Statistics I – Chapter 3 Describing Data through Statistics

Ling-Chieh Kung

Department of Information Management
National Taiwan University

September 19, 2012

Describing data through statistics

- In Chapter 2, we introduced how to summarize data through graphs.
- In this chapter, we will discuss how to summarize data through **numbers**.
  - These "numbers" are called **statistics** for samples and parameters for **populations**.

Road map

- **Central tendency for ungrouped data**.
- Variability for ungrouped data.
- Grouped data.
- Measures of shape.

Central tendency for ungrouped data

- Measures of central tendency yields information about the **center** or middle part of a group of numbers.
  - Where the center is ("center" must be defined)?
  - Where the middle part is ("middle part" must be defined)?
- They provide **summaries** to data.
  - Analogy: The determinant and eigenvalues are "summaries" of a matrix.

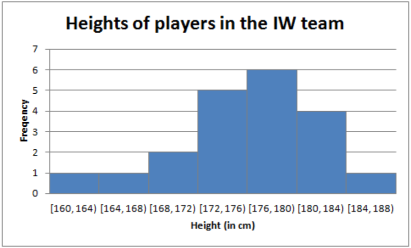
Central tendency for ungrouped data

- We will discuss five measures of central tendency:
  - Modes.
  - Medians.
  - Means.
  - Percentiles.
  - Quartiles.
- We first focus on **ungrouped data**. They are raw data without any categorization.

Central tendency for ungrouped data

- In the IW baseball team, players' heights (in cm) are:

178	172	175	184	172	175	165	178	177	175
180	182	177	183	180	178	179	162	170	171



- Let's try to describe the central tendency of this data.

Modes

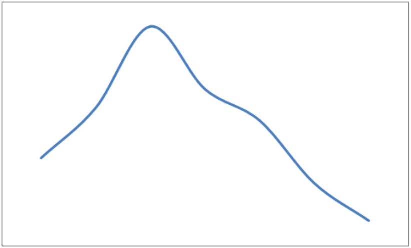
- The **mode**(s) is (are) the **most frequently** occurring value(s) in a set of data.
  - In the team, the modes are 175 and 178. See the sorted data:

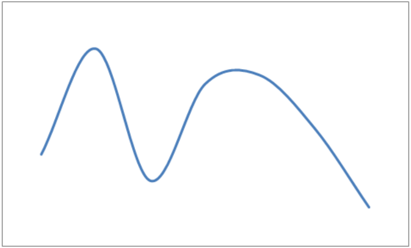
162	165	170	171	172	172	175	175	175	177
177	178	178	178	179	180	180	182	183	184

  - We thus know that most people are of 175 and 178 cm.

The number of modes

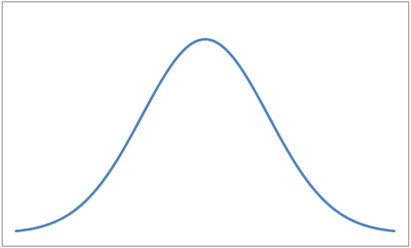
- The data of the IM team is **bimodal**.
- In general, data may be **unimodal**, bimodal, or **multimodal**.
  - When the mode is unique, the data is unimodal.
  - When there are two modes or two values of similar frequencies that are more dominant than others, the data is bimodal.





Bell shaped curve

- A particularly important type of unimodal curves is the **bell shaped curves**.



- Normal distributions, which will be defined in Chapter 5, is bell shaped.

Medians

- The **median** is the **middle** value in an ordered set of numbers.
  - For the median, **at least half** of the numbers are weakly below and **at least half** are weakly above it.[1]
- To find the median, suppose there are N numbers:
  - If N is odd, the median is the $\frac{N+1}{2}$th large number.
  - If N is even, the median is the **average** of the $\frac{N}{2}$th and the $(\frac{N}{2}+1)$th large number.

[1] "Weekly below (above)" means "no greater (less) than".

Medians

- In the IW team, the median is $\frac{177+177}{2} = 177$ cm.

162	165	170	171	172	172	175	175	175	177
177	178	178	178	179	180	180	182	183	184

- For the following team, the median is $\frac{175+177}{2} = 176$ cm.

