

MEC 20005 Summer 2024

경영경제통계



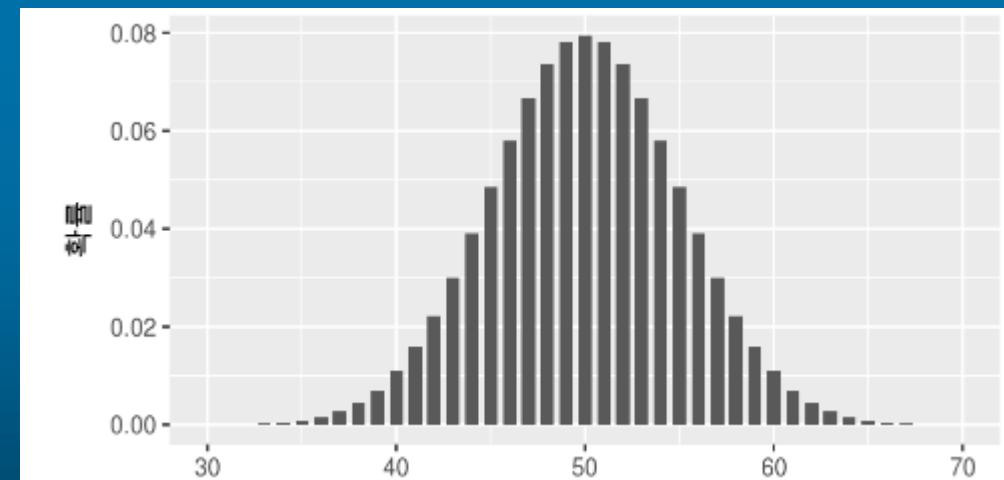
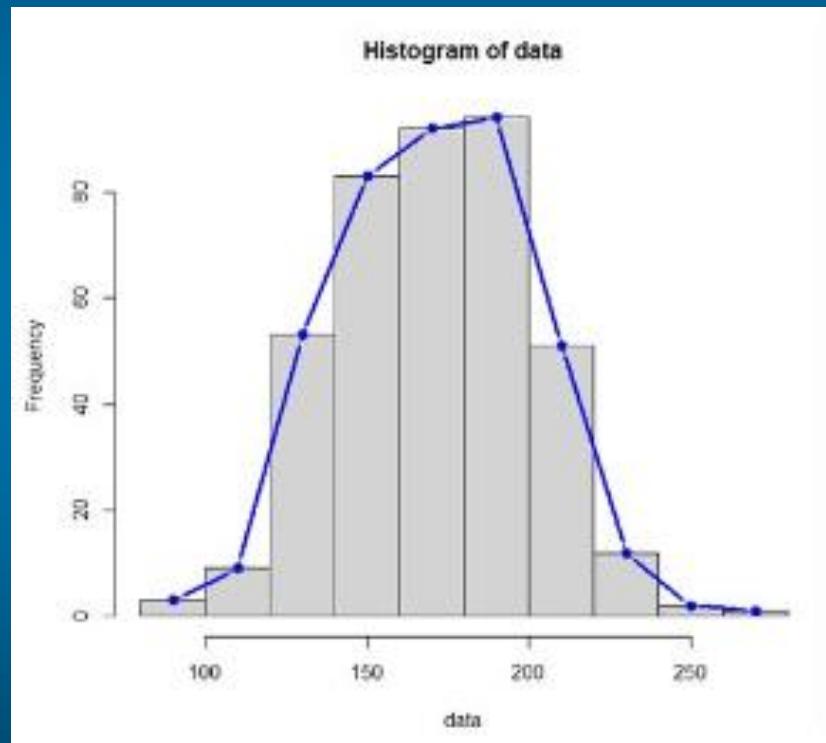
오주희 교수

한동대학교 경영경제학부

E-mail: Jooheeoh@handong.edu

June 26th 2024

도수분포, 상대도수분포, 확률분포



Descriptive Statistics



University of Pennsylvania
ScholarlyCommons

Operations, Information and Decisions Papers

Wharton Faculty Research

8-1998

Beyond the Productivity Paradox: Computers are the Catalyst for Bigger Changes

Erik Brynjolfsson

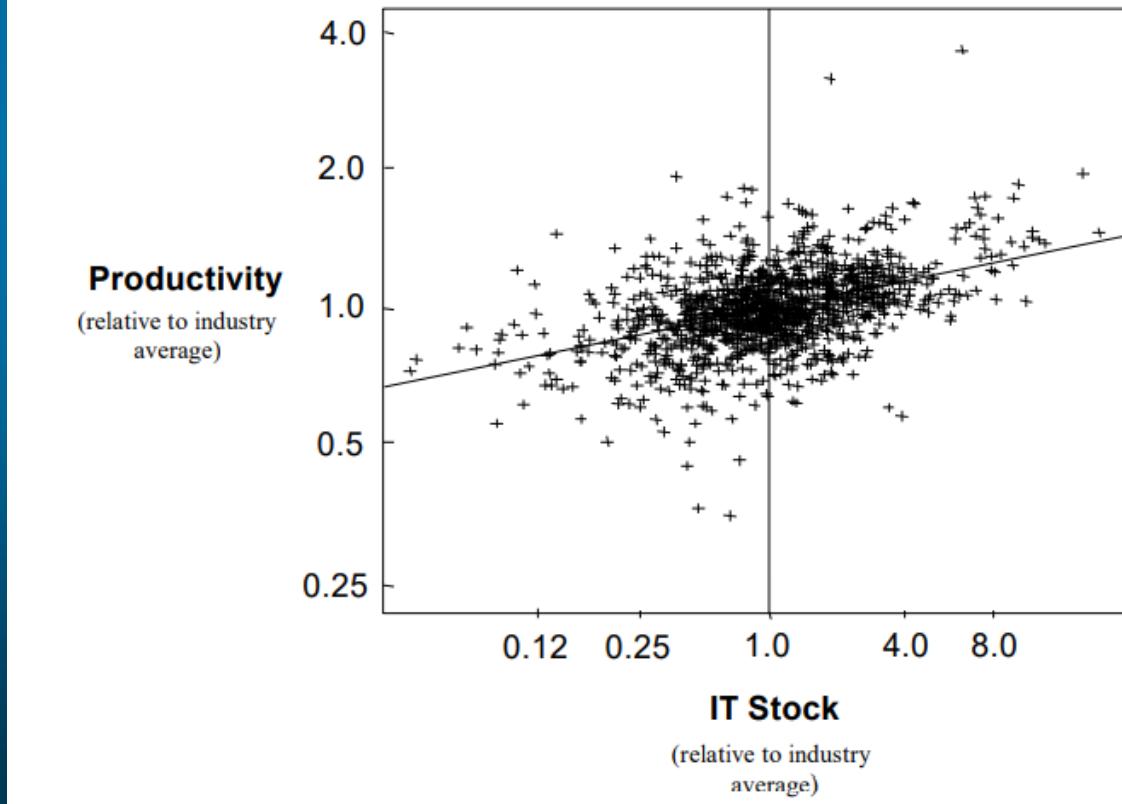
Lorin M. Hitt

University of Pennsylvania

Scatter plot

Figure 2: Variation in productivity and IT investment across firms

Caption: The vertical axis (labeled "Productivity") is multifactor productivity, defined as output divided by a weighted sum of inputs (in constant 1990 dollars). The horizontal axis (labeled "IT Stock") represents the total IT inputs in a firm. Both productivity and IT input are centered at the industry average. Note that some of the variation in IT Stock is due to differences in firm size. The points represent an individual firm in a particular year. There are approximately 1300 data points in this graph.



