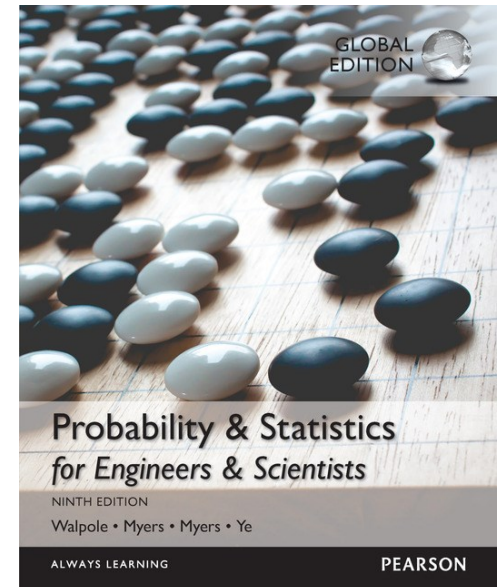


# Chapter 1

# Introduction to Statistics and Data Analysis

School of Computing, Gachon Univ.  
Joon Yoo



# Introduction

- People have always had to cope with uncertainty
  - E.g., the weather, their food supply, traffic conditions, and other aspects of their environment
    - Even in gambling (Dice game, Coin toss, Lotteries ...)
- and have made efforts to reduce this uncertainty and its effects
- **Probability theory** is devoted to the study of *uncertainty* and *variability*
  - Probability quantifies how uncertain we are about future events

# What is Statistics?

- Statistics is a discipline which is concerned with:
  - summarizing information to aid understanding,
  - drawing conclusions from data,
  - estimating the present or predicting the future, and
  - designing experiments and other data collection.

- **Why Study Statistics?**

**Answers provided by statistical approaches can provide the basis for making decisions or choosing actions.**

- A supplier fills cans of soda marked 12 ounces. How much soda does each can really contain?



- It is very unlikely any one can contains exactly 12 ounces of soda = There is **variability** in any process.
  - ✓ Some cans contain a little more than 12 ounces, and some cans contain a little less.
- On the average, there are 12 ounces in each can.
- The supplier hopes there is little variability in the process, that most cans contain close to 12 ounces of soda.

# Measure and Variability

- **Data**

Data consists of information coming from observation, counts, measurements, or responses.

- **Variability in Scientific Data**

There will always be **variability** in the data.

[**Think!**] If the observed data were always the same and always on the target, there would be no need for statistical methods.

**One of the primary objectives of statistics:** measuring and characterizing variability.

# Population, Sample

- **Population** \*

- A *population* is the collection of all outcomes, responses, measurements, or counts that are of interest.

- **Sample** \*\*

- A *sample* is a subset of a population.
- Information is gathered in the form of **samples**, or collections of **observations**.



■ 조사: 글로벌리서치, 의뢰: JTBC,  
2월 28일~3월 1일 전국 성인 1,006명 조사  
(신뢰수준 95%, 표본오차  $\pm 3.1\%$ )

- **Parameter** \*\*\*

- A *parameter* is numerical description of a population characteristics.
  - E.g., mean, variance, max value, ...

**\*모집단 \*\*표본 \*\*\*모수(파라미터)**

# Example: Population, Sample

## Population

(Some Unknown Parameters)

Example:

High school students in Korea  
(Height Mean)

**N=Population Size**



## Sample = Observations

(We calculate Some Statistics)

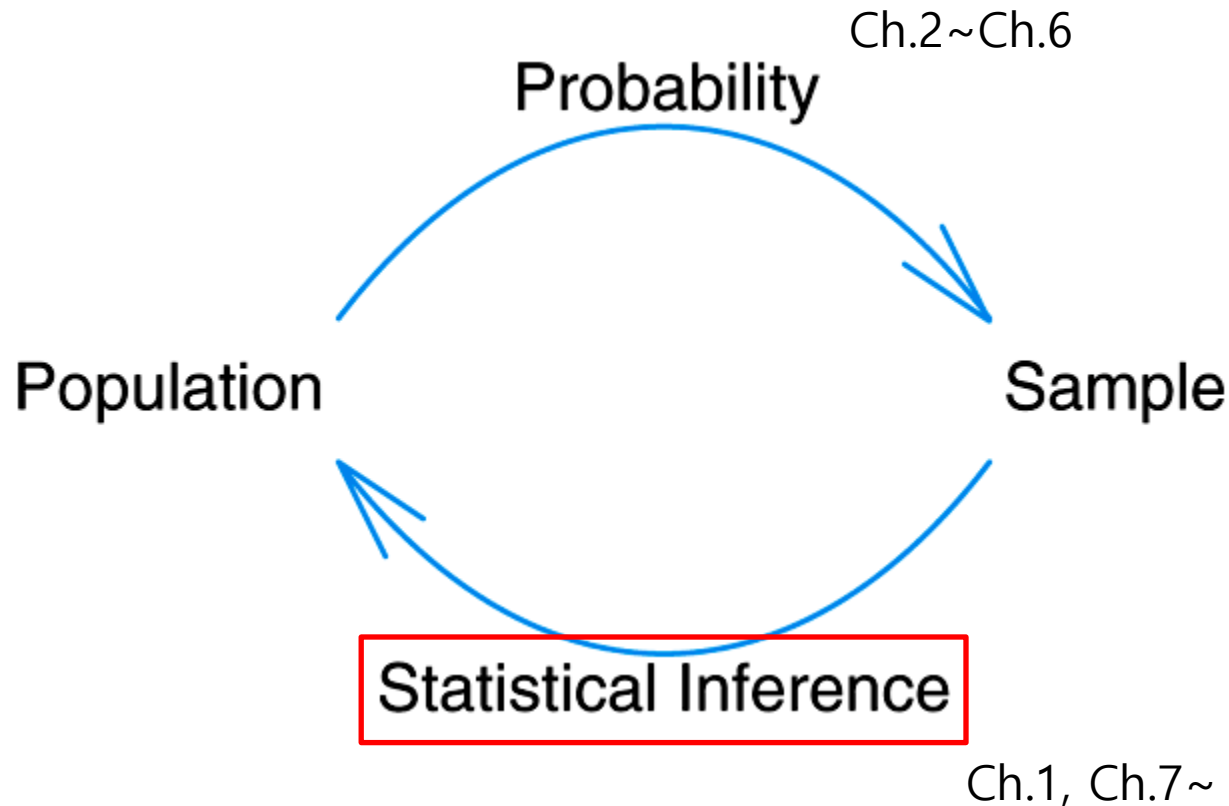
Example: 1000 high school students  
(Sample Mean)

**n = Sample Size**

- Let  $X_1, X_2, \dots, X_N$  be the **population values** (in general, they are unknown)
- Let  $x_1, x_2, \dots, x_n$  be the **sample values** (these values are known by measurements or observations )
- **Statistics** obtained from the **sample** are used to estimate (approximate) the **parameters** of the **population**.

■ 조사: 글로벌리서치, 의뢰: JTBC,  
2월 28일~3월 1일 전국 성인 1,006명 조사  
(신뢰수준 95%, 표본오차  $\pm 3.1\%$ )

# Fundamental relationship between probability and inferential statistics





# Comparison of **Probability** and **Statistics**

**Probability**: Properties of the **population** are assumed known. Answer questions about the **sample** based on these properties.