162	165	170	171	172	172	175	175	175	**175**
177	178	178	178	179	180	180	182	183	184

- For the following team, the median is 177 cm.

162	165	170	171	172	172	175	175	175	175	
177	178	178	178	179	180	180	182	183	184	**188**

Medians

- A median is unaffected by the magnitude of extreme values:
  - For the following team, the median is still 177 cm.

162	165	170	171	172	172	175	175	175	175	
177	178	178	178	179	180	180	182	183	184	**238**

- Unfortunately, a median does not use all the information contained in the numbers.
  - While data may be of interval or ratio scales, a median only treat the data as ordinal.

Means

- The (arithmetic) **mean** is the **arithmetic average** of a group of data.
  - For the IW team, the mean is

$$\frac{162 + 165 + 170 + \cdots + 183 + 184}{20} = 175.65 \text{ cm.}$$

- In Statistics, means are the most commonly used measure of central tendency.
- Do people consider geometric means in Statistics?

Population means v.s. sample means

- Let $\{x_i\}_{i=1,\ldots,N}$ be a population with N as the **population size**. The **population mean** is

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

- Let $\{x_i\}_{i=1,\ldots,n}$ be a sample with $n < N$ as the **sample size**. The **sample mean** is

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

- Throughout this year (and the whole Statistics world), we use the above notations.

Population means v.s. sample means

- Isn't these two means the same?
  - From the perspective of calculation, yes.
  - From the perspective of statistical inference, **no**.
- In practice, typically the population mean of a population is **unknown**.
  - We use **inferential Statistics** to estimate or test for the population mean.
  - To do so, we start from the sample mean.

Some remarks for means

- Do not try to find the mean for ordinal or nominal data.
- A mean uses all the information contained in the numbers.
- Unfortunately, a mean will be affected by extreme values.
  - Therefore, using the mean and median **simultaneously** can be a good idea.
  - We should try to identify **outliers** (extreme values that seem to be "strange") before calculating a mean (or any statistics).
  - Any outlier here?

16	165	170	171	172	172	175	175	175	177
177	178	178	178	179	180	180	182	183	184

Quartiles

- The range of a set of data is determined by the two extreme values. It says nothing about the other numbers.
  - For uniformly distributed data, the range is representative.
  - For other types of distribution, especially bell shaped distributions, the range ignores most of the data.
- Sometimes we want to know the range of the middle 50% values. This motivates us to define **quartiles**.
- For the qth quartile,
  - **at least** $\frac{q}{4}$ of the values are weakly below it and
  - **at least** $1 - \frac{q}{4}$ of the values are weakly above it.

Quartiles

- To calculate the qth quartile, $q = 1, 2, 3$, first calculate $i = \frac{q}{4}N$. Then we have the qth quartile as

$$Q_i \equiv \begin{cases} \dfrac{x_i + x_{i+1}}{2} & \text{if } i \in \mathbb{N} \\ x_i & \text{otherwise} \end{cases} .$$

- Find the quartiles for the IW team:

162	165	170	171	172	172	175	175	175	177
177	178	178	178	179	180	180	182	183	184

  - How many numbers are below the qth quartile?
  - What is the proportion of numbers below the qth quartile?

Some remarks for quartiles

- The **interquartile range** (IQR), is defined as the difference between the first and third quartiles.
  - What is the proportion of numbers in the interquartile range?
- What is the **second quartile**?
- The textbook says that, for the qth quartile, **at most** $1 - \frac{q}{4}$ of the values are weakly above it. What do you think?

Percentiles

- The idea of quartiles can be generalized to **percentiles**.
- For the Pth percentile,
  - at least $\frac{P}{100}$ of the values are weakly below it and
  - at least $1 - \frac{P}{100}$ of the values are weakly above it.
- In theory, P can be any **real** number between 0 and 100.
- In practice, typically only **integer** values of P are of interest.

Percentiles

- To calculate the Pth percentile, $P \in [0, 100]$, first calculate $i = \frac{P}{100}N$. Then we have the Pth percentile as

$$P_i \equiv \begin{cases} \dfrac{x_i + x_{i+1}}{2} & \text{if } i \in \mathbb{N} \\ x_i & \text{otherwise} \end{cases} .$$

  - The 25th percentile is the first quartile.
  - The 50th percentile is the median.
  - The 75th percentile is the third quartile.

