Statistics and Data Analysis

Descriptive Statistics (2): Summarization

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Summarizing the data with numbers

- Descriptive Statistics includes some common ways to describe data.
  - **Visualization** with graphs.
  - **Summarization** with numbers.
- This is always the **first step** of any data analysis project: To get intuitions that guide our directions.
- Today we talk about summarization.
  - For a set of (a lot of) numbers, we use a few numbers to summarize them.
  - For a population: these numbers are **parameters**.
  - For a sample: these numbers are **statistics**.
- We will talk about three things:
  - Measures of **central tendency** for the center or middle part of data.
  - Measures of **variability** for how variable the data are.
  - Measures of **correlation** for the relationship between two variables.

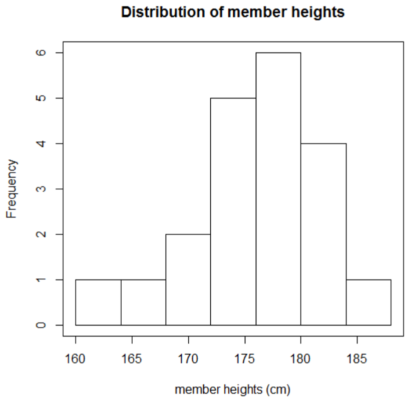
Road map

- **Describing central tendency**.
- Describing variability.
- Describing correlation.

Central tendency

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

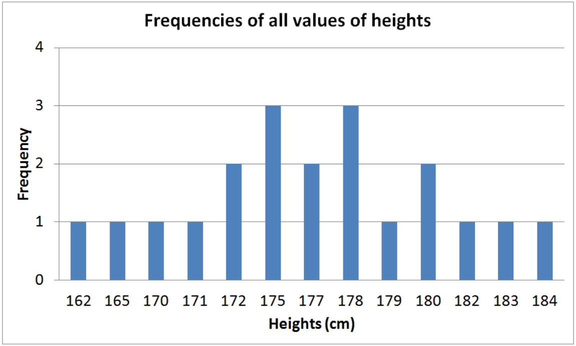
178	172	175	184
172	175	165	178
177	175	180	182
177	183	180	178
179	162	170	171



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



- It is better to look for a mode in a set of **qualitative** data.
  - Otherwise, maybe all values are modes!

Medians

- The **median** is the **middle** value in an ordered set of numbers.
  - Roughly speaking, **half** of the numbers are below and **half** are above it.
- 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.
- For example:
  - The median of $\{1, 2, 4, 5, 6, 8, 9\}$ is 5.
  - The median of $\{1, 2, 4, 5, 6, 8\}$ is $\frac{4+5}{2} = 4.5$.

Medians

- A median is unaffected by the magnitude of extreme values:
  - The median of $\{1, 2, 4, 5, 6, 8, 9\}$ is 5.
  - The median of $\{1, 2, 4, 5, 6, 8, 900\}$ is still 5.
- Medians may be calculated from **quantitative** or **ordinal** data.
  - It cannot be calculated from nominal data.
- Unfortunately, a median uses only **part** of the information contained in these numbers.
  - For quantitative data, a median only treats them as ordinal.

Means

- The **mean** is the **average** of a set of data.
  - Can be calculated only from quantitative data.
  - The mean of $\{1, 2, 4, 5, 6, 8, 9\}$ is

$$\frac{1+2+4+5+6+8+9}{7} = 5.$$

- A mean uses **all** the information contained in the numbers.
- Unfortunately, a mean will be affected by extreme values.
  - The mean of $\{1, 2, 4, 5, 6, 8, 900\}$ is $\frac{1+2+4+5+6+8+900}{7} \approx 132.28$!
  - 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).

Population means vs. sample means

- Let $\{x_i\}_{i=1,...,N}$ be a population with N as the **population size**. The **population mean** is
$$\mu \equiv \frac{\sum_{i=1}^{N} x_i}{N}.$$
- Let $\{x_i\}_{i=1,...,n}$ be a sample with $n < N$ as the **sample size**. The **sample mean** is
$$\bar{x} \equiv \frac{\sum_{i=1}^{n} x_i}{n}.$$
- People use μ and $\bar{x}$ in almost the whole statistics world.

