frame (Discrete)
- Number of gears (Discrete)
- Price (Continuous)

Describing Categorical Variables

Categorical variables (both nominal and ordinal) are discrete, and magnitude does not matter.

- Arithmetic operations are not applicable.
- Thus, we typically summarize them by **counting** the number of occurrences in each category.

The most straightforward way to describe catagorical data is by the **frequency**.

Frequency

The number of times a particular value occurs in a data set.

Suitable when there is a relatively small number of distinctive values in the variable.

Example: Second hand bicycle sales

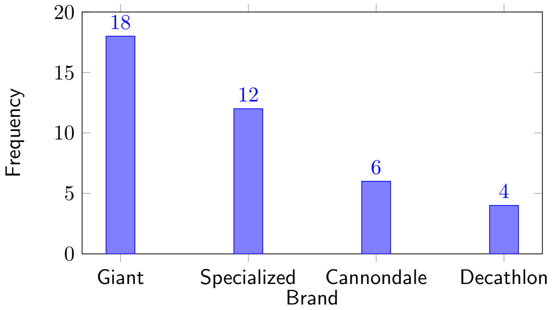
- Brand: Giant, Specialized, Cannondale, etc.
- Frequency table:

Brand	Frequency	Relative frequency
Giant	18	45%
Specialized	12	30%
Cannondale	6	15%
Decathlon	4	10%
Total	40	100%

Bar Chart

Bar Charts:

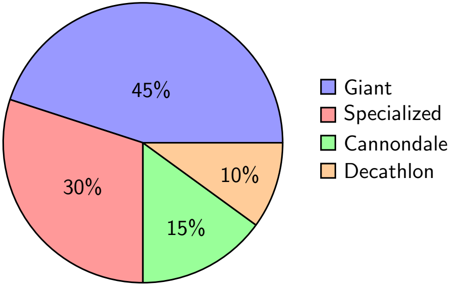
- The height of each block is proportional to the frequency.
- Best for comparing frequencies between different categories.



Pie Chart

Pie Charts:

- The angle of each slice is proportional to the frequency.
- Ideal for showing how one category compares to the whole.

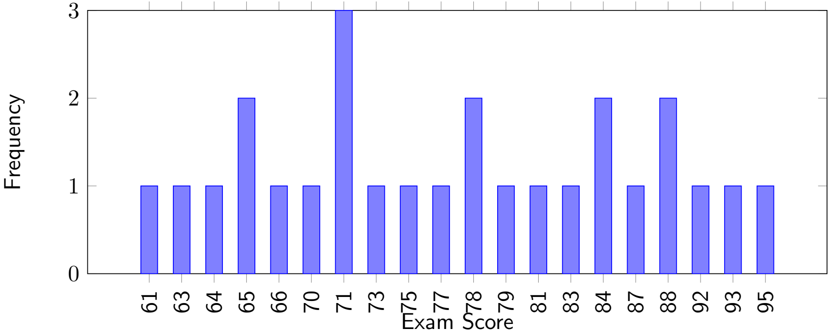


Describing Numeric Variables

Now we turn our focus to numeric variables (both discrete and continuous).

- **Key characteristics:** Ordered and arithmetic calculations are allowed.
- **Ordering** allows us to arrange tables and charts in a systematic manner.
- **Arithmetic calculations** enable deeper quantitative analysis (more on this later).

Bar Chart: Exam Scores



What problem do you see in this bar chart?

Limitations of Bar Charts for Many Distinct Values

- When a discrete or continuous variable has lots of distinct values and each value appears only a few times, the bar chart provides little information.
- Frequencies may be similar, making differences hard to discern.
- The distance between bars does not reflect the true numeric distance between observed values.
- Alternative visualizations may be more informative.

To address this issue, we may consider **grouping the data**.

Stem-and-Leaf Diagram for Ordinal Variables

A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set $x_1, x_2, ..., x_n$ where each number x_i consists of at least two digits.

- Divide the each number x_i into:
  - a **stem**: one or more leading digits; and
  - a **leaf**: the remaining digits.
- List the stems in a column, in ascending order.
- Record the leafs in the rows beside their stems, in ascending order.

Example: Exam scores, 61, 63, 64, 65, 65, 66, 70, 71, 71, 71, 73, 75, 77, 78, 78, 79, 81, 83, 84, 84, 87, 88, 88, 92, 93, 95.