Some final remarks

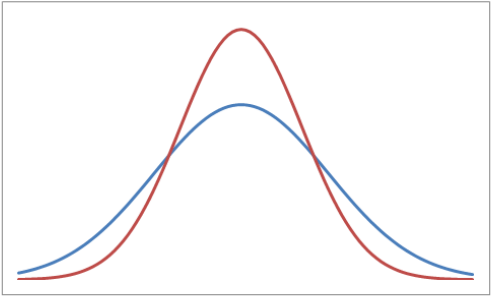
- Five measures of central tendency for ungrouped data: modes, medians, means, quartiles, percentiles.
- Each measure provide a certain summary of the data.
- To better describe a set of data, combine some of these measures.

Road map

- Central tendency for ungrouped data.
- **Variability for ungrouped data**.
- Grouped data.
- Measures of shape.

Variability for ungrouped data

- **Measures of variability** describe the **spread** or **dispersion** of a set of data.
- Especially useful when two sets of data have the same center.



Variability for ungrouped data

- We will discuss seven measures of central tendency:
  - Ranges.
  - Interquartile ranges.
  - Mean absolute deviations.
  - Variances.
  - Standard deviations.
  - z scores.
  - Coefficients of variation.
- We first focus on ungrouped data. They are raw data without any categorization.

Ranges and Interquartile ranges

- The **range** of a set of data $\{x_i\}_{i=1,\ldots,N}$ is

$$\max_{i=1,\ldots,N}\{x_i\} - \min_{i=1,\ldots,N}\{x_i\}.$$

  - In applications that require strict "guarantees," such as quality control, the range is important.
- The **interquartile range** of a set of data is the difference of the first and third quartile.
  - It is the range of the **middle 50%** of data.

Deviations from the mean

- Consider a set of population data $\{x_i\}_{i=1,\ldots,N}$ with mean μ.
- Intuitively, a way to measure the dispersion is to examine how each number **deviates from the mean**.
- For x_i, the deviation from the mean is defined as

$$x_i - \mu.$$

- For a **sample**, the deviations from the mean are defined based on the sample mean $\bar{x}$.

Deviations from the mean

- How to combine the N deviations into **a single number**?
- Intuitively, we may sum them up:

$$\sum_{i=1}^{N}(x_i - \mu).$$

- What will happen?
- How would you design a way to combine these deviations?

Deviations from the mean

- Instead of summing them up, we have the following two alternative options:
  - Summing up the absolute values of the deviations:

$$\sum_{i=1}^{N} |x_i - \mu|.$$

  - Summing up the squares of the deviations.

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

Mean absolute deviations

- The mean absolute deviation (MAD) of a population $\{x_i\}_{i=1,\ldots,N}$ is the average of the **absolute values** of the deviations from the mean:

$$\frac{\sum_{i=1}^{N} |x_i - \mu|}{N}.$$

  - It is always **nonnegative**. As long as any two numbers are different, it is **positive**.
  - The larger the MAD is, the more dispersed the data is.

Mean absolute deviations

- In the WI baseball team, there are with only six players. In one game, their scores are 3, 5, 6, 10, 12, and 18 points.
- Find the MAD of the population:
  - First, find the population size:

$$N = 6.$$

  - Second, find the population mean:

$$\mu = \frac{\sum_{i=1}^{6} x_i}{6} = 9.$$

Mean absolute deviations

- Third, find the sum of absolute deviations by constructing the following table:

	x_i	$x_i - \mu$	$\lvert x_i - \mu \rvert$
	3	-6	6
	5	-4	4
	6	-3	3
	10	1	1
	12	3	3
	18	9	9
Total	54	0	26

- Finally, the mean absolute deviation is $\frac{26}{6} = \frac{13}{3} \approx 4.33$.