Descriptive Statistics



<http://pubsonline.informs.org/journal/mksc/>

MARKETING SCIENCE

Vol. 37, No. 1, January–February 2018, pp. 5–21

ISSN 0732-2399 (print), ISSN 1526-548X (online)

Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery

Hannes Datta,^a George Knox,^a Bart J. Bronnenberg^{a,b}

^a Tilburg University, 5000 LE Tilburg, Netherlands; ^b Centre for Economic Policy Research, London EC1V 0DX, United Kingdom

Contact: h.datta@tilburguniversity.edu, <http://orcid.org/0000-0002-8723-6002> (HD); g.knox@tilburguniversity.edu (GK); bart.bronnenberg@tilburguniversity.edu (BJB)

Received: February 9, 2016

Revised: October 19, 2016; February 13, 2017

Accepted: March 15, 2017

Published Online in Articles in Advance:
September 11, 2017

<https://doi.org/10.1287/mksc.2017.1051>

Copyright: © 2017 The Author(s)

Abstract. Instead of purchasing individual content, streaming adopters rent access to libraries from which they can consume content at no additional cost. In this paper, we study how the adoption of music streaming affects listening behavior. Using a unique panel data set of individual consumers' listening histories across many digital music platforms, adoption of streaming leads to very large increases in the quantity and diversity of consumption in the first months after adoption. Although the effects attenuate over time, even after half a year, adopters play substantially more, and more diverse, music. Relative to music ownership, where experimentation is expensive, adoption of streaming increases new music discovery. While repeat listening to new music decreases, users' best discoveries have higher play rates. We discuss the implications for consumers and producers of music.

History: Avi Goldfarb served as the senior editor and Catherine Tucker served as associate editor for this article.

Open Access Statement: This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to download this work and share with others for any purpose, except commercially, if you distribute your contributions under the same license as the original, and you must attribute this work as "Marketing Science. Copyright © 2017 The Author(s). <https://doi.org/10.1287/mksc.2017.1051>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>."

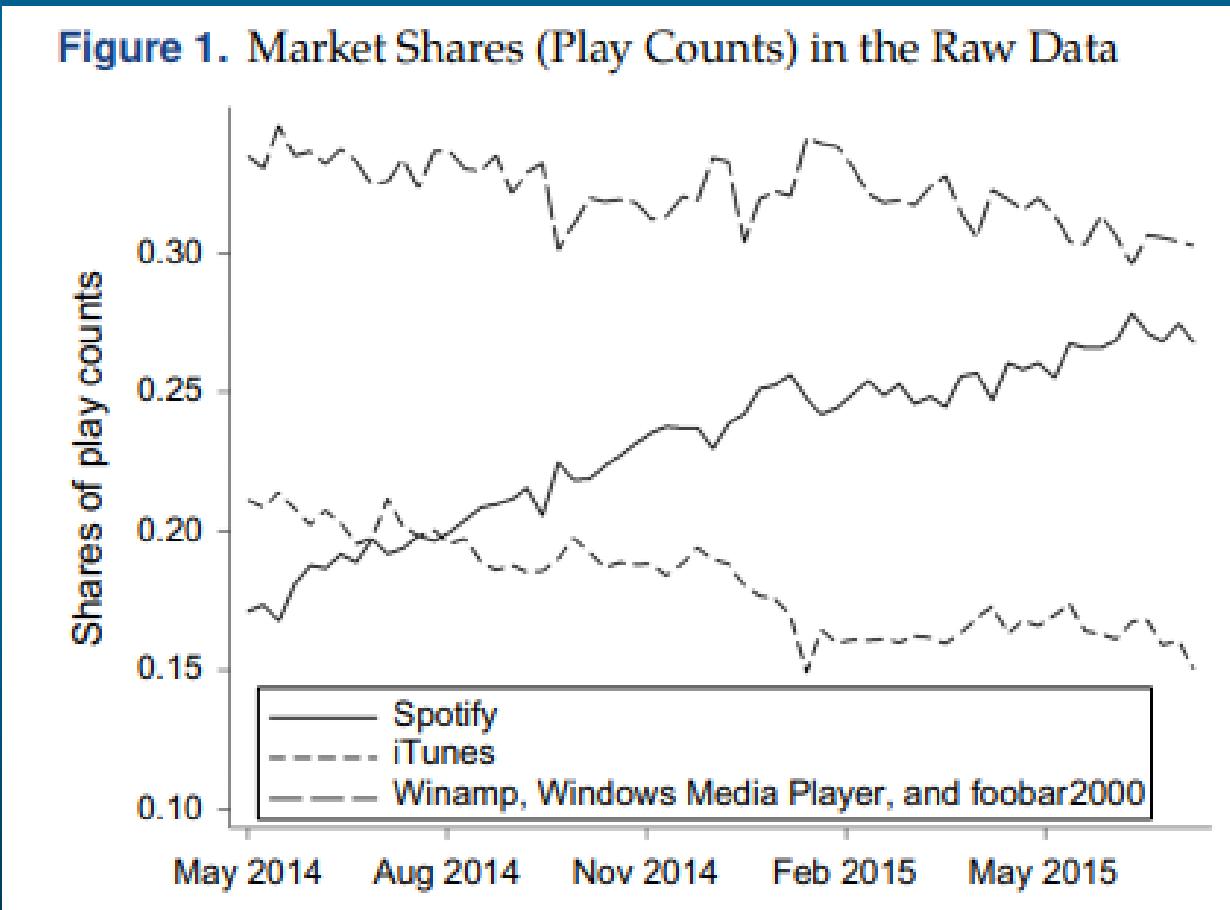
Funding: The authors acknowledge financial support from the Netherlands Foundation for Science [NWO 453-09-004] and from the Marketing Science Institute [MSI 4-1854].