**Statistics**: Use information in the **sample** to draw a conclusion about the **population**.



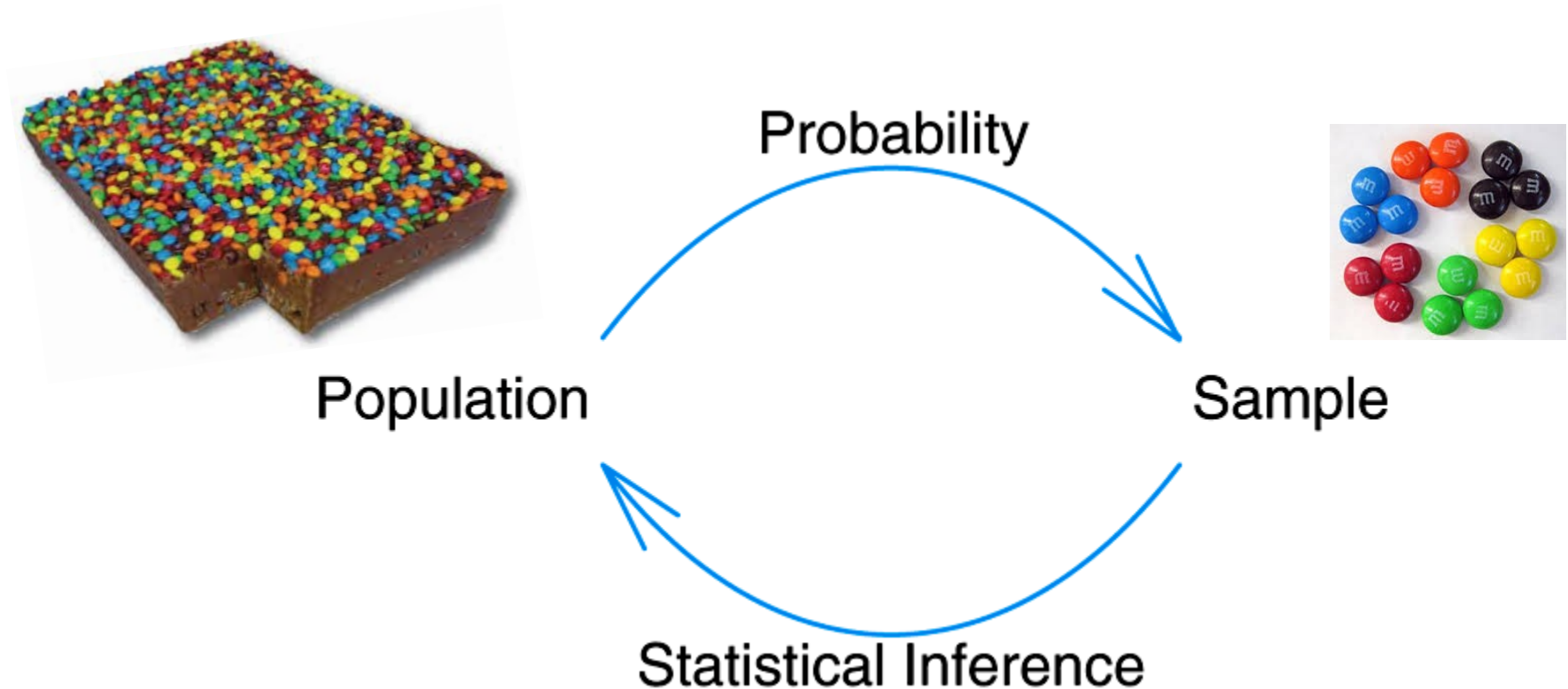
**Example:** A jar of M&M's contains 100 candy pieces, 15 are red. A handful of 10 is selected.

**Probability question:** What is the probability that 3 of the 10 selected are red?

**Example:** A handful of 10 M&M's is selected from a jar containing 1000 candy pieces. Three M&M's in the handful are red.

**Statistics question:** What is the proportion of red M&M's in the entire jar?

# Fundamental relationship between probability and inferential statistics



# Population and samples

- Is the sample informative about the total population?
- Does the sample share the same probability model with the total population?



- Choose the sample in a totally **random** fashion without any prior considerations

# Flipped learning assignment

- Enroll Introduction to Probability and Data in Coursera (choose audit (청강))
- View the video regarding “*sampling and sources of bias*”

The screenshot shows the Coursera interface for the course 'Introduction to Probability and Data'. The left sidebar lists the course overview and designing studies sections. The main video player shows a lecture titled 'a few sources of sampling bias' with a subtitle 'our study would suffer from convenience bias.' and a poll source citation.

**Course Overview**

- Introduction to Statistics with R (2 min)
- More about Introduction to Probability and Data (10 min)
- Feedback Surveys (10 min)

**Designing Studies**

- Lesson Learning Objectives (10 min)
- Introduction (3 min)
- Data Basics (5 min)
- Observational Studies & Experiments (4 min)
- Sampling and sources of bias (8 min)**
- Experimental Design (2 min)
- (Spotlight) Random Sample Assignment (3 min)

**Video Content:**

a few sources of sampling bias

- ▶ **Convenience sample:** Individuals who are easily accessible are more likely to be included in the sample

our study would suffer from convenience bias.

Poll source: edition.cnn.com, August 29, 2013

Sampling and sources of bias

<https://www.coursera.org/lecture/probability-intro/sampling-and-sources-of-bias-Y96uT>

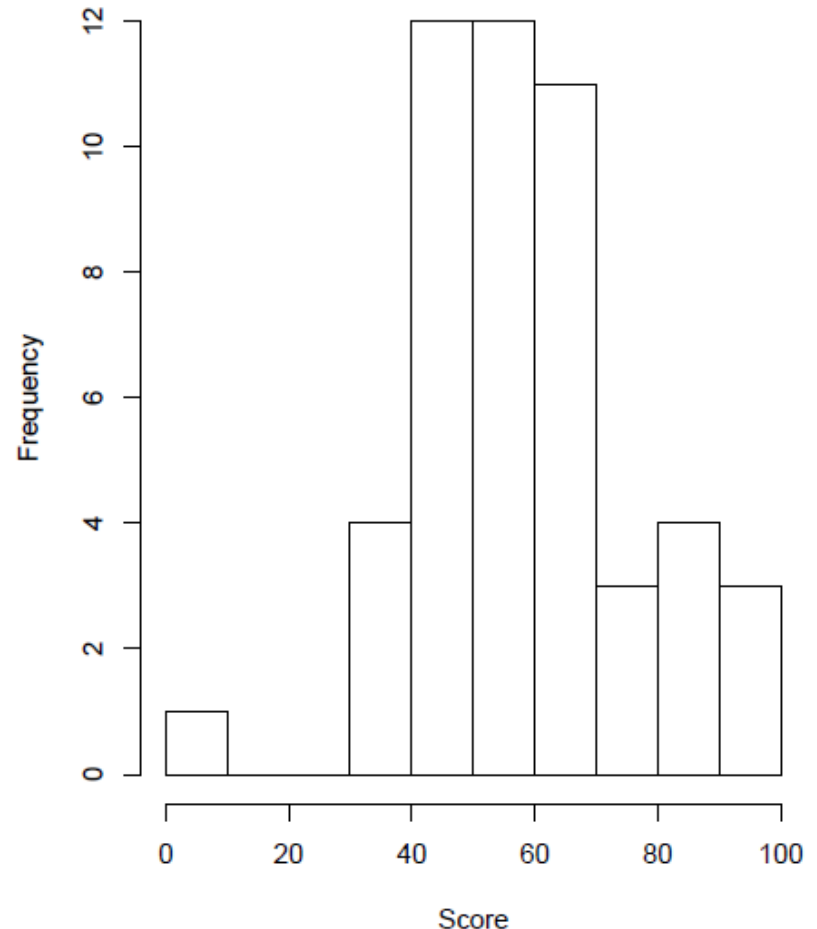
# Descriptive vs. Inferential Statistics

- **Descriptive statistics** (기술통계)
  - the branch of statistics that involves the description, organization, and summarization of data.
- **Inferential statistics** (추론통계)
  - the branch of statistics that involves using a sample to draw conclusions about a population.
  - A basic tool in the study of inferential statistics is **probability.**

# Descriptive statistics (기술통계)

- Mean (average)
- Median
- Variability : Sample range (Max – Min), Variance, standard deviation
- Distribution

Histogram of Scores for Class A

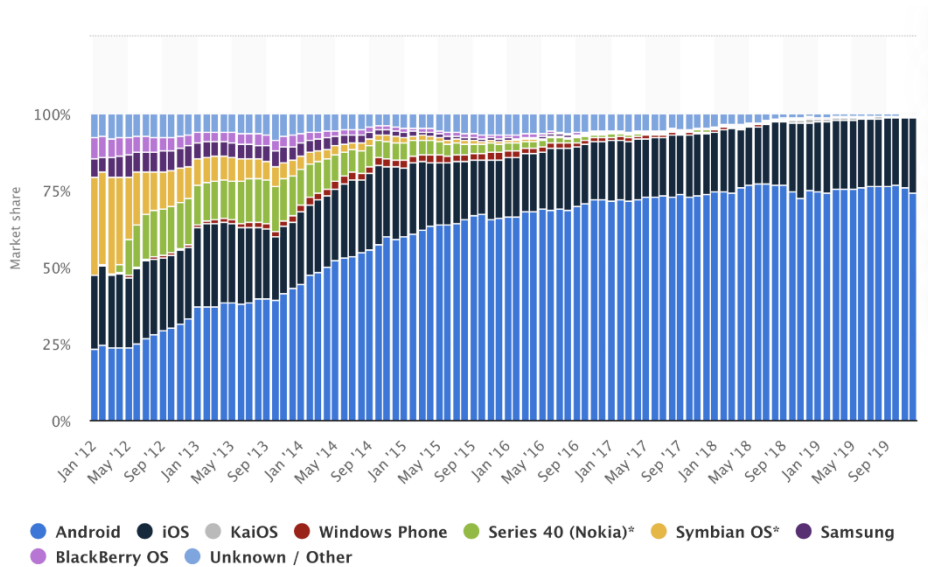


# Descriptive statistics

- Mobile OS market share

- <https://www.designveloper.com/vi/blog/mobile-app-development-android-vs-ios/>
- [https://www.researchgate.net/figure/Smartphone-Market-Share-for-Android-Device-and-iOS-15\\_fig2\\_345342497](https://www.researchgate.net/figure/Smartphone-Market-Share-for-Android-Device-and-iOS-15_fig2_345342497)