Stem	Leaf
6	1 3 4 5 5 6
7	0 1 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

Stem-and-Leaf Diagram: Choosing the Number of Digits

The number of digits should be chosen such that:

- The "tree" is not too tall and skinny.
- The "tree" is not too short and fat.
- Both undesirable cases result in less information regarding the "shape of the distribution" of the data.

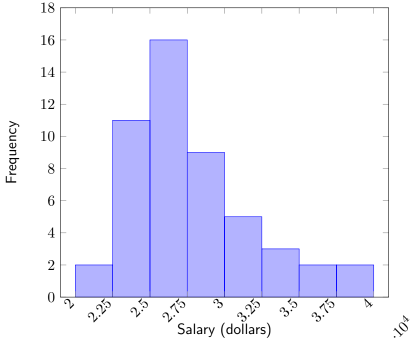
Limitation: However, it may not always be possible to construct a "normally-shaped" stem-and-leaf diagram, because grouping is done in an inflexible way by choosing a fixed number of digits to serve as the stem.

Histogram

Histogram is a more compact summary of data (compare with stem-and-leaf diagram).

- Divide the range of the data into intervals, which are called **bins**. Bins usually have equal length.
- Label the bin boundaries on x-axis.
- Mark and label the y-axis with the frequencies (or the relative frequencies).
- Above each bin, draw a rectangle where height is equal to the frequency (or the relative frequency) corresponding to that bin.

The per capita income of the states in the US.



Histogram – Intervals with Different Width

- Sometimes, it is necessary to draw histograms with uneven bins.
- For the previous dataset, we may want

Range	Frequency
$[20000, 25000]$	13
$[25000, 27500]$	16
$[27500, 30000]$	9
$[30000, 32500]$	5
$[32500, 35000]$	3
$[35000, 37500]$	2
$[37500, 40000]$	2

- What happens if we plot the frequency directly as the height of the histogram?
- Creates misperception that long intervals have large frequency.

Histogram – Intervals with Different Width

- The area rather than the height should be proportional to the frequency.
- Introduce the density scale: Draw the bar so that the height is

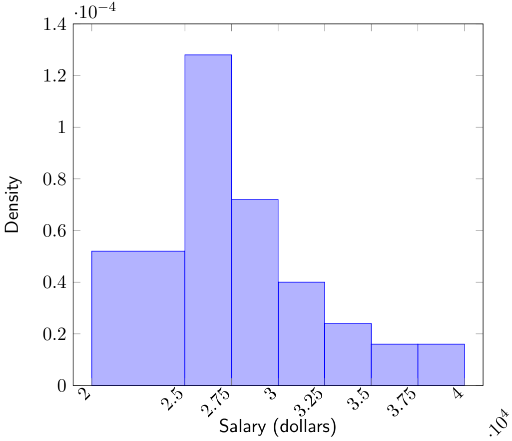
(relative frequency)/(width of the bin)

Range	Frequency	Density
$[20000, 25000]$	13	$5.2e{-}5$
$[25000, 27500]$	16	$1.3e{-}4$
$[27500, 30000]$	9	$7.2e{-}5$
$[30000, 32500]$	5	$4.0e{-}5$
$[32500, 35000]$	3	$2.4e{-}5$
$[35000, 37500]$	2	$1.6e{-}5$
$[37500, 40000]$	2	$1.6e{-}5$

- The total area is 100%.

Histogram – Density Scale

- The histogram in density scale is the sample distribution.
- If the sample size is large and representative of the population, then it is close to the population distribution.
- Roughly tell us the range and shape of the distribution.
- Tell us the mode $[25000, 27500]$.



Histogram – Number of Bins

Recall in the stem-and-leaf plot, we discussed the number of digits used in the stems.

For the number of bins in a histogram

- Too large and too small are both problematic.
- No single "best" formula.
- Usual choices includes $\sqrt{n}$ and $\log_2 n$, where n is the sample size.

Quantitative Summaries of Data

We saw plots that visualize the distribution of data (e.g., histograms).

Some plots require preliminary steps to summarize data:

- Counting frequencies.
- Grouping data.

Summarizing data using numeric values is very useful.

- Once a large dataset is collected, the first step is to understand its basic characteristics:
  - **Location:** Where is the center?
  - **Dispersion:** How spread out is the data?
  - **Shape:** What is the form of the distribution?