Supplemental Material: Data and the online appendix are available at <https://doi.org/10.1287/mksc.2017.1051>.

Keywords: digital distribution • online streaming • entertainment industry • music consumption • variety

Descriptive Statistics

- Market shares for the major platforms in terms of play counts
- Usage of Spotify grew steadily while iTunes, other platforms declined



Correlation

Table 1. Variable Operationalization

Dimension	Operationalization
(1) <i>Quantity</i>	<ul style="list-style-type: none"> • Log number of song plays
(2) <i>Variety</i>	
<i>Breadth</i>	<ul style="list-style-type: none"> • Log number of unique artists, songs, and genres listened to
<i>Concentration in common favorites</i> (superstars)	<ul style="list-style-type: none"> • Number of unique artists in the top 20, top 100, and top 500 in a user's geographic region^a (ranked according to a rolling window of a year, lagged by four weeks; $t - 55, \dots, t - 4$), divided by the number of unique artists listened to over the same time period
<i>Concentration in personal favorites</i>	<ul style="list-style-type: none"> • Herfindahl index (sum of squared listening shares), computed over a user's weekly plays of artists, songs, and genres
(3) <i>Discovery</i>	
<i>New content consumption</i>	<ul style="list-style-type: none"> • Number of distinct new artists, songs, and genres listened to by a user for the first time,^b divided by the total number of distinct artists, songs, and genres listened to
<i>Repeat consumption</i>	<ul style="list-style-type: none"> • Number of unique new artists, songs, and genres played more than once, divided by the total number of unique new artists, songs, and genres listened to
<i>Best discoveries</i>	<ul style="list-style-type: none"> • Amount of plays of the top 1 new artist, song, and genre in an 8-week period subsequent to discovery ($t + 1, \dots, t + 8$) ranked in order of plays, divided by the amount of plays of the overall (not necessarily new) top 1 artist, song, and genre over the same time period

Note. All variables are computed at the user-week level.

Correla

Table 2. Summary Statistics

	N	Mean	SD	Min.	Max.
User characteristics					
<i>Gender (female = 1)</i>	1,978	0.24	0.43	0.00	1.00
<i>Age</i>	1,978	23.70	6.23	11.00	70.00
<i>European Union (dummy)</i>	1,978	0.32	0.47	0.00	1.00
<i>South America (dummy)</i>	1,978	0.21	0.41	0.00	1.00
<i>United States/Canada (dummy)</i>	1,978	0.10	0.30	0.00	1.00
<i>Other geographic region (dummy)</i>	1,978	0.37	0.48	0.00	1.00
Quantity of consumption					
<i>Play counts on all platforms</i>	122,636	188.89	269.43	0.00	3,862.00
<i>Play counts on Spotify</i>	122,636	11.19	69.38	0.00	2,439.00
<i>Play counts on iTunes</i>	122,636	48.43	148.59	0.00	3,857.00
<i>Play counts on Winamp, Windows Media Player, and foobar2000</i>	122,636	94.32	211.80	0.00	2,945.00
<i>Play counts on other platforms</i>	122,636	34.96	110.17	0.00	2,432.00
Breadth of variety					
<i>Number of unique artists</i>	97,924	36.64	50.92	1.00	882.00
<i>Number of unique songs</i>	97,924	150.44	169.79	1.00	2,603.00
<i>Number of unique genres</i>	97,924	14.00	13.49	1.00	209.00
Concentration of variety					
<i>Top 20 artists (share of unique artists)</i>	97,924	0.04	0.09	0.00	1.00
<i>Top 100 artists (share of unique artists)</i>	97,924	0.12	0.17	0.00	1.00
<i>Top 500 artists (share of unique artists)</i>	97,924	0.29	0.25	0.00	1.00
<i>Artist concentration (Herf.)</i>	97,924	0.21	0.23	0.00	1.00
<i>Song concentration (Herf.)</i>	97,924	0.05	0.11	0.00	1.00
<i>Genre concentration (Herf.)</i>	97,924	0.35	0.24	0.03	1.00
Discovery of new content					
<i>New artists (share of unique artists)</i>	97,924	0.20	0.22	0.00	1.00
<i>New songs (share of unique songs)</i>	97,924	0.37	0.27	0.00	1.00
<i>New genres (share of unique genres)</i>	97,924	0.05	0.09	0.00	1.00
<i>New artists played more than once</i> (share of unique new artists)	75,116	0.59	0.37	0.00	1.00
<i>New songs played more than once</i> (share of unique new songs)	91,579	0.22	0.25	0.00	1.00
<i>New genres played more than once</i> (share of unique new genres)	33,617	0.56	0.45	0.00	1.00
<i>Top 1 new artist to overall top 1 artist</i> (share of plays)	80,061	0.22	0.33	0.00	1.00
<i>Top 1 new song to overall top 1 song</i> (share of plays)	77,239	0.55	0.41	0.00	1.00
<i>Top 1 new genre to overall top 1 genre</i> (share of plays)	83,445	0.01	0.07	0.00	1.00

기술통계: 수치척도

- 위치, 산포도, 형태, 연관성에 관한 수치적 척도 소개
- 위치척도: 평균, 중앙값, 최빈값, 백분위수, 사분위수
- 변동성척도: 범위, 사분위 범위, 분산, 표준편차, 변동성 계수
- 분포형태, 상대적 위치, 극단값의 척도
- 두 변수간 관련성 측정: 공분산, 상관계수
- 가중평균과 그룹화 자료

- 표본통계량: 표본으로부터 추출된 자료로 부터 계산된 수치척도
- 모집단 모수: 모집단 자료로부터 도출된 척도

기술 통계량

- 자료의 중심위치 (대표값): 평균, 중앙값, 최빈값
- 자료의 상대적 위치: 사분위수, 백분위수
- 자료의 변동성 척도: 분산, 표준편차, 범위, 사분위 범위
- 자료의 분포 형태: 왜도 (분포 대칭성)
첨도 (분포 뾰족한 정도/꼬리 부분의 두터움 정도)

- 자료의 중심위치(대표값): 자료의 중심위치 혹은 자료를 대표하는 값
- 자료의 모든 값을 잘 대표할 수 있어야 함
- 거리를 정하는 측도의 정의에 따라 다른 대표값이 정의:
 - 평균: 편차의 제곱의 합 최소화
$$\min_a \sum (X_i - a)^2$$
 - 중앙값: 편차의 절대값의 합 최소화
$$\min_a \sum |X_i - a|$$