Graph



Table

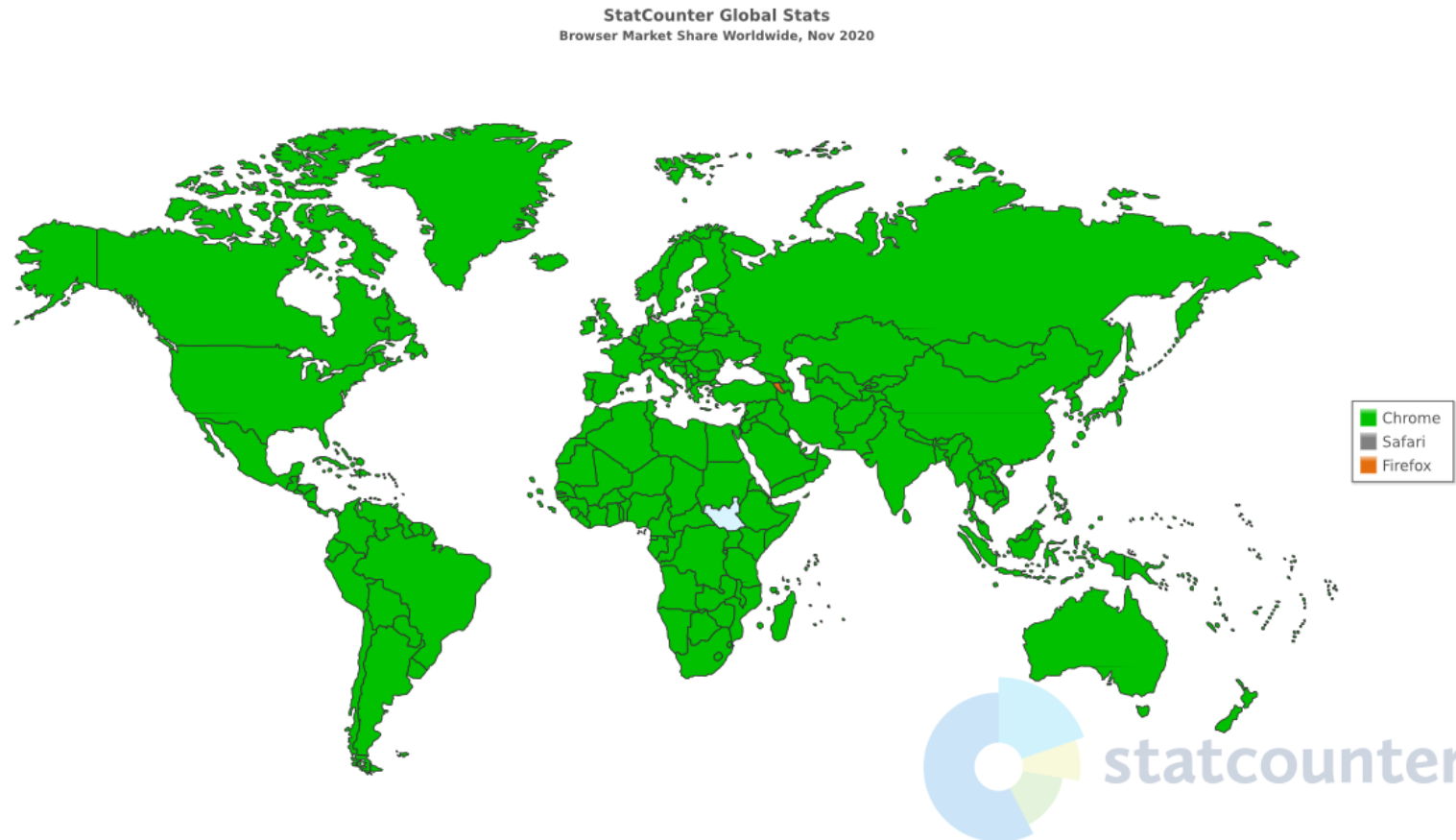
Year	Android	iOS	Other	Total
2017	85.1%	14.7%	0.2%	100.0%
2018	85.1%	14.9%	0.0%	100.0%
2019	86.6%	13.4%	0.0%	100.0%
2020	86.6%	13.4%	0.0%	100.0%
2021	86.9%	13.1%	0.0%	100.0%
2022	87.0%	13.00%	0.0%	100.0%
2023	87.15%	12.9%	0.0%	100.0%



# Descriptive statistics

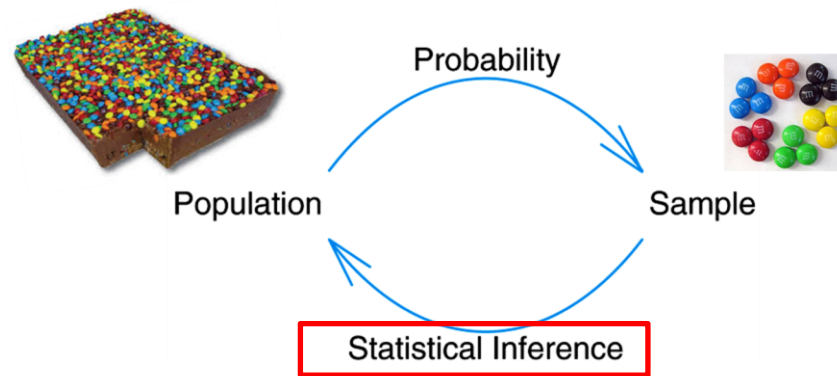
- “Market share of web browsers” [Wikipedia]

## Geographical trend



# Inferential statistics (추론통계)

- Inferential statistics
  - The branch of statistics that involves using a sample to draw conclusions about a population.




# Example (Descriptive vs. Inferential Statistics)

- A large sample of men, aged 48, was studied for 18 years. For unmarried men, 60% to 70% were alive (생존) at age 65. For married men, 90% were alive at age 65.
- Which part of the study represents the descriptive (기술) branch statistics? What conclusions might be drawn from this study using inferential (추론) statistics?

# Answer:

- Descriptive statistics:
  - For unmarried men, 60% to 70% were alive at age 65.  
For married men, 90% were alive at age 65.
- A possible inference:
  - Being married is associated with a longer life for men.

# Example 1.1

Suppose that an engineer encounters data from a manufacturing process in which 100 items are sampled and 10 are found to be defective. It is expected and anticipated that occasionally there will be defective items. Obviously these 100 items represent the sample. However, it has been determined that in the long run, the company can only tolerate 5% defective in the process. Now, the elements of probability allow the engineer to determine how conclusive the sample information is regarding the nature of the process. In this case, the **population** conceptually represents all possible items from the process. Suppose we learn that *if the process is acceptable*, that is, if it does produce items no more than 5% of which are defective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process. This small probability suggests that the process does, indeed, have a long-run rate of defective items that exceeds 5%. In other words, under the condition of an acceptable process, the sample information obtained would rarely occur. However, it did occur! Clearly, though, it would occur with a much higher probability if the process defective rate exceeded 5% by a significant amount. 