Describing Central Tendency

- **Central tendency** is the tendency of observations to "pile up" around a particular value.
- For *categorical variables*, the most appropriate measure of central tendency is the **sample mode**.

Sample mode

The sample mode is the most frequently occurring data value(s).

Brand	Frequency	Relative frequency	
Giant	18	45%	← "Sample mode"
Specialized	12	30%	
Cannondale	6	15%	
Decathlon	4	10%	
Total	40	100%	

Describing Central Tendency for Quantitative Data

- For categorical data, the mode is pretty much the best we can have.
- For quantitative variables (discrete or continuous), arithmetic calculations are possible!

Example: Noise levels measured at 36 different times directly outside Grand Central Station in Manhattan:

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85, 69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65.

- The mode probably won't tell us anything informative. The modes are 65,74,83,90,94.
- Instead, we have two choices, based on **ordering** and **arithmetic**, respectively.

Describing Central Tendency: The Median

When the data is arranged in increasing order, one obvious choice for measuring central tendency is to identify the value that "sits in the middle."

Sample median

Let n denote the number of observations and sort the data in increasing order:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

$$\text{Median} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{if } n \text{ is odd;} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & \text{if } n \text{ is even.} \end{cases}$$

Example: Consider the data set: 3, 5, 7, 8, 8, 9, 10, 12, 23, 35 (where $n = 10$). The median is the average of the 5$^{\text{th}}$ and 6$^{\text{th}}$ observations $\frac{8+9}{2} = 8.5$.

New Median with an Additional Observation

Example: Noisey Manhattan:

60, 65, 65, 68, 69, 72, 74, 74, 75, 77, 78, 82, 83, 83, 85, 87, 88, 89, 90, 90, 91, 94, 94, 95, 97, 100, 102, 107, 108, 110, 112, 114, 115, 122, 124, 125.

- **Original Data ($n = 36$):** Sorted order gives the 18th value = 89 and the 19th value = 90, so the median is 89.5.
- **Case 1: Additional observation = 91 ($n = 37$)** When 91 is added and the data are re-sorted, the median becomes the 19th observation. In this case, the 19th value is 90, so the new median is 90.
- **Case 2: Additional observation = 125 ($n = 37$)** When 125 is added (it is the largest value), the ordering of the lower values remains unchanged. The 19th observation is still 90, so the new median remains 90.

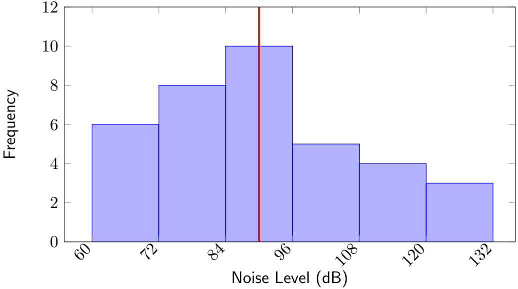
The median is robust to extreme values.

Data Compression and Loss of Information

We can easily identify the median from the stem-and-leaf diagram, as the previous example demonstrates.

Can you find the median if you are only given the histogram?

- **The answer is no!**
- A histogram abstracts out specific values to achieve a simpler description of the data.
- Although it maintains essential information, there is a loss of detailed data due to this compression.



There is a trade-off between information and simplicity.

Describing Central Tendency: Sample Mean

A far more commonly used central tendency measure is the **(arithmetic) mean**. In statistics, we usually call it the **sample mean**.

Sample mean

The sum of all the observations divided by the number of observations, n. Let $x_1, x_2, \ldots, x_n$ be a sample. The sample mean is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Mean vs. Median Sensitivity: Manhattan Noise Example

Example: Noisey Manhattan:

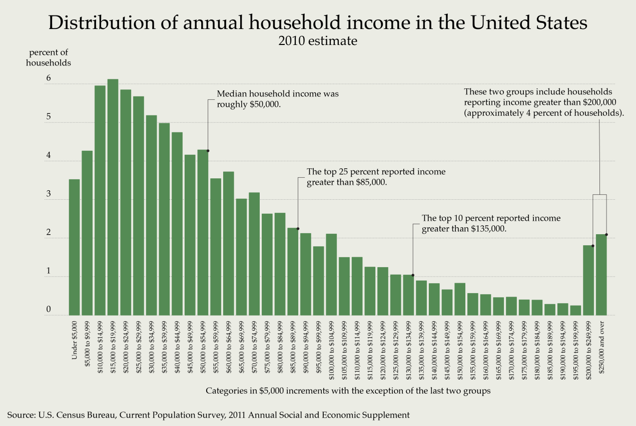
60, 65, 65, 68, 69, 72, 74, 74, 75, 77, 78, 82, 83, 83, 85, 87, 88, 89, 90, 90, 91, 94, 94, 95, 97, 100, 102, 107, 108, 110, 112, 114, 115, 122, 124, 125.