표본 평균

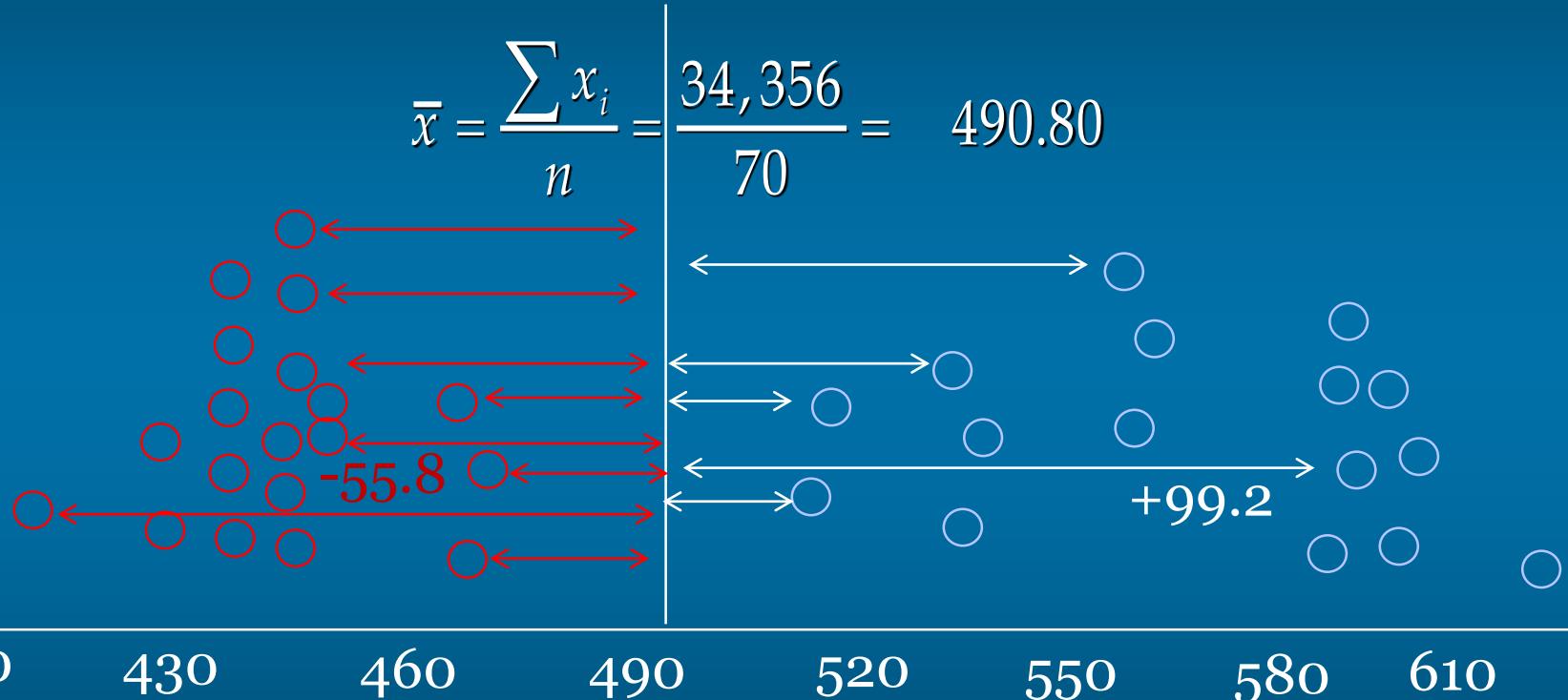
$$\bar{x} = \frac{\sum x_i}{n} = \frac{34,356}{70} = 490.80$$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

표본평균

- 거리의 개념으로 차이의 제곱의 합을 최소화 → 평균
- 자료의 편차 [각 관측치와 평균 차이]의 합은 항상 0이 됩니다.

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$



- 단점: 이상치 (outlier)가 존재하는 경우 평균은 이상치의 영향을 받음.
- 예: 관측자료: 3,3,3,3 → 평균=3
- 관측자료: 3,3,3,3, 100 → 평균=22.4

중앙값



- 자료를 크기 순서대로 나열했을 때 중간순위의 값.
- 중앙값의 장점: 극단치, 이상치(outlier)의 영향을 적게 받음.
- 계산:
 - 자료의 개수 n이 짝수: 중앙에 위치한 값
 - 자료의 개수 n이 짝수: $n/2$ 와 $(n/2)+1$ 순위의 평균.
예: 관측치 개수가 70개인 경우, 35, 36번째 값의 평균
$$\text{중앙값} = (475 + 475)/2 = 475$$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

노트: 데이터는 오름차순 정렬되어 있음

최빈값 (Mode)

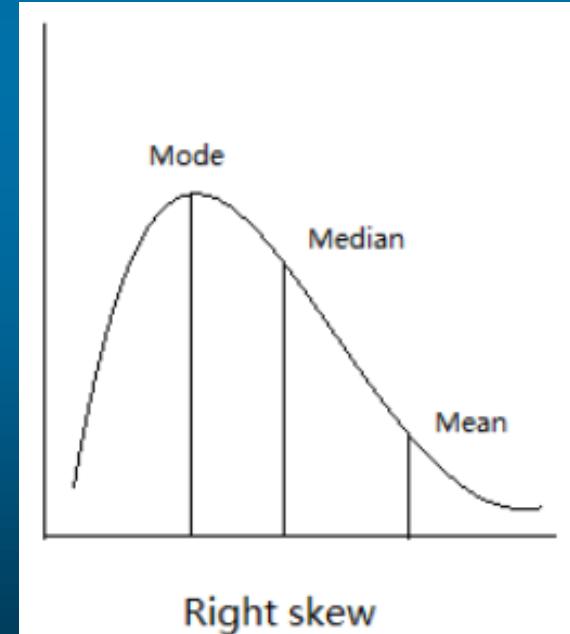
- 자료 중 관측 빈도가 가장 많은 값
- 장점: 자료에 이상치(outlier)가 존재해도 영향을 덜 받음
 - 예: 관측자료: 3,3,3,3 → 평균=3, 최빈값=3
 - 관측자료: 3,3,3,3, 100 → 평균=22.4, 최빈값=3
 - 평균과 중앙값은 양적 자료에만 사용할 수 있는 반면에, 최빈값은 질적 자료나 양적 자료 모두에 사용할 수 있음.
 - 평균과 중앙값은 단 하나의 값으로 계산되는 반면에, 최빈값은 없거나 하나 이상의 값이 있을 수 있음.