# Example 1.2

- **Table 1.1** Data Set for Example 1.2

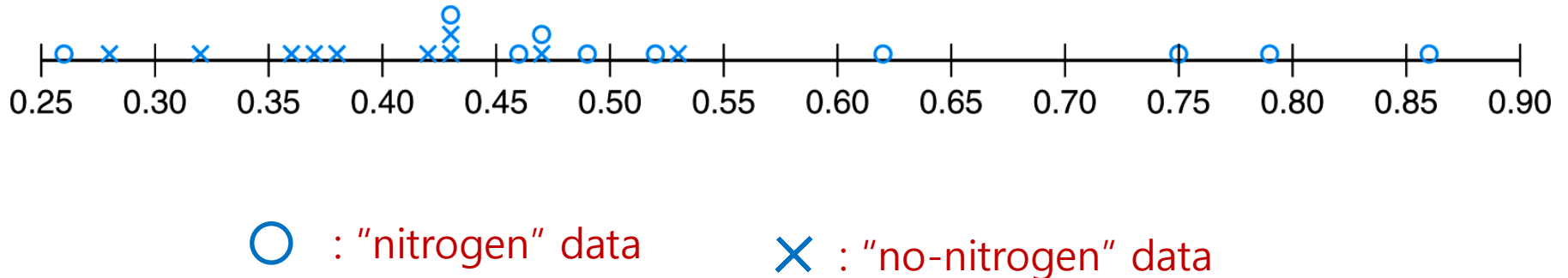
Stem weight data (in grams)



No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

# Example 1.2

- **Figure 1.1** A dot plot of stem weight data



- **What can you observe from this experiment ?**
  - Does the use of nitrogen have effect ?
  - How can this be quantified ?

# Summarizing Data Sets

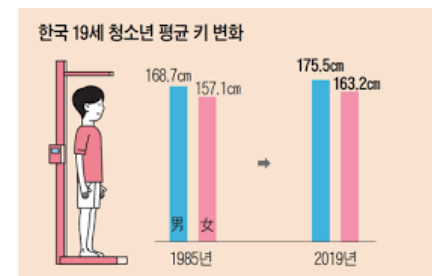




# Sample Mean (평균)

- Summarizing Statistics
  - A numerical quantity whose value is determined by data
- Suppose that we have  $n$  numerical values  $(x_1, x_2, x_3, x_4, \dots, x_n)$ 
  - Sample mean  $\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$
  - For constants  $a$  and  $b$ , if  $y_i = ax_i + b$  for  $i=1, \dots, n$   
Then sample mean of the data set  $y_1, \dots, y_n$  is,

$$\bar{y} = a\bar{x} + b$$



# Example 1.3

- Example : The winning scores in the U.S. Masters golf tournament in the years from 1999 to 2008 were as follows:

280, 278, 272, 276, 281, 279, 276, 281, 289, 280

Compute sample mean. Answer:  $\bar{x} = \mathbf{279.2}$

# Example 1.4

- Example : Compute mean

Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

**SOLUTION**

$$\bar{x} = (15 \cdot 2 + 16 \cdot 5 + 17 \cdot 11 + 18 \cdot 9 + 19 \cdot 14 + 20 \cdot 13) / 54 \approx 18.24 \quad \blacksquare$$

$$\bar{x} = \sum_{i=1}^k v_i f_i / n$$

By writing the preceding as

$$\bar{x} = \frac{f_1}{n} v_1 + \frac{f_2}{n} v_2 + \cdots + \frac{f_k}{n} v_k$$

# Sample Median

- Purpose: Reflect the central tendency of the sample that is uninfluenced by extreme values or outliers
- If  $n$  is odd
  - the value in position  $(n+1)/2$
  - Example: 1 2 3
- If  $n$  is even
  - the average of the values in positions  $n/2$  and  $n/2 + 1$
  - Example: 1 2 3 4

**Definition 1.2:** Given that the observations in a sample are  $x_1, x_2, \dots, x_n$ , arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

# Example 1.5

As an example, suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9.$$

Clearly, the mean is influenced considerably by the presence of the extreme observation, 14.7, whereas the median places emphasis on the true “center” of the data set.

# Example 1.6

- Data from Example 1.2



Stem weight data (in grams)

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

$$\bar{x} \text{ (no nitrogen)} = 0.399 \text{ gram,}$$

$$\tilde{x} \text{ (no nitrogen)} = \frac{0.38 + 0.42}{2} = 0.400 \text{ gram,}$$

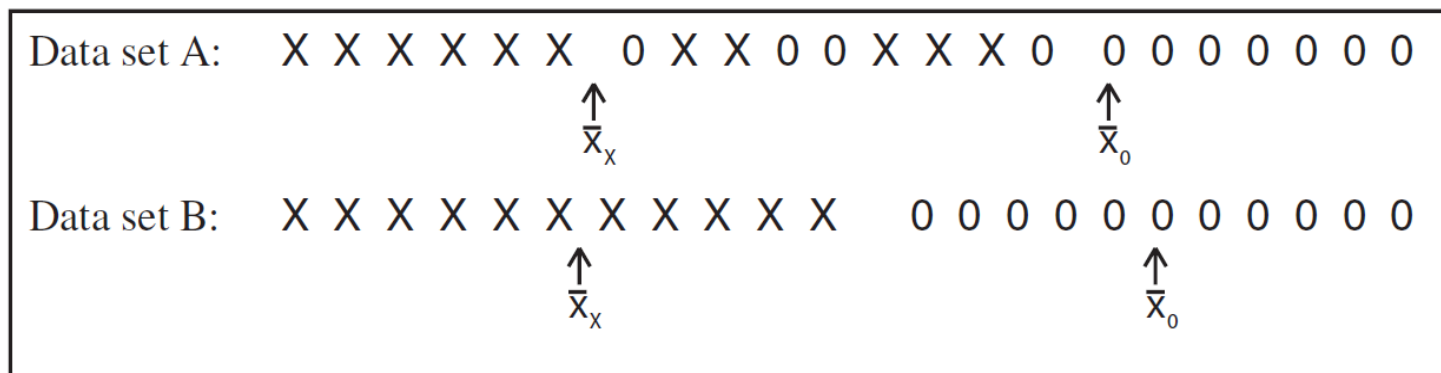
$$\bar{x} \text{ (nitrogen)} = 0.565 \text{ gram,}$$

$$\tilde{x} \text{ (nitrogen)} = \frac{0.49 + 0.52}{2} = 0.505 \text{ gram.}$$

- What can you observe from this experiment ?
  - Does the use of nitrogen have effect ?
  - How can this be quantified ?

# Variability

As another example, contrast the two data sets below. Each contains two samples and the difference in the means is roughly the same for the two samples, but data set B seems to provide a much sharper contrast between the two populations from which the samples were taken. If the purpose of such an experiment is to detect differences between the two populations, the task is accomplished in the case of data set B. However, in data set A the large variability *within* the two samples creates difficulty. In fact, it is not clear that there is a distinction *between* the two populations.

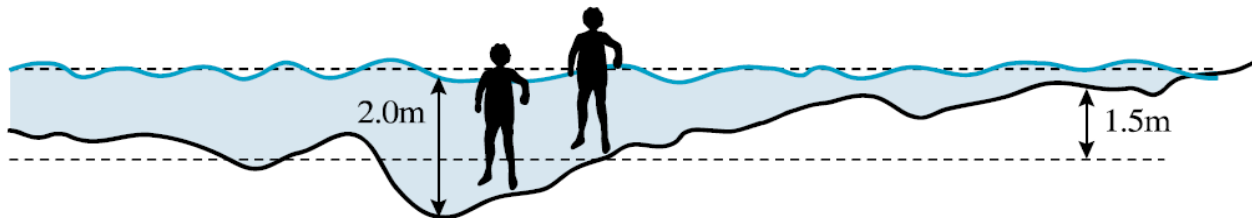


# Sample Variance

- Sample variance

- The spread or variability of the data values

❖ “평균키 180cm 병사들이 평균수심 150cm 강 건너다 빠져 죽은 건...”