Population means v.s. sample means

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

- Isn't these two means the same?
  - From the perspective of calculation, yes.
  - From the perspective of statistical inference, **no**.
- Typically the population mean is **fixed but unknown**.
  - The sample mean is **random**: We may get different values of $\bar{x}$ today and tomorrow.
  - To start from $\bar{x}$ and use **inferential statistics** to estimate or test μ, we need to apply **probability**.

Quartiles and percentiles

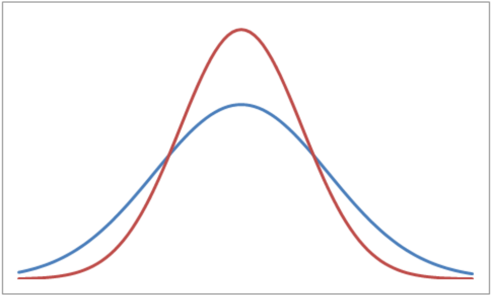
- The median lies at the middle of the data.
- The **first quartile** lies at the middle of the **first half** of the data.
- The **third quartile** lies at the middle of the **second half** of the data.
- For the pth **percentile**:
  - $\frac{p}{100}$ of the values are below it.
  - $1 - \frac{p}{100}$ of the values are above it.
- Median, quartiles, and percentiles:
  - The 25th percentile is the first quartile.
  - The 50th percentile is the median (and the second quartile).
  - The 75th percentile is the third quartile.

Road map

- Describing central tendency.
- **Describing variability**.
- Describing correlation.

Variability

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



Ranges and Interquartile ranges

- The **range** of a set of data $\{x_i\}_{i=1,...,N}$ is the difference between the maximum and minimum numbers, i.e.,

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

- 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.
  - It excludes the effects of extreme values.

Deviations from the mean

- Consider a set of population data $\{x_i\}_{i=1,...,N}$ 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 population mean is defined as

$$x_i - \mu.$$

- For a **sample**, the deviation from the sample mean of x_i is

$$x_i - \bar{x}.$$

i	x_i	deviation
1	1	$1-5=-4$
2	2	$2-5=-3$
3	4	$4-5=-1$
4	5	$1-5=0$
5	6	$6-5=1$
6	8	$8-5=3$
7	9	$9-5=4$
Mean	5	

Mean deviations

- May we summarize the N deviations into a single number to summarize the aggregate deviation?
- Intuitively, we may sum them up and then calculate the **mean deviation**:

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

- Is it always 0?

i	x_i	deviation
1	1	$1-5=-4$
2	2	$2-5=-3$
3	4	$4-5=-1$
4	5	$1-5=0$
5	6	$6-5=1$
6	8	$8-5=3$
7	9	$9-5=4$
Mean	5	0

Adjusting mean deviations

- People use two ways to adjust it:
  - Mean **absolute** deviations (MAD):
    $$\frac{\sum_{i=1}^{N}|x_i-\mu|}{N}.$$
  - Mean **squared** deviations (variance):
    $$\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}.$$

i	x_i	deviation d_i	$\lvert d_i \rvert$	d_i^2
1	1	$1-5=-4$	4	16
2	2	$2-5=-3$	3	9
3	4	$4-5=-1$	1	1
4	5	$1-5=0$	0	0
5	6	$6-5=1$	1	1
6	8	$8-5=3$	3	9
7	9	$9-5=4$	4	16
Mean	5	0	2.29	7.43

Measuring variability

- **Larger** MADs and variances means the data are **more disperse**.
- Consider two 7-student groups and their grades:
  - Group 1: 70, 72, 75, 76, 78, 80, 81.
  - Group 2: 58, 63, 68, 74, 82, 90, 97.

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	70	-6	6	36
2	72	-4	4	16
3	75	-1	1	1
4	76	0	0	0
5	78	2	2	4
6	80	4	4	16
7	81	5	5	25
Mean	76	0	3.14	14

i	x_i	d_i	$\lvert d_i \rvert$	d_i^2
1	58	-18	18	324
2	63	-13	13	169
3	68	-8	8	64
4	74	-2	2	4
5	82	6	6	36
6	90	14	14	196
7	97	21	21	441
Mean	76	0	11.71	176.29

MADs vs. variances

- The main difference:
  - An MAD puts the same weight on all values.
  - A variance puts more weights on **extreme values**.
- They may give different ranks of dispersion:

i	x_i	d_i	$\|d_i\|$	d_i^2
1	0	−5	5	25
2	4	−1	1	1
3	5	0	0	0
4	6	1	1	1
5	10	5	5	25
Mean	5	0	2.4	10.4

i	x_i	d_i	$\|d_i\|$	d_i^2
1	1	4	4	16
2	2	3	3	9
3	5	0	0	0
4	8	3	3	9
5	9	4	4	16
Mean	5	0	2.8	10

- In general, people use variances more than MADs.
  - But MADs are still popular in some areas, e.g., demand forecasting.
  - It is the analyst's discretion to choose the appropriate one.