Some remarks for MAD

- Mean absolute deviations are intuitive, easy to calculate, and reasonably representative.
- Unfortunately, as the absolute function is NOT **differentiable**, it is hard to derive rigorous statistical properties for it.
- Mean absolute deviations are thus less useful in Statistics.
  - In this semester, you will not see them again ...
  - In some applications, such as forecasting, mean absolute deviations are still adopted.

Variances

- The variance of a population $\{x_i\}_{i=1,\ldots,N}$ is the average of the **squared values** of the deviations from the mean:

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

  - It is always **nonnegative**. As long as any two numbers are different, it is **positive**.
  - A larger variance implies a more dispersed set of data.
  - It emphasizes on huge deviations.
  - It is **differentiable**.

Variances

- Find the variance of the WI team players' scores $\{3, 5, 6, 10, 12, 18\}$. Note that this is a **population**.
- Again, we construct the following table:

	x_i	$x_i - \mu$	$(x_i - \mu)^2$
	3	-6	36
	5	-4	16
	6	-3	9
	10	1	1
	12	3	9
	18	9	81
Total	54	0	152

- The population variance is thus $\sigma^2 = \frac{152}{6} = \frac{76}{3} \approx 25.33$.

Some remarks for variances

- The population variance 25.33 is much larger than the mean absolute deviation 4.33. Is this always true?
- While the mean absolute deviation is 4.33 **points**, the population variance is 25.33 **squared points**.
- The main disadvantage of using variances is that the unit of measurement is the square of the original one.

Population v.s. sample variances

- The symbol σ^2 is always used as the **population variance**.
- For a sample $\{x_i\}_{i=1,...,n}$, the **sample variance** is defined as

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

  Notice that $n-1$!!
- You probably want to ask something ...

Standard deviations

- To fix the problem of having a squared unit of measurement when using variances, we define **standard deviations**.
- For either a population or a sample, the standard deviation is the **square root** of the variance:

$$\sigma \equiv \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \quad \text{and} \quad s \equiv \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

- Standard deviations have the same unit of measurement as the raw data.

Standard deviations

- As we will see, standard deviations play a very important role in statistical inference.
- Before that, let's study two interesting rules regarding standard deviations.

Chebyshev's theorem

- **Chebyshev's theorem** provides a lower bound on the proportion of data that are "close to" the mean:

Proposition 1 (Chebyshev's theorem)

For any set of data with mean μ and standard deviation σ, if $k \geq 1$, at least

$$1 - \frac{1}{k^2}$$

proportion of the values are within $[\mu - k\sigma, \mu + k\sigma]$.

- So 75% of data are within 2σ, 89% are within 3σ, etc.
- The power of Chebyshev's theorem is that it applies to **any** set of data.

Chebyshev's theorem

- Let's verify Chebyshev's theorem by investigating the WI team players' scores $\{3, 5, 6, 10, 12, 18\}$.
  - $\mu = 9$ and $\sigma \approx 5.03$.
  - For $k = 2$: $[-1.06, 19.06]$ contains $100\% > 1 - \frac{1}{2^2} = 75\%$.
  - For $k = 1.5$: $[1.46, 16.55]$ contains $83.3\% > 1 - \frac{1}{(1.5)^2} = 55.6\%$.
- We will prove this theorem when studying Chapter 6.
- As Chebyshev's theorem applies to any set of data, the bounds it provide are typically **loose** for most data.
- The next theorem does better for bell shaped data.

The empirical rule

- The **empirical rule** estimates the approximate proportion of values that are "close to" the mean:

Observation 1 (The empirical rule)

For a bell shaped set of data, approximately 68%, 95%, and 99.7% of the values are within 1σ, 2σ, *and* 3σ *from* μ.

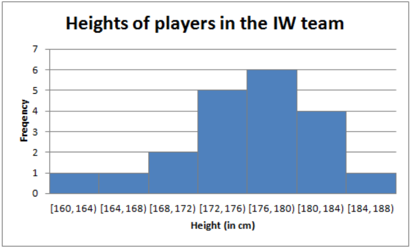
- For the scores $\{3, 5, 6, 10, 12, 18\}$:
  - $\mu = 9$ and $\sigma \approx 5.03$.
  - For 1σ: $[3.97, 14.03]$ contains $66.7\% \approx 68\%$.
  - For 2σ: $[-1.06, 19.06]$ contains $100\% \approx 95\%$.