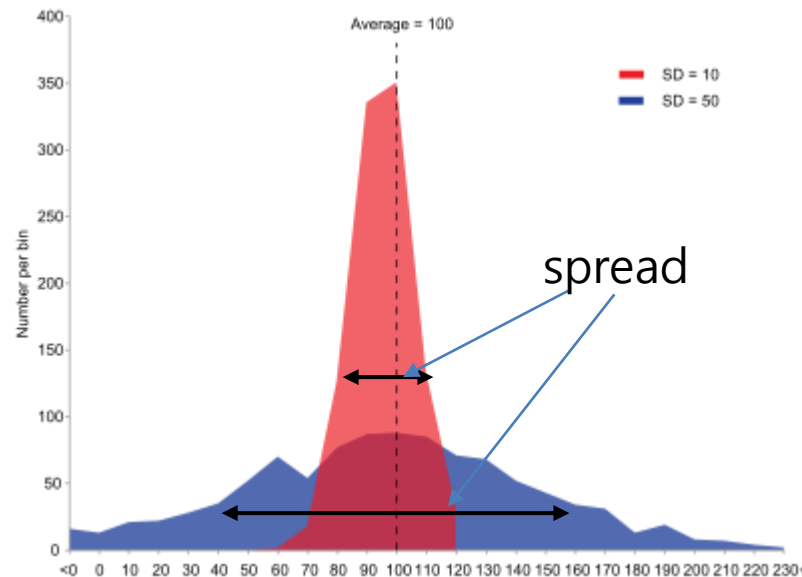


Considering the average value alone, the characteristics of the collected data cannot be fully summarized!



# Sample Variance

- Sample variance
  - The spread or variability of the data values



Q: How can we formulize the concept of “spread”?

# Sample Variance

- **Sample Range:**  $X_{max} - X_{min}$ .

## Definition 1.3:

The **sample variance**, denoted by  $s^2$ , is given by

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

The **sample standard deviation**, denoted by  $s$ , is the positive square root of  $s^2$ , that is,

$$s = \sqrt{s^2}.$$

# Why (n-1)?

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

- n-1: degrees of freedom associated with the variance estimate
  - depict the number of **independent** pieces of information available for computing variance

For example, suppose that we wish to compute the sample variance and standard deviation of the data set (5, 17, 6, 4). The sample average is  $\bar{x} = 8$ . The computation of the variance involves

$$(5 - 8)^2 + (17 - 8)^2 + (6 - 8)^2 + (4 - 8)^2 = (-3)^2 + 9^2 + (-2)^2 + (-4)^2.$$

The quantities inside parentheses sum to zero. In general,  $\sum_{i=1}^n (x_i - \bar{x}) = 0$

Then the computation of a sample variance does not involve  $n$  **independent squared deviations** from the mean  $\bar{x}$ . In fact, since the last value of  $x - \bar{x}$  is determined by the initial  $n - 1$  of them, we say that these are  $n - 1$  “pieces of information” that produce  $s^2$ . Thus, there are  $n - 1$  degrees of freedom rather than  $n$  degrees of freedom for computing a sample variance.

# Example 1.7

- Find the sample variance of the data sets **A** and **B**

$$\mathbf{A}: 3, 4, 6, 7, 10 \quad \mathbf{B}: -20, 5, 15, 24$$

**SOLUTION** As the sample mean for data set **A** is  $\bar{x} = (3 + 4 + 6 + 7 + 10)/5 = 6$ , it follows that its sample variance is

$$s^2 = [(-3)^2 + (-2)^2 + 0^2 + 1^2 + 4^2]/4 = 7.5$$

The sample mean for data set **B** is also 6; its sample variance is

$$s^2 = [(-26)^2 + (-1)^2 + 9^2 + (18)^2]/3 \approx 360.67$$

Thus, although both data sets have the same sample mean, there is a much greater variability in the values of the **B** set than in the **A** set. ■

# Describing data sets

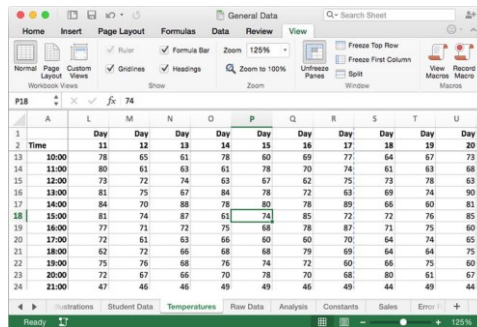


# Describing data sets

- Numerical findings should be presented
  - Clearly
  - Concisely
  - Easy to find out the characteristics of data

[Tables]

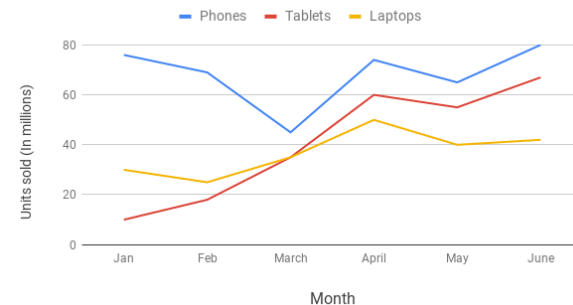
- Ways



<raw data>

Month	Sales	Cost	Profit	ROI
Jan	10	6	4	66.67%
Feb	20	15	5	33.33%
Mar	30	24	6	25.00%
Apr	40	33	7	21.21%
May	50	42	8	19.05%
Jun	60	51	9	17.65%

Device sales



[Graphs]

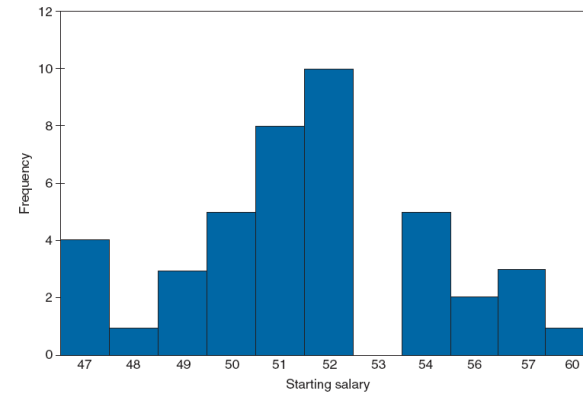
# Frequency tables and graphs

- A data set with relatively small number of distinct values
  - Frequency table
  - Bar graph
  - Frequency polygon