Standard deviations

- One drawback of using variances is that the unit of measurement is the **square** of the original one.
- For the baseball team, the variance of member heights is 34.05 $\mathbf{cm}^2$. What is it?!
- People take the square root of a variance to generate a **standard deviation**.
- The standard deviation of member heights is
$$\sqrt{34.05} \approx 5.85 \text{ cm}.$$
- A standard deviation typically has more managerial implications.

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

z-scores

- Consider a set of sample data $\{x_i\}_{i=1,...,n}$ with sample mean $\bar{x}$ and sample standard deviation s. For x_i, the **z-score** is

$$z_i = \frac{x_i - \bar{x}}{s}.$$

- In a set of population data $\{x_i\}_{i=1,...,N}$ with population mean μ and population standard deviation σ, the z-score of x_i is

$$z_i = \frac{x_i - \mu}{\sigma}.$$

- A value's z-score measures for **how many standard deviations** it deviates from the mean.

z-scores vs. outliers

- For detecting **outliers**, one common way is double check whether x_i is an outlier if
$$|z_i| = \left|\frac{x_i - \mu}{\sigma}\right| > 3.$$
  - It is quite rare for a value's magnitude of z-score to be so large.
  - For sample data, use $\frac{x_i - \bar{x}}{s}$.
- Some people propose the use of median and MAD is a similar way: double check whether x_i is an outlier if[1]
$$\left|\frac{x_i - \text{median}}{\text{MAD}}\right| > 3.$$
- The above rules only **suggest** one to investigate some extreme values again. These rules are neither sufficient nor necessary for outliers.

[1] The "MAD" here can be mean absolute deviation from mean, mean absolute deviation from median, median absolute deviation from median, etc.

Population v.s. sample variances

- Recall that the formulas for population and sample means are

$$\mu \equiv \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{and} \quad \bar{x} \equiv \frac{\sum_{i=1}^{n} x_i}{n}, \text{ respectively.}$$

  - Formula-wise there is no difference.
- However, **population** and **sample variances** are

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

  - Note the difference between N and $n - 1$!
  - Population and sample standard deviations are $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$, respectively.
- People use σ^2, σ, s^2, and s in almost the whole statistics world.

Coefficient of variation

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

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

- When will you use coefficients of variation?

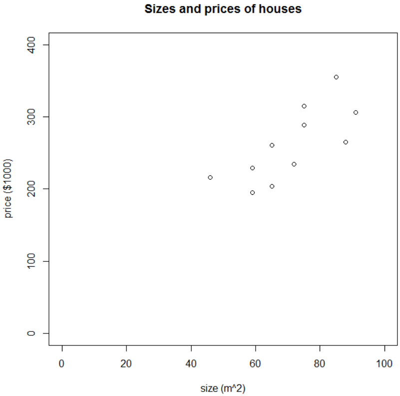
Road map

- Describing central tendency.
- Describing variability.
- **Describing correlation.**

Introduction

- Consider the size of a house and its price in a city:

Size (in m^2)	Price (in \$1000)
75	315
59	229
85	355
65	261
72	234
46	216
107	308
91	306
75	289
65	204
88	265
59	195



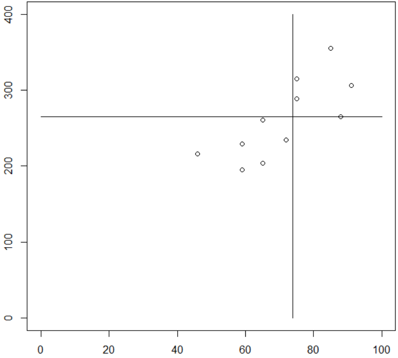
- How do we measure/describe the **correlation** (linear relationship) between the two variables?