- **Original Measures:**
  - Sample mean = 90.67 dB.
  - Sample median = 90 dB.
- **Case 1: Additional sample of 91 dB**
  - New mean = 90.68 dB.
  - New median remains 90 dB.
- **Case 2: Additional sample of 125 dB**
  - New mean = 91.59 dB.
  - New median remains 90 dB.

The sample mean is more sensitive to changes in data than the median.

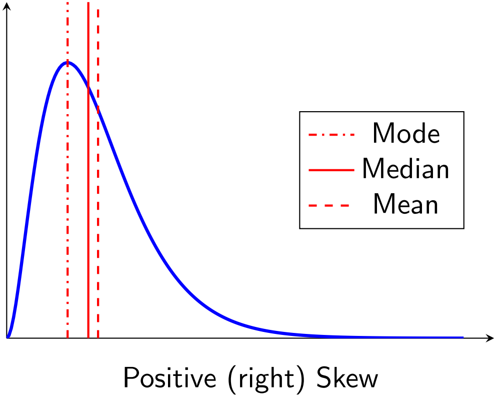
Example: Household Income: Mean vs. Median

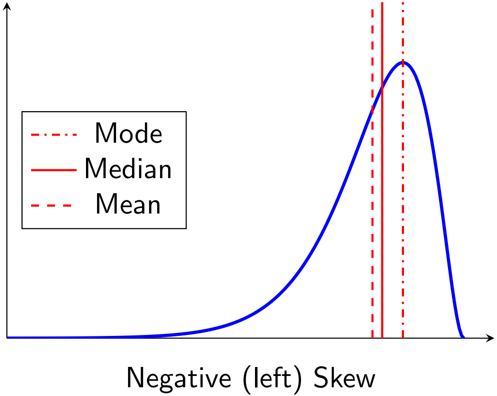
- In 2020, the average (mean) household income was US$97,026.
- The median household income was US$67,521.
- The top 1% of earners in the U.S. reported adjusted gross incomes over US$546,000 per year (2019) – more than seven times the median.
- **Observation:** The mean is sensitive to outliers.



Shape of Distribution: Symmetry / Skewness

We say that the histogram is skewed.





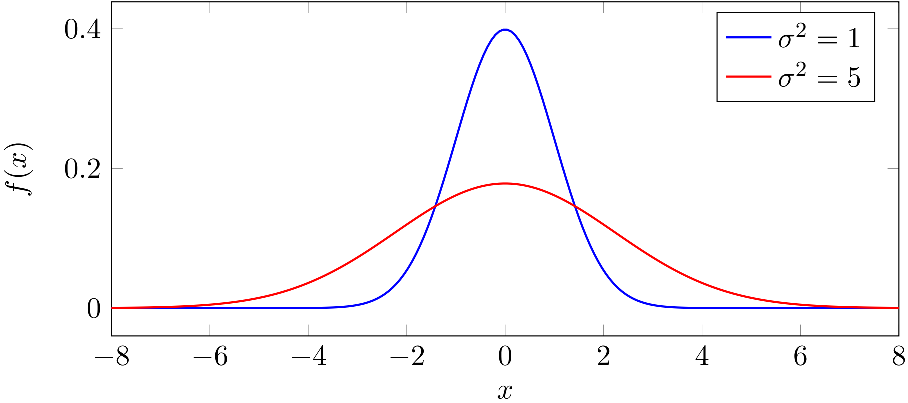
Describing the Center of the Data – Which to Use?

To describe the "center" of the data set

- For categorical data, use the mode.
- For ordered data, use the average.
- If the histogram is *highly skewed*, use the median.

Measures of Variation/Dispersion

In statistics, **dispersion** (also called **variability**, **scatter**, or **spread**) is the extent to which the "distribution" of the data is *stretched* or *squeezed*.



Range and Its Sensitivity to Outliers

Range

The **range** is the difference between the largest and smallest observations in a sample:

$$\text{Range} = \text{largestsample} - \text{smallestsample}.$$

Example: Consider the sample:

$$3, 5, 7, 8, 8, 9, 10, 12, 23, 35.$$

The range of the sample is: $35 - 3 = 32$.