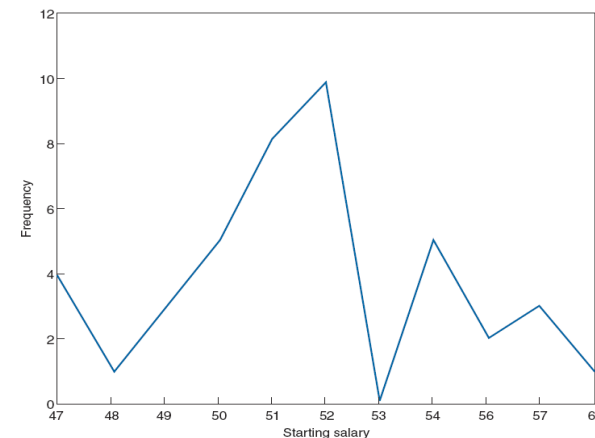
TABLE 2.1 *Starting Yearly Salaries*

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

Frequency table



Bar graph

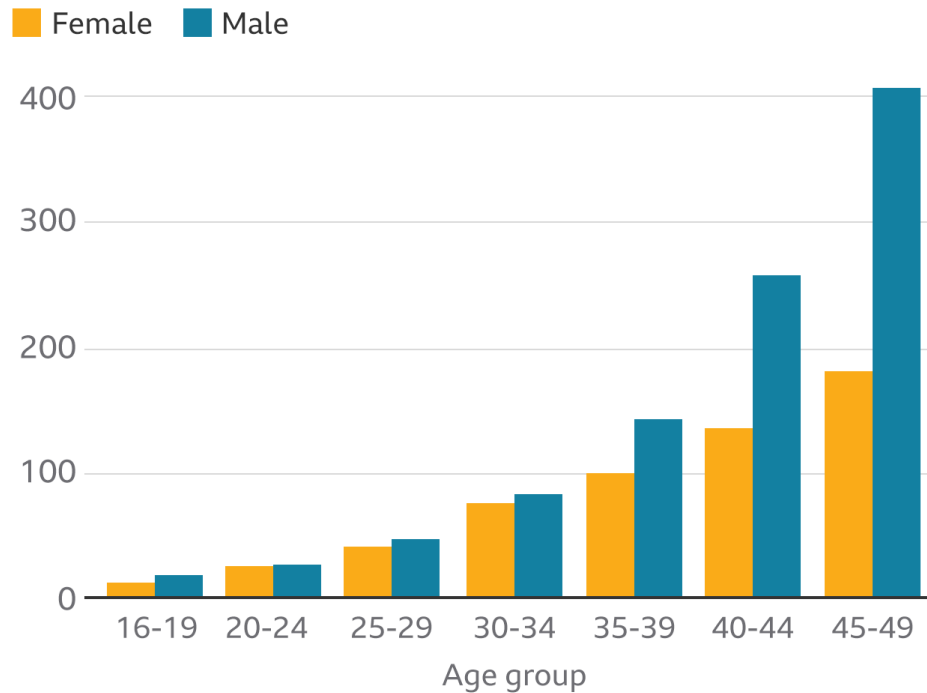


Frequency polygon

# Frequency graphs: Example

## Critically ill patients with Covid-19, Aug to Jan

Number of patients under 50 admitted per 1m population



Source: Public Health England

BBC

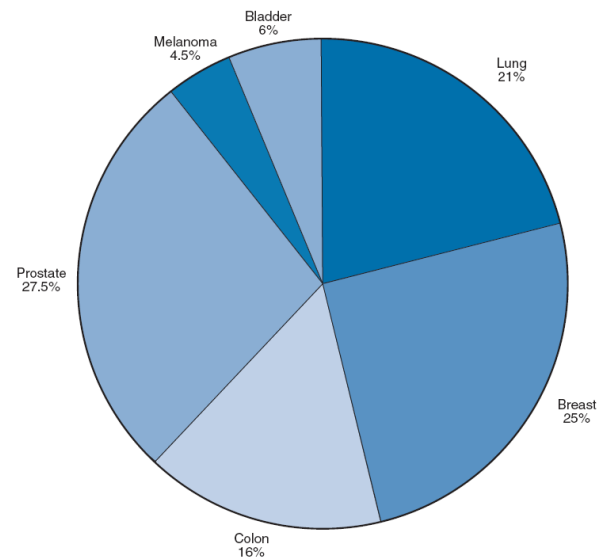
<https://www.bbc.com/news/health-56208674>



# Relative frequency tables and graphs

- Given
  - Data set consisting of  $n$  values
  - Frequency  $f$  of a particular value
- **Relative frequency**
  - The ratio  $f/n$
  - The proportion of the data having that value
- Graphs
  - Relative frequency line, bar, polygon graphs
- Pie chart
  - To show relative frequency

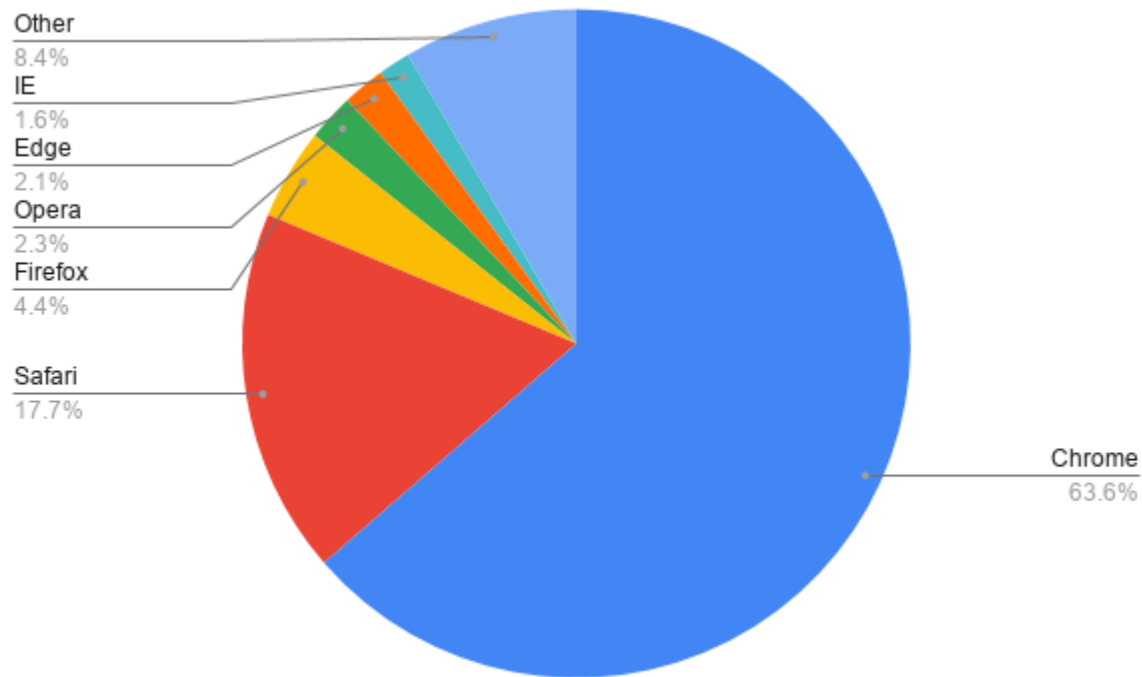
Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06



<Pie chart>

# Pie chart: Example

## Web browser market share 2020



<https://www.wearegecko.co.uk/>

# Grouped data and histogram

- Large number of distinct values
  - Divide the values into groupings
  - Class intervals
- The number of class intervals
  - Too few: losing too much information
  - Too many: the frequencies of each class being too small
  - 5 to 10 class intervals are typical
    - But is a subjective choice
- Class boundaries
  - The end points of a class interval
  - **Left end** inclusion convention
    - 20-30 means  $20 \leq x < 30$

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

# Frequency histogram

Item Lifetimes

1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

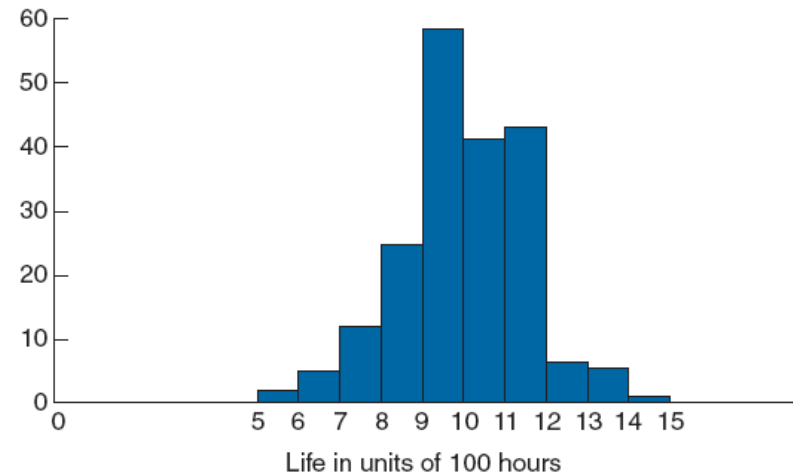
Class frequency table

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1



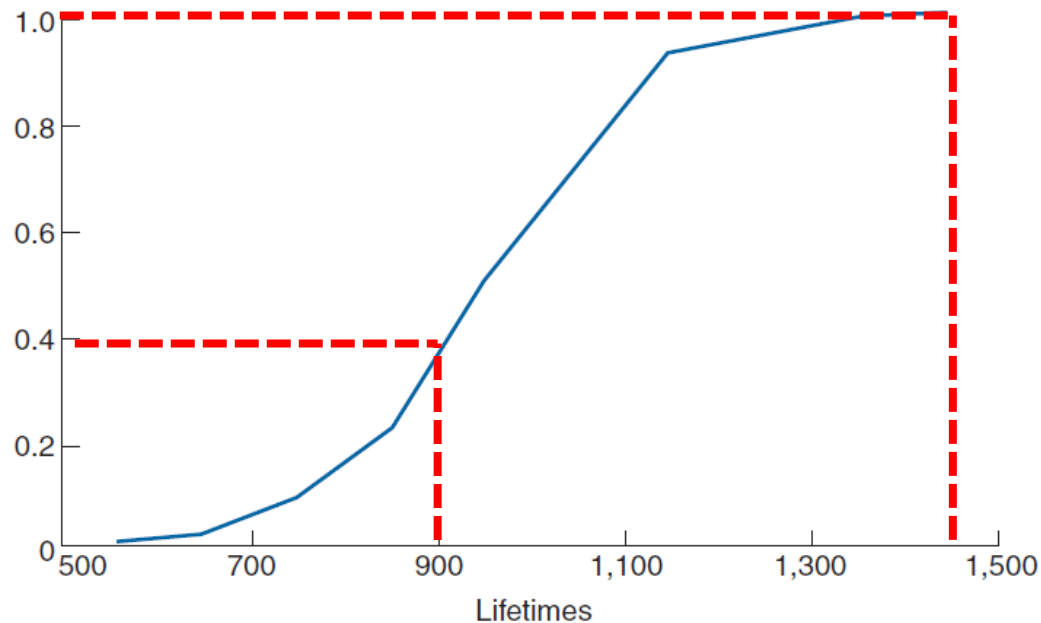
**Histogram:** Bar graph plot of class data