Intuition

- Consider a set of paired data $\{(x_i, y_i)\}_{i=1,...,N}$.
- When one variable goes up, does the other one **tend to** go up or down?
- More precisely, if x_i is larger than μ_x (the mean of the x_is), is it more likely to see $y_i > \mu_y$ or $y_i < \mu_y$?
- Let's highlight the two means on the scatter plot.

Intuition

- The scatter plot with the two means:



- We say that the two variables have a **positive** correlation.
  - If one goes up when the other goes down, there is a **negative** correlation.

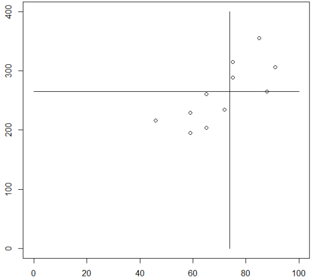
Covariances

- We define the **covariance** of a set of two-dimensional **population** data as

$$\sigma_{xy} \equiv \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}.$$

  - If most points fall in the first and third quadrants, most $(x_i - \mu_x)(y - \mu_y)$ will be positive and σ_{xy} tends to be positive.
  - Otherwise, σ_{xy} tends to be negative.
- The **sample covariance** is

$$s_{xy} \equiv \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$



Example: house sizes and prices

- For our example:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
75	315	1.08	50.25	54.44
59	229	−14.92	−35.75	533.27
85	355	11.08	90.25	1000.27
65	261	−8.92	−3.75	33.44
72	234	−1.92	−30.75	58.94
46	216	−27.92	−48.75	1360.94
107	308	33.08	43.25	1430.85
91	306	17.08	41.25	704.69
75	289	1.08	24.25	26.27
65	204	−8.92	−60.75	541.69
88	265	14.08	0.25	3.52
59	195	−14.92	−69.75	1040.44
$\bar{x} = 73.92$	$\bar{y} = 264.75$	–	–	$s_{xy} = 617.16$

- So the covariance of house size and price is 617.16.
- Is it large or small?
  - This depends on how variable the two variables themselves are.

Correlation coefficients

- To take away the auto-variability of each variable itself, we define the population and sample **correlation coefficients** as

$$\rho \equiv \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \text{and} \quad r \equiv \frac{s_{xy}}{s_x s_y},$$

  - σ_x and σ_y are the population standard deviations of x_is and y_is.
  - s_x and s_y are the sample standard deviations of x_is and y_is.
  - In our example, we have $r = \frac{617.16}{16.78 \times 50.45} \approx 0.729$.
- It can be shown that we always have

$$-1 \leq \rho \leq 1 \quad \text{and} \quad -1 \leq r \leq 1.$$

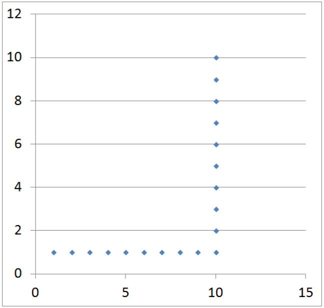
  - $\rho > 0$ $(s > 0)$: Positive correlation.
  - $\rho = 0$ $(s = 0)$: No correlation.
  - $\rho < 0$ $(s < 0)$: Negative correlation.

Magnitude of correlation

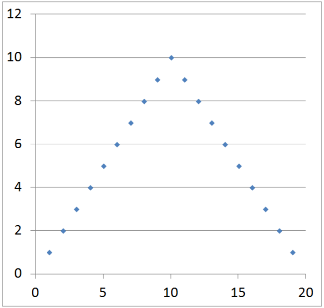
- In practice, people often determine the degree of correlation based on $|\rho|$ or $|s|$:
  - $0 \leq |\rho| < 0.25$ or $0 \leq |s| < 0.25$: A weak correlation.
  - $0.25 \leq |\rho| < 0.5$ or $0.25 \leq |s| < 0.5$: A moderately weak correlation.
  - $0.5 \leq |\rho| < 0.75$ or $0.5 \leq |s| < 0.75$: A moderately strong correlation.
  - $0.75 \leq |\rho| \leq 1$ or $0.75 \leq |s| \leq 1$: A strong correlation.

Correlation vs. independence

- A correlation coefficient only measures how one variable **linearly** depends on the other variable.



$(r = 0.5973)$



$(r = 0)$

- Being **uncorrelated** does not mean being **independent**!