- If an additional data point of 99 is included, the new range is $99 - 3 = 96 \gg 32$.

The range is highly sensitive to outliers!

Quartiles: A Robust Measure of Spread

The range is sensitive to outliers. A more robust alternative is to use **quartiles**.

Sample quartile

- The **first quartile** Q_1 is the value below which roughly 25% of the observations fall.
- The **second quartile** Q_2, also known as the median, is the value below which roughly 50% of the observations fall.
- The **third quartile** Q_3 is the value below which roughly 75% of the observations fall.
- Quartiles divide the data into *four roughly equal parts*.

Calculation of Quartiles

There are many ways to precisely define the quartiles.

Example: The default option in `R` and `Python`.

- Sort the n samples in *increasing order* $\Rightarrow \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ **ordered statistics**.
- The k-th quartile Q_k is at position

$$i_k = 1 + (n-1) \times \frac{k}{4}, \quad \text{for } k = 1, 2, 3.$$

- If i above positions are not integers, use *linear interpolation*:

$$Q_k = x_{(\lfloor i_k \rfloor)} + (i_k - \lfloor i_k \rfloor) \times (x_{(\lceil i_k \rceil)} - x_{(\lfloor i_k \rfloor)}), \quad \text{for } k = 1, 2, 3.$$

  - **Floor function**: $\lfloor x \rfloor$ is the largest integer that is smaller than or equal to x.
  - **Ceiling function**: $\lceil x \rceil$ is the smallest integer that is larger than or equal to x.

Calculation of Quartiles

Example: Noisy Manhattan:

60, 65, 65, 68, 69, 72, 74, 74, 75, 77, 78, 82, 83, 83, 85, 87, 88, 89, 90, 90, 91, 94, 94, 95, 97, 100, 102, 107, 108, 110, 112, 114, 115, 122, 124, 125.

First Quartile (Q_1):

$$i_1 = 1 + \frac{36-1}{4} = 9.75, \quad Q_1 = x_{(9)} + (9.75 - 9)\Big(x_{(10)} - x_{(9)}\Big) = 75 + 0.75 \times (77 - 75) = 76.5.$$

Second Quartile (Q_2, the median):

$$i_2 = 1 + \frac{36-1}{2} = 18.5, \quad Q_2 = x_{(18)} + (18.5 - 18)\Big(x_{(19)} - x_{(18)}\Big) = 89 + 0.5 \times (90 - 89) = 89.5.$$

Third Quartile (Q_3):

$$i_3 = 1 + \frac{3(36-1)}{4} = 27.25, \quad Q_3 = x_{(27)} + (27.25 - 27)\Big(x_{(28)} - x_{(27)}\Big) = 102 + 0.25 \times (107 - 102) = 103.25.$$

Measure of Dispersion: Interquartile Range

Interquartile range (IQR)

The interquartile range (IQR) is defined as:

$$\mathrm{IQR} = Q_3 - Q_1$$

It measures the range of the center 50% of the data.

- Quartiles are insensitive to outliers.

Example: Noisy Manhattan: Given:

$$Q_3 = 103.25 \quad \text{and} \quad Q_1 = 76.5,$$

the IQR is:

$$\mathrm{IQR} = 103.25 - 76.5 = 26.75.$$

Useful Data Features – Sample Percentile and Quantile

Sample percentile and quantile

The k-th **percentile** for $k = 1, \ldots, 99$ is at the position $1 + (n-1) \times k/100$.

The q **quantile**[a] for $0 < q < 1$ is at the position $1 + (n-1) \times q$.

[a] Not quartile!

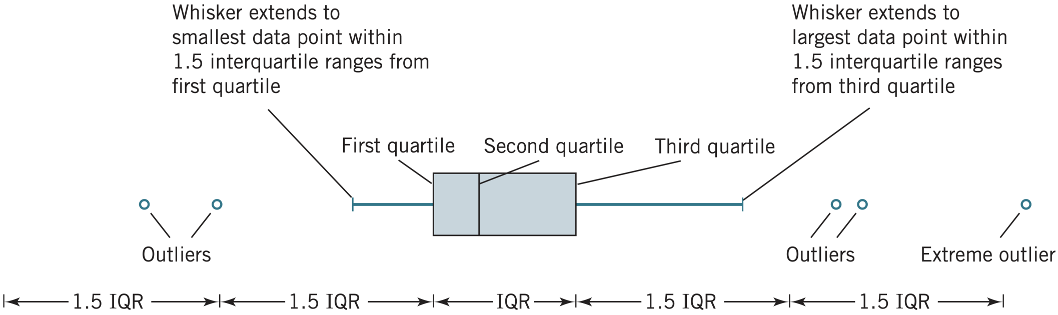
- the median is the 0.5 quantile, the first quartile is the 0.25 quantile, the 51-th percentile is the 0.51 quantile.