Number of occurrences

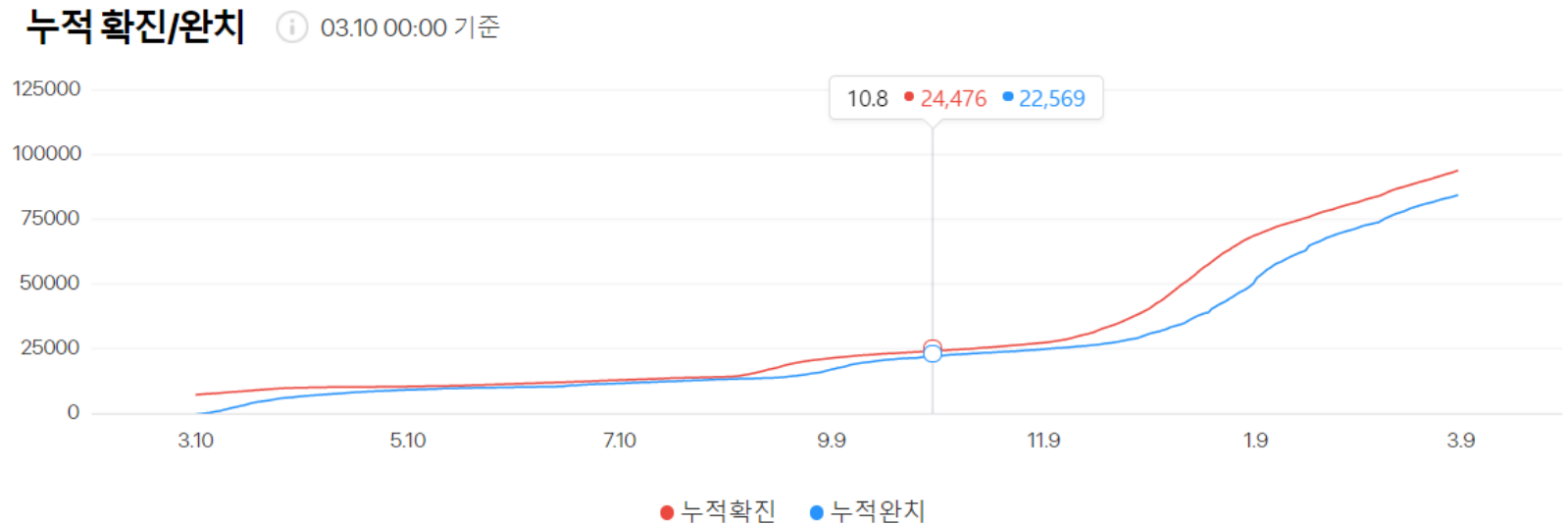


# Ogives

- Ogive
  - Cumulative (누적) frequency plot



# Ogives: Example



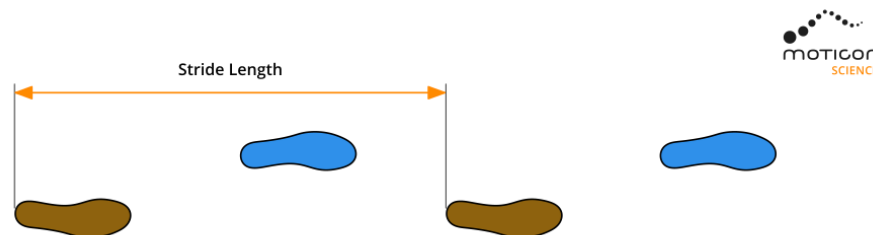
<https://news.daum.net/covid19>

# Data for Histogram

- Example: stride lengths (보폭) of 25 male students were determined, with the following results:

Stride Length (inch)				
28.6	26.5	30.0	27.1	27.8
26.1	29.7	27.3	28.5	29.3
28.6	28.6	26.8	27.0	27.3
26.6	29.5	27.0	27.3	28.0
29.0	27.3	25.7	28.8	31.4

- What can we learn about the distribution (shape) of stride lengths for this sample?



# Histogram construction

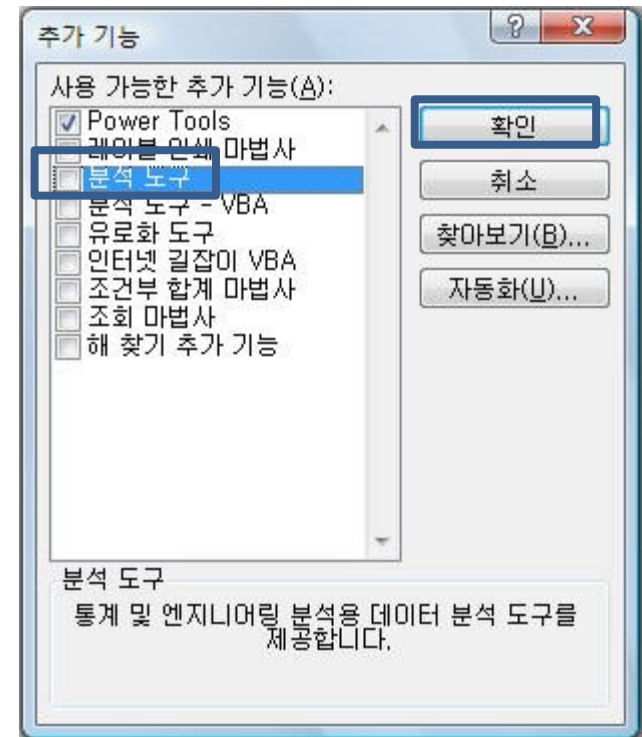
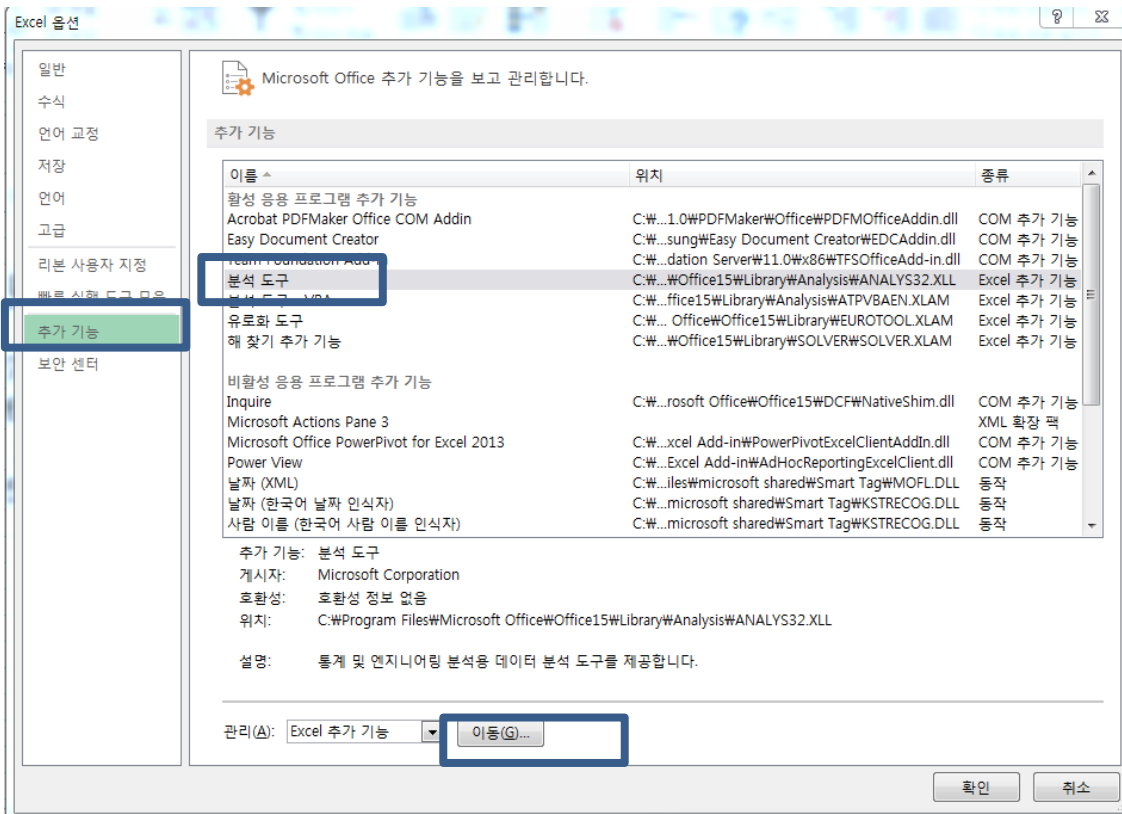
- Determining frequencies and relative frequencies