Chebyshev's theorem v.s. empirical rule

- Recall that the IW team players' heights

162	165	170	171	172	172	175	175	175	177
177	178	178	178	179	180	180	182	183	184

are approximately bell shaped:



Chebyshev's theorem v.s. empirical rule

- Let's apply the two rules on the IW team players' heights:
  - $\mu = 175.65$ and $\sigma = 5.54$.
  - The result:

	1σ	2σ	3σ
Chebyshev's theorem	0%	75%	88.9%
Empirical rule	68%	95%	99.7%
Real proportion	70%	95%	100%

Some remarks for the empirical rule

- It is a **rule of thumb**! All we have are approximations.
- The approximation is precise for normally distributed data.
- The approximation is good only for bell shaped data.
- What kind of data may make the approximation bad?

Standard scores

- For a number x_i, we define its **z score** (standard scores or z value) as
$$z = \frac{x_i - \mu}{\sigma}.$$

- A z score represents the **number of standard deviations** that the value deviates from the mean.
- z scores are particularly important for normal distributions. This will be discussed extensively later in the semester.

Coefficient of variation

- The coefficient of variation is the **ratio** of the standard deviation to the mean:

$$\text{Coefficient of variation} = \frac{\sigma}{\mu}.$$

- Why do we want to use coefficients of variation? Is using standard deviation not enough?
- When will you use coefficients of variation? Is it when you have one or multiple sets of data?

Road map

- Central tendency for ungrouped data.
- Variability for ungrouped data.
- **Grouped data**.
- Measures of shape.

Grouped data

- A set of **grouped data** contains values that are divided into **several classes**.
  - One example is frequency distributions.
  - When you survey people's income ...
- We now introduce how to calculate the mean, median, mode, variance, and standard deviation for a set of grouped data.

Means for grouped data

- In calculating the mean for a set of grouped data, the **class midpoint** are used to represent all the values in that class.
- For the IW team, suppose we only have the frequency table:

Class	Frequency
$[160, 164)$	1
$[164, 168)$	1
$[168, 172)$	2
$[172, 176)$	5
$[176, 180)$	6
$[180, 184)$	4
$[184, 188)$	1

Means for grouped data

- The mean of this set of grouped data is calculated as follows:

Class	Frequency (f_i)	Class midpoint (M_i)	f_iM_i
$[160, 164)$	1	162	162
$[164, 168)$	1	166	166
$[168, 172)$	2	170	340
$[172, 176)$	5	174	870
$[176, 180)$	6	178	1068
$[180, 184)$	4	182	728
$[184, 188)$	1	186	186
Total	20		3520

Then the mean is $\frac{3520}{20} = 176$ cm.

Means for grouped data

- For a set of grouped data with k classes, let M_i be the midpoint and f_i be the frequency of class i. The mean of this set of data is

$$\mu_{\text{grouped}} = \frac{\sum_{i=1}^{k} f_i M_i}{\sum_{i=1}^{k} f_i}.$$

- The mean for grouped data is just an approximation.
- It is hard to do better if we do not know more about the distribution of the data.

Variances for grouped data

- For variances, we still use the **class midpoint** to represent all numbers in each class.
- For a set of grouped data with mean μ and k classes, let M_i be the midpoint and f_i be the frequency of class i. The variance of this set of data is

$$\sigma^2_{\text{grouped}} = \frac{\sum_{i=1}^{k} f_i (M_i - \mu)^2}{\sum_{i=1}^{k} f_i}.$$

- Verify by yourself that the variance of the IW team's grouped heights is 32.8 cm^2.

Standard deviations for grouped data

- For standard deviations, we still use the **class midpoint** to represent all numbers in each class.
- For a set of grouped data with mean μ and k classes, let M_i be the midpoint and f_i be the frequency of class i. The variance of this set of data is

$$\sigma_{\text{grouped}} = \sqrt{\frac{\sum_{i=1}^{k} f_i (M_i - \mu)^2}{\sum_{i=1}^{k} f_i}}.$$

- Verify by yourself that the standard deviation of the IW team's grouped heights is 5.73 cm.