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

450이 가장 많다 (7 번) → 최빈값 = 450

분포 모양에 따른 평균, 중앙값, 최빈값

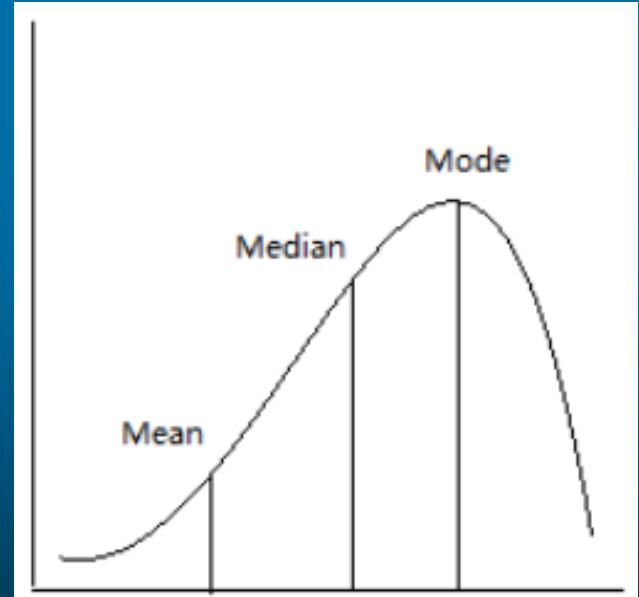
- 분포가 좌우 대칭: 평균=중앙값=최빈값
- 분포가 오른쪽으로 치우쳤을 때 (skewed to the right):
 - 오른쪽 꼬리가 더 길다는 것은, 대부분의 자료보다 큰 값을 갖는 이상치들이 존재한다는 것을 의미함.
 - 평균은 이상치들에 치우쳐서 크게 계산되는 반면, 중위수나 최빈값은 이상치들의 영향을 덜 받음.
 - 중위수, 최빈값 < 평균
- 봉우리가 왼편에 위치하므로 가장 관측빈도가 높은 최빈값은 왼편에 위치
 - 중간순위의 값인 중앙값은 최빈값보다 상대적으로 오른쪽에 위치함
 - 최빈값 < 중앙값
 - 오른쪽으로 치우친 분포의 경우, 최빈값<중앙값<평균



분포 모양에 따른 평균, 중앙값, 최빈값

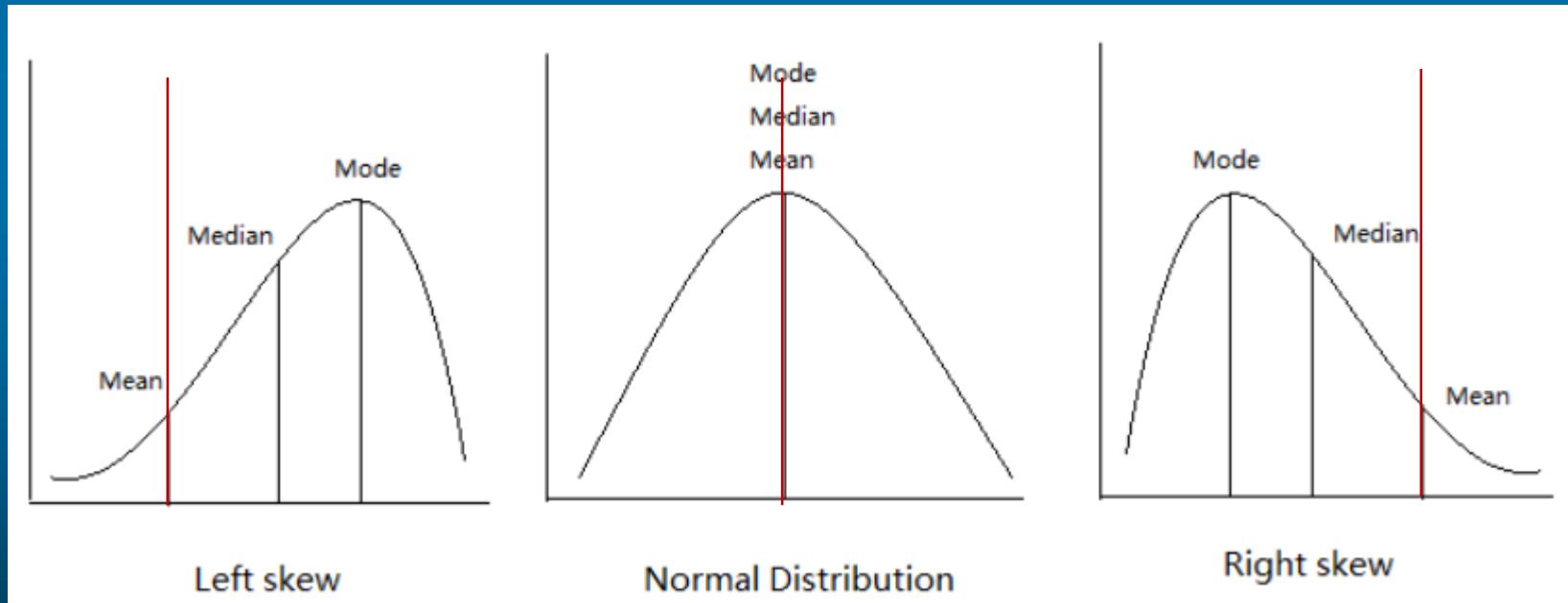
- 분포가 왼쪽으로 치우쳤을 때 (skewed to the left):
 - 왼쪽 꼬리가 더 길다는 것은, 대부분의 자료보다 작은 값을 갖는 이상치들이 분포 왼쪽에 존재한다는 것을 의미함.
 - 평균은 이상치들에 치우쳐서 작게 계산되는 반면, 중위수나 최빈값은 이상치들의 영향을 덜 받음.
 - 평균 < 중위수, 최빈값

- 봉우리가 오른편에 위치하므로 가장 관측빈도가 높은 최빈값은 오른편에 위치
 - 중간순위의 값인 중앙값은 최빈값보다 상대적으로 왼쪽에 위치함
 - 중앙값 < 최빈값
 - 왼쪽으로 치우친 분포의 경우,
평균 < 중앙값 < 최빈값



분포 모양에 따른 평균, 중앙값, 최빈값

- Skewed to the left: Mode > Median > Mean
- Skewed to the right: Mean > Median > Mode



백분위수

- p 백분위수는 적어도 관찰값의 p 퍼센트가 그 값과 같거나 작은 값이다. 적어도 관찰값의 $(100-p)$ 퍼센트는 p 백분위수와 같거나 더 큰 값을 가진다.

▶ 자료를 크기 순서대로 배열한다(오름차순).

지표 (index) i , p 백분위수의 위치를 계산한다.

▶ 위치지표 i
$$i = (p/100)n$$

▶ 만약 i 가 정수가 아니라면, 올림한다. p 백분위수는 i 번째 위치한 값이다.

▶ 만약 i 정수라면, p 백분위수는 i 와 $i+1$ 번째 위치한 값의 평균이다.