Lower	Upper	<i>Midpoint</i>	Frequency	Relative Frequency
24.85	26.20	25.525	2	0.08
26.20	27.55	26.875	10	0.40
27.55	28.90	28.225	7	0.28
28.90	30.25	29.575	5	0.20
30.25	31.60	30.925	1	0.04

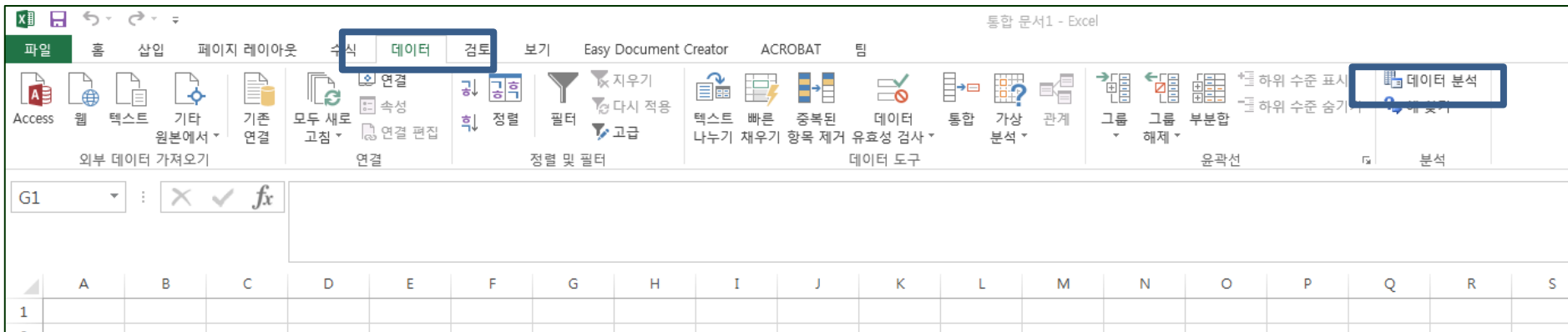


# Constructing a Histogram

- Using Excel
  - Enabling analysis tool (파일->추가기능->분석도구)



- Data → data analysis



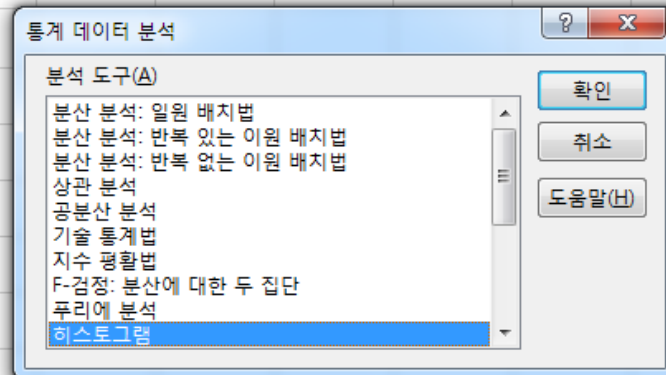
Stride Length					
28.6	26.5	30	27.1	27.8	
26.1	29.7	27.3	28.5	29.3	
28.6	28.6	26.8	27	27.3	
26.6	29.5	27	27.3	28	
29	27.3	25.7	28.8	31.4	

통계 데이터 분석

분석 도구(A)

- 분산 분석: 일원 배치법
- 분산 분석: 반복 있는 이원 배치법
- 분산 분석: 반복 없는 이원 배치법
- 상관 분석
- 공분산 분석
- 기술 통계법
- 지수 평활법
- F-검정: 분산에 대한 두 집단
- 주기에 분석
- 히스토그램**

확인  
취소  
도움말(H)



## 계급값 작성

## Stride Length

28.6	26.5	30	27.1	27.8
26.1	29.7	27.3	28.5	29.3
28.6	28.6	26.8	27	27.3
26.6	29.5	27	27.3	28
29	27.3	25.7	28.8	31.4

계급값

26.2  
27.6  
28.9  
30.3  
31.6

?

히스토그램

입력

입력 범위(I):

\$C\$4:\$G\$8

계급 구간(B):

\$I\$4:\$I\$8

☐ 이름표(L)

출력 옵션

☒ 출력 범위(O): \$K\$5:\$Q\$16

☐ 새로운 워크시트(P):

☐ 새로운 통합 문서(W)

☐ 파레토 순차적 히스토그램(A)
 

☒ 누적 백분율(M)
 ☒ 차트 출력(C)

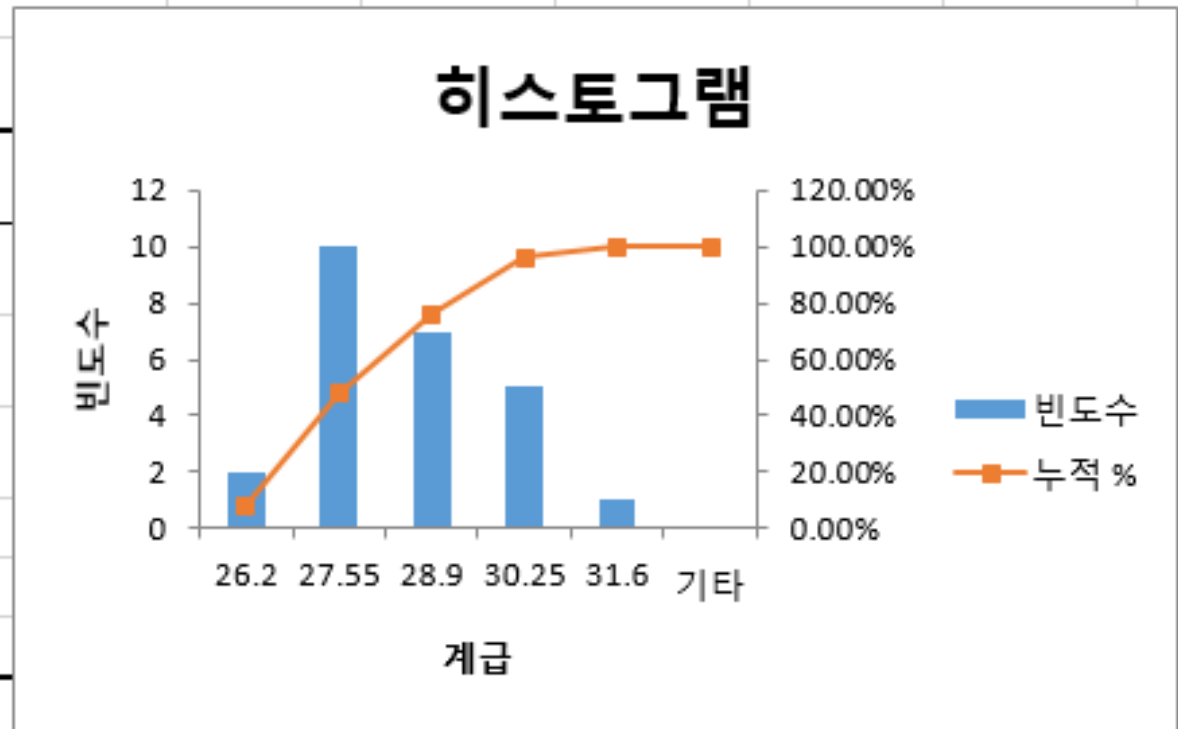
확인

취소

도움말(H)

# Constructing a Histogram

계급	빈도수	누적 %
26.2	2	8.00%
27.55	10	48.00%
28.9	7	76.00%
30.25	5	96.00%
31.6	1	100.00%
기타	0	100.00%



End of chapter



<https://www.psycom.net/bipolar-questions-answers>

# Pre-lecture Assignment

- Coursera:  
<https://www.coursera.org/learn/probability-intro/lecture/07vL4/introduction>
- Watch the 4 videos on the introduction of probability: Introduction(5min), Disjoint Events + General Addition Rule (9min), Independence (9min), Probability Examples (9min), (Spotlight) Disjoint vs. Independent (2min)