For samples

- When the grouped data form a sample, change the denominator from $\sum_{i=1}^{k} f_i$ to $\sum_{i=1}^{k} f_i - 1$.

Modes for grouped data

- The mode for grouped data is the **class midpoint** of the modal class.
- Verify by yourself that the mode of the IW team's grouped heights is 178 cm.

Medians for grouped data

- Calculating medians for grouped data does NOT use class midpoints!
- It involves the following steps:
  - Given the size N, find the **median class**: the class in which the $\frac{N}{2}$th term locates.
  - Determine the **position in the class** of the $\frac{N}{2}$th term.
  - Do an **interpolation** within the median class based on the position and the frequency of the class.

Medians for grouped data

Class	Frequency
$[160, 164)$	1
$[164, 168)$	1
$[168, 172)$	2
$[172, 176)$	5
$[176, 180)$	6
$[180, 184)$	4
$[184, 188)$	1

- $\frac{N}{2} = 10$.
- The tenth term locates in the class $[176, 180)$. It is the first term of the median class.
- As the class starts from 176, ends at 180, and has six terms, the interpolation puts the first term at

$$176 + \frac{1}{6}(180 - 176) \approx 176.67.$$

- So the median is 176.67 cm.

Road map

- Central tendency for ungrouped data.
- Variability for ungrouped data.
- Grouped data.
- **Measures of shape**.

Skewness

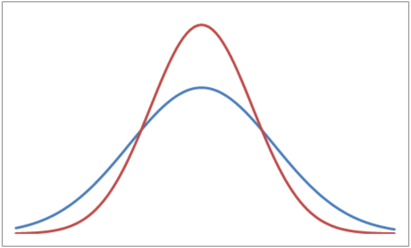
- In describing the distribution of a set of data, the **shape** is also important.
- There are two common statistical descriptions on the shape of a set of data:
  - Skewness.
  - Kurtosis.

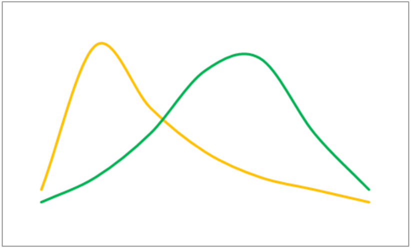
Skewness

- A distribution is **symmetric** if its right half is the **mirror** image of its left half.
- A distribution is **skewed** (asymmetric) if it is not symmetric.
- There are two types of skewness, depending on where the tail goes:
  - **Positively skewed** or **skewed to the right**.
  - **Negatively skewed** or **skewed to the left**.

Skewness

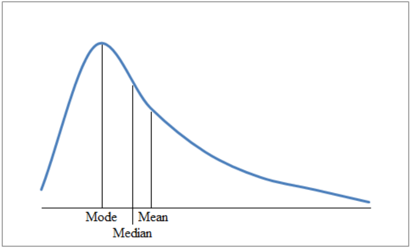
- Which curve is symmetric?
- Which is skewed to the left?
- Which is skewed to the right?

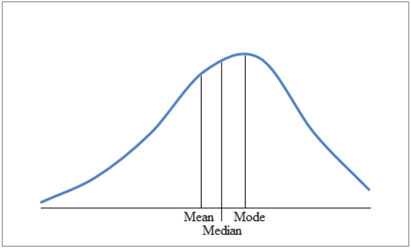




Skewness

- If a distribution is unimodal, the relationship among the mean, median, and mode gives hints to the skewness.
  - Symmetric: mean = median = mode.
  - Skewed to the left: mean < median < mode.
  - Skewed to the right: mean > median > mode.





Coefficients of skewness

- Many different coefficients of skewness have been defined.
- A coefficient of skewness is a function of the data values such that the function is:
  - symmetric if the coefficient is 0,
  - skewed to the right if the coefficient is positive, and
  - skewed to the left if the coefficient is negative.
- No one says which coefficient of skewness dominates all others.

Kurtosis

- **Kurtosis** describes the **degree of peakedness** of a distribution.
- Many different **coefficients of kurtosis** have been defined.
- No one says which coefficient of kurtosis dominates all others.