Five-number Summary and Boxplot

Five-number Summary

The five-number summary consists of minimum, maximum, and three quartiles.

Boxplot: A concise display of the range, center, skewness, variation and outliers.



Box Plot and Outliers

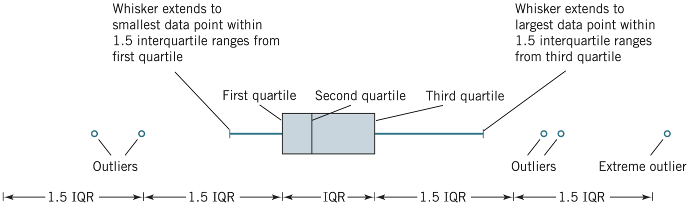
Outlier

- An **outlier** is an observation x such that either

$$x > Q_3 + 1.5 \times \text{IQR} \quad \text{or} \quad x < Q_1 - 1.5 \times \text{IQR}.$$

- An **extreme outlier** is an observation x such that either

$$x > Q_3 + 3 \times \text{IQR} \quad \text{or} \quad x < Q_1 - 3 \times \text{IQR}.$$



Measure of Dispersion: Sample Variance

Let $\{x_i,\, i = 1, 2, \ldots, n\}$ be a sample and $\bar{x}$ be the sample mean.

Sample variance and standard deviation

The **sample variance** is given by:

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

The **sample standard deviation** is:

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

Why divide by $n - 1$ instead of n? We will get into that later.

Sample Variance

The **sample standard deviation** is used to describe the variation of the data.

- Roughly speaking, S measures how a typical data point deviates from the average.
- The standard deviation of the income data is $\$4659$. In other words, the per capita income of a typical state is around $\$28054 \pm 4659$.
- If a data point is more than $2s$ away from the average, then it is considered quite extreme. (We should see very few of them. For normal random variable, less than 5%.)

Properties of Sample Variance and Standard Deviation

Let c, c_1, and c_2 be constants.

Translation Invariance

If $y_i = x_i + c$ for $i = 1, 2, \ldots, n$, then $s_y^2 = s_x^2$.

Scaling Property

If $y_i = c\,x_i$ for $i = 1, 2, \ldots, n$, then $s_y^2 = c^2 s_x^2$.

Combined Transformation

If $y_i = c_1\,x_i + c_2$ for $i = 1, 2, \ldots, n$, then $s_y^2 = c_1^2\,s_x^2$.

An Equivalent Formula for Sample Variance

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

Proof:

$$\begin{aligned} s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{2}{n-1}\bar{x}\sum_{i=1}^{n} x_i + \frac{n}{n-1}\bar{x}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{2n}{n-1}\bar{x}^2 + \frac{n}{n-1}\bar{x}^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2. \end{aligned}$$

Variance and Sample Variance

Sample variance, as the name suggests, is a sample version of variance.

Variance (Probability Perspective)

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Calculation of the variance requires knowledge of the distribution.

Sample Variance (Statistical Perspective)

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

Calculation of the sample variance requires only observations.

Measure of Dispersion: Coefficient of Variation (CV)

Definition

The coefficient of variation (CV) is defined by

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

where s is the sample standard deviation and $\bar{x}$ is the sample mean.

- **Dimensionless:** The CV remains the same regardless of the units used.
- **Scale invariant:** It is useful for comparing variability across data sets with different scales.

Summary

To describe the central tendency of the data, we use

- Mean.
- Median.
- Mode.
- Midrange.

To describe the variation of the data, we use

- Range.
- IQR.
- Variance.
- Standard deviation.
- Coefficient of variation.

Extended Reading and Exercises

- Chapter 6 of **Douglas C. Montgomery and George C. Runger, Applied Statistics and Probability for Engineers, 7th Ed.**