80th 백분위수



$$i = (p/100)n = (80/100)70 = 56$$

위치지표 i: 정수인 경우 56, 57번째 값의 평균:

$$80 \text{ 백분위수} = (535 + 549)/2 = 542$$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

노트: 데이터는 오름차순으로 정렬

80th 백분위수



▶ “적어도 관찰값의
80%가
542 보다 작거나
같다.”

▶ “적어도 관찰값의
20%가
542 보다 크거나
같다.”

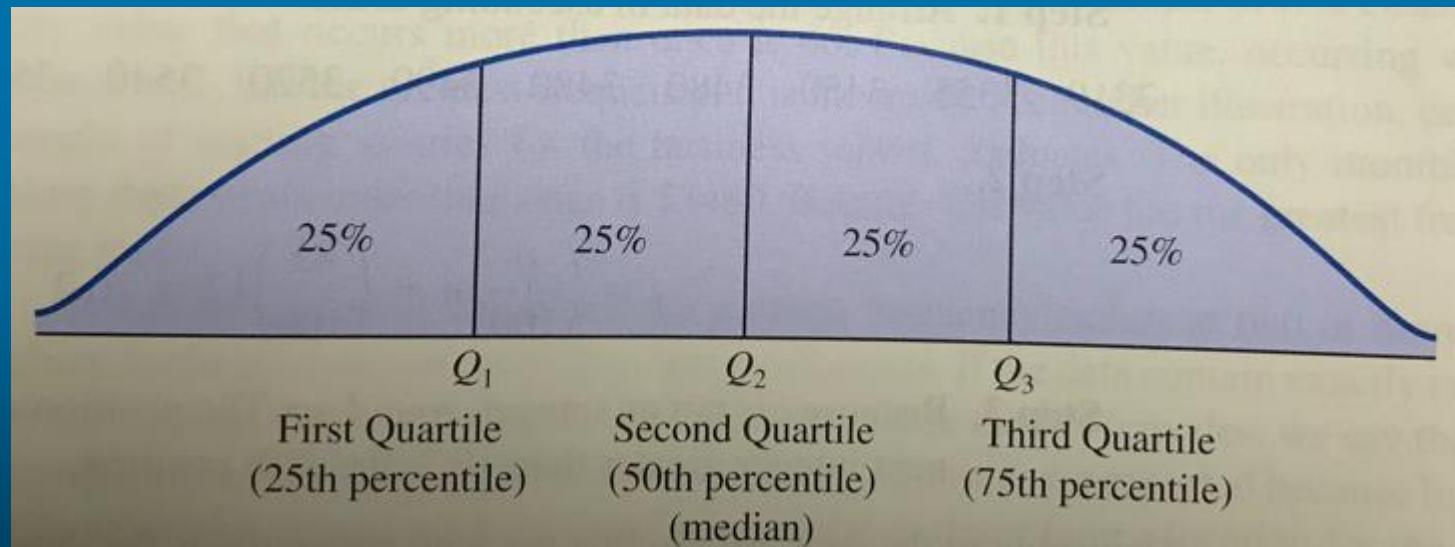
▶ $56/70 = .8$ 또는 80%

▶ $14/70 = .2$ 또는 20%

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

사분위수(quartiles)

- ▶ 사분위수는 특정 백분위수이다.
- ▶ 1사분위수 = 25th 백분위수
- ▶ 2사분위수 = 50th 백분위수 = 중앙값
- ▶ 3사분위수 = 75th 백분위수



3 사분위수



3사분위수 = 75th 백분위수

$$i = (p/100)n = (75/100)70 = 52.5 = 53$$

3사분위수 = 525

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

노트: 데이터는 오름차순으로 정렬

사분위수 범위(IQR:interquartile range)

- ▶ 사분위수 범위는 3사분위수와 1사분위수의 차이이다.

$$IQR = Q_3 - Q_1$$

- ▶ 사분위수 범위는 자료의 중앙 50%의 범위를 의미한다.
- ▶ 이는 극단값의 영향을 줄일 수 있다.

사분위수 범위



3사분위수 ($Q3$) = 525

1사분위수 ($Q1$) = 445

사분위수 범위 = $Q3 - Q1 = 525 - 445 = 80$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

노트: 데이터는 오름차순으로 정렬

용어: 적률 (Moment)

- Moments are a set of statistical parameters to measure a distribution. 분포를 측정하는 통계적 파라미터
- 정의: $E(X^i)$, $i=1,2,3, \dots$ → i번째 적률 (the i-th moment)

- 대표적인 4가지 많이 사용되는 적률 척도:

- (1) 1차 적률: $E(X)$ 대표값에 관한 척도

$$\mu = E(X)$$

- (2) 2차 적률: $E(X^2)$ 분산에 관한 척도

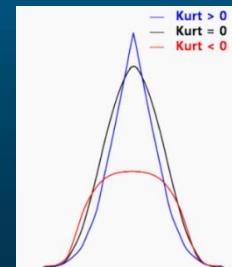
$$\sigma^2 = E(X - E(X))^2$$

- (3) 3차 적률: $E(X^3)$ 왜도에 관한 척도

$$\mu_3 = \frac{E(X - \mu)^3}{\sigma^3}$$

- (4) 4차 적률: $E(X^4)$ 첨도에 관한 정보

$$\mu_4 = \frac{E(X - \mu)^4}{\sigma^4}$$



분포 형태, 상대적 위치, 극단값

- 분포 형태
- z-값
- 체비셰프의 원리
- 경험법칙
- 극단값 찾기

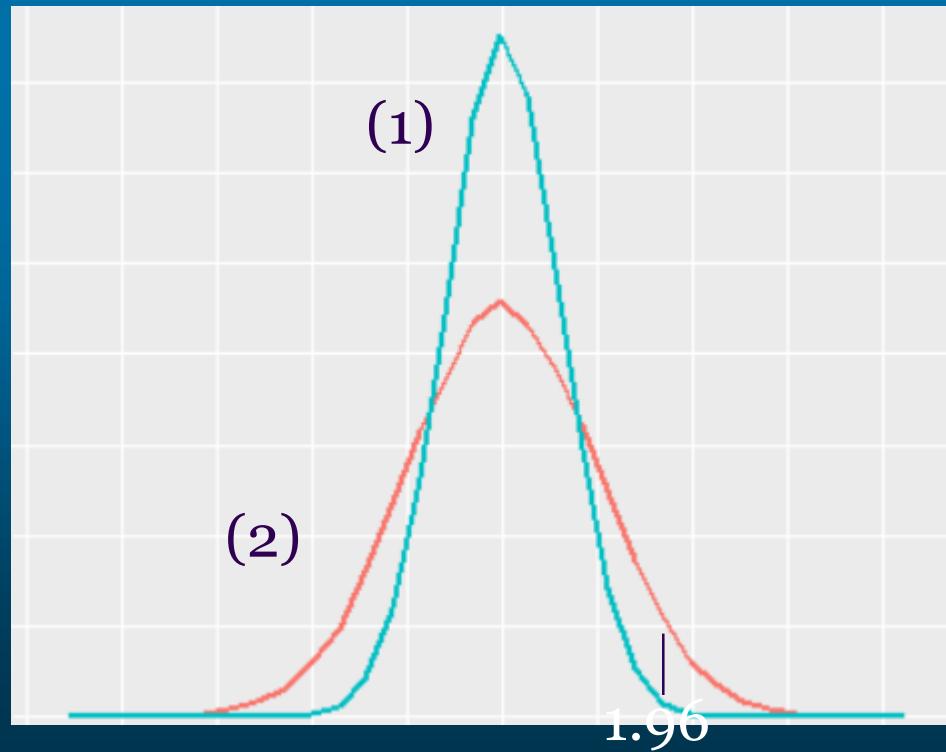
Z -값

- 관찰값의 Z 값: 자료에서 관찰값의 상대위치를 측정하는 척도
- 배경: 단위가 다르거나 분산이 다른 변수의 분포는 서로 비교할 수 없다.

1. 단위 차이: 석유 가격=50 달러/배럴, 금가격=1,100 달러/온스

→ 단위가 다르기 때문에 국제시장에서 한국의 상품 상대 가격이 어느쪽이 더 높은지 알기 어려움 → 단위제거 (평균 차감)

2. 분산 차이: 상대위치척도로서 값이 1.96인 경우, 단위가 같아도 분산이 다르면, 같은 값이라도 분포에서 갖는 상대적 위치가 다르므로 비교 불가



Z-값: 표준화된 값

표준화: (자료값-평균)/표준편차



정의: 자료 x_i 의 Z-값 \rightarrow

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

(자료값-표본 평균)
(표본 표준편차)



특성: (1) 단위가 없다.
(2) 평균은 0이고, 분산은 1이다.



Z-값을 종종 ‘표준화(된) 값’이라고 한다.

z-값: 표준화된 값

- 의미: 평균을 중심으로 자료의 상대적 크기(위치) 표시
 - 만약, z-값이 2라면, 자료 X_i 는 표본 평균보다 표준편차의 2배만큼 크다.

$$z_i = \frac{X_i - \bar{X}}{s} = 2 \quad \rightarrow \quad X_i = \bar{X} + 2s$$

- 두 개의 다른 변수가 같은 z-값을 가졌다면, 각각 평균으로부터 같은 크기의 표준편차 배수만큼 떨어져 있다는 점에서 같은 상대적 크기를 가진다.
(예) Y: 경영반의 1번 학생 몸무게(kg)
X: 경제반의 5번 학생 키(cm)
 $\rightarrow z_Y = \frac{Y - \bar{Y}}{s_Y}, \quad z_X = \frac{X - \bar{X}}{s_X}$
 \rightarrow 만약, $z_Y = z_X$ 이면, Y와 X는 같은 상대적 크기를 가진다.

체비세프의 정리(부등식)

- 의미: 평균값을 중심으로 특정 표준편차 크기의 범위 안에 있는 값들의 최소 비율

평균 $\pm z\sigma$ 표준편차 크기의 범위 안에 있는 관측치들의 비율은 적어도 $(1 - 1/z^2)$ 이다. 단, $z > 1$ 임.

$$P\{|X-\mu| \leq z\sigma\} \geq 1 - 1/z^2$$

- 다음 범위에 있을 관측치 비율을 말해줌:
- 평균 $- z\sigma$ 표준편차 \leq [관측치 비율] \leq 평균 $+ z\sigma$ 표준편차

- 체비세프 정리를 이용할 때 z 값 계산.
- 예) 전체 관측치 중 평균에서 2 표준편차 범위 안에 있는 관측치들의 비율
- $\rightarrow z=2$ 이므로, 관측치 비율: $1 - 1/2^2 = 0.75$

체비셰프의 정리

- 예) 100명의 경영경제통계학 학생의 중간시험 점수의 평균이 70점, 표준편차가 5라고 할 때, 몇 명의 학생이 60점과 80점 사이에 있는가?

$$60\text{점}: z_{60} = \frac{60-70}{5} = -2$$

$$80\text{점}: z_{80} = \frac{80-70}{5} = 2$$

→ 결국, “평균 $\pm 2 \times$ 표준편차” 범위에 있을 학생들의 비율

→ $z=2 \rightarrow$ 관측치 비율: $1 - 1/2^2 = 0.75$

→ 적어도 75% 이상의 학생들이 60점과 80점 사이에 있을 것임.

체비셰프의 원리(Chebyshev's theorem)

- ▶ 적어도 자료값들의 **75%** 는 평균에서
 $z = 2$ 표준편차 범위 안에 있어야 한다.
- ▶ 적어도 자료값들의 **89%** 는 평균에서
 $z = 3$ 표준편차 범위 안에 있어야 한다.
- ▶ 적어도 자료값들의 **94%** 는 평균에서
 $z = 4$ 표준편차 범위 안에 있어야 한다.

경험 법칙(Empirical rule)

자료가 종모양의 분포를 가지고 있는 경우,

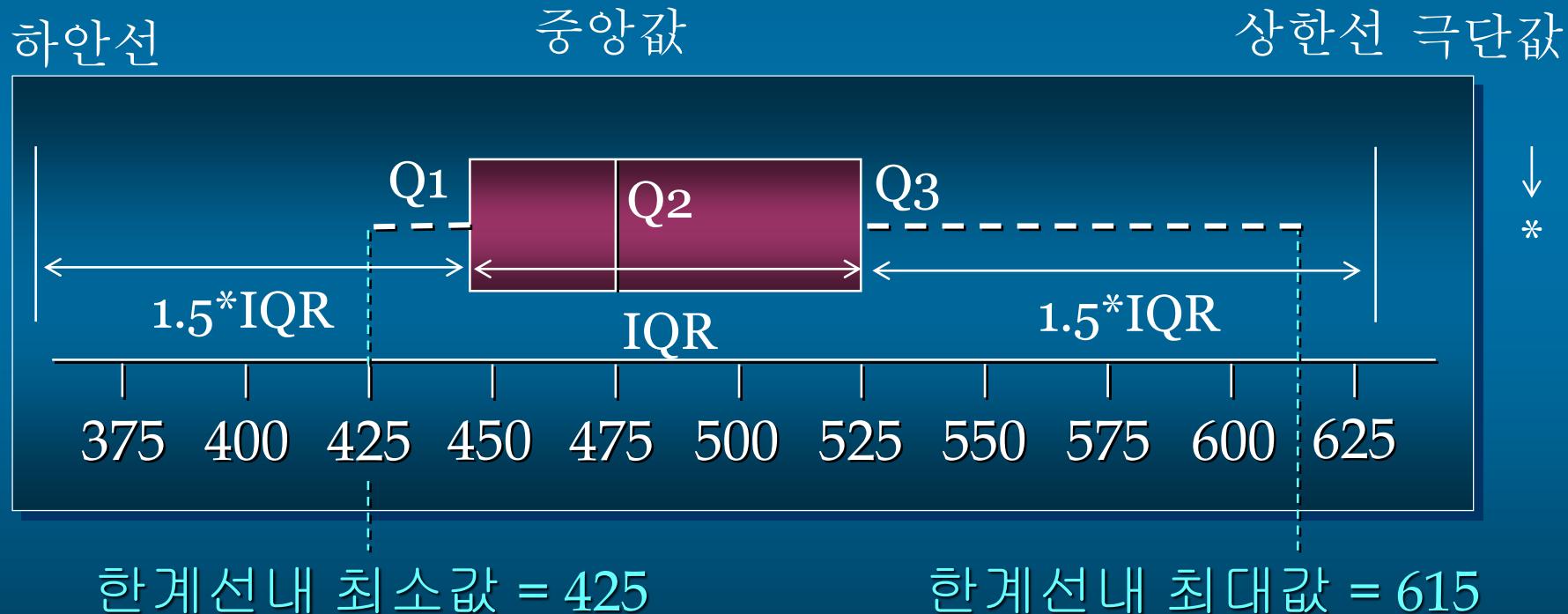
- ▶ 경험법칙은 평균의 특정 표준편차 내에 있어야 하는 자료값들의 비율을 결정하는데 사용될 수 있다.

- ▶ 경험법칙은 제6장에서 다루어 질 정규분포에 기반한다.

분포탐색: 상자 그림

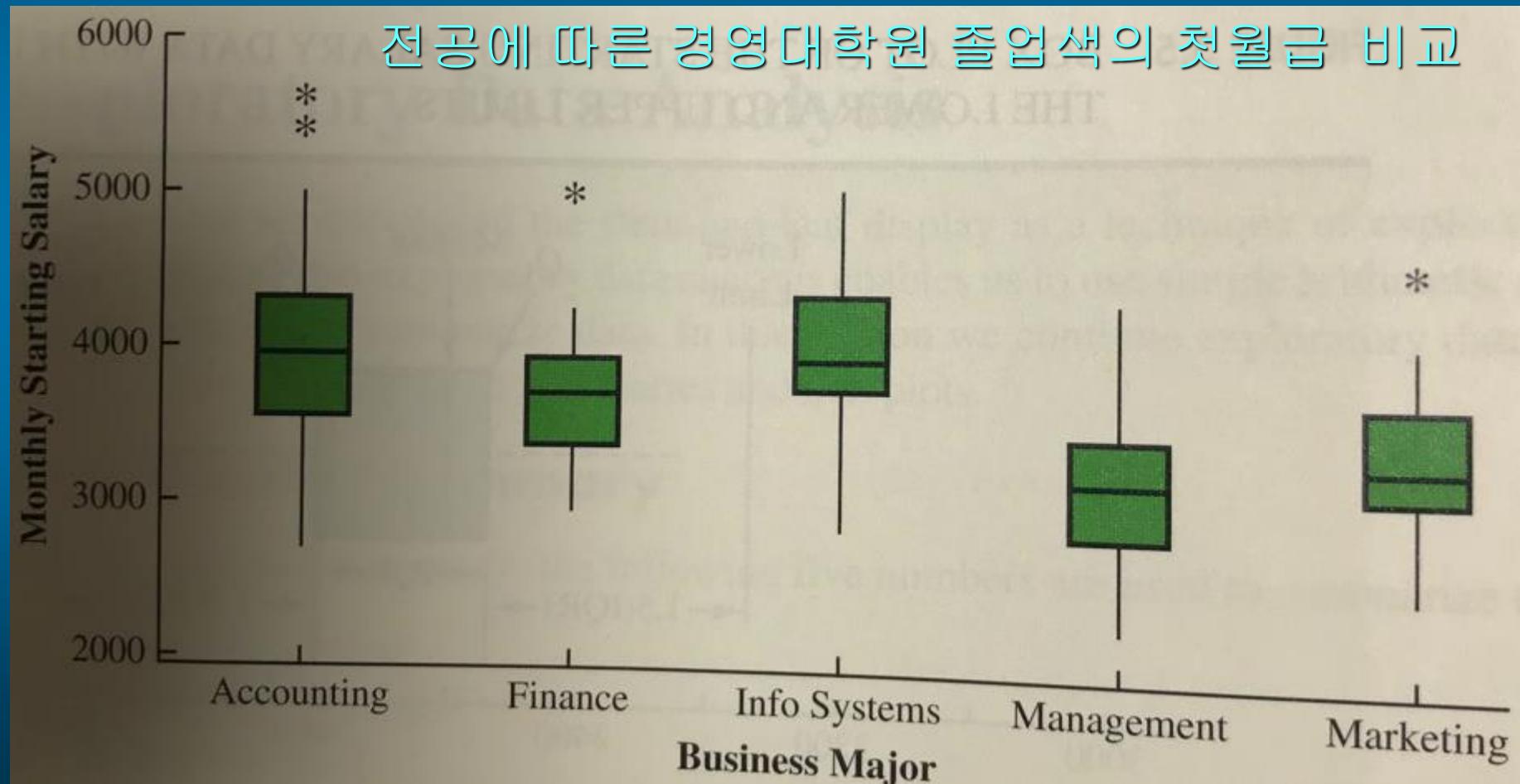


- ▶ 다섯수치 요약: 최소값, Q1, 중앙값(Q2), Q3, 최대값
상자의 양끝에서 한계선내에 최소값과 최대값까지
점선(whiskers)을 그린다.



상자 그림

- 중앙값 비교: 회계, 경영정보시스템
- 자료 중앙의 50% 분포 높이에 따라 편차 비교: 재무(소)
- 극단치 존재여부: 회계, 재무, 마케팅



두 변수간의 연관성 측정 (Measures of association between two variables)

- ▶ 지금까지 하나의 변수에 대한 자료를 요약하기 위한 수치적 방법들을 살펴보았다.
- ▶ 경영자나 의사결정자는 종종 두 변수의 관계에 관심을 가진다.
- ▶ 두 변수의 관계에 관한 기술 측정치로 공분산과 상관계수가 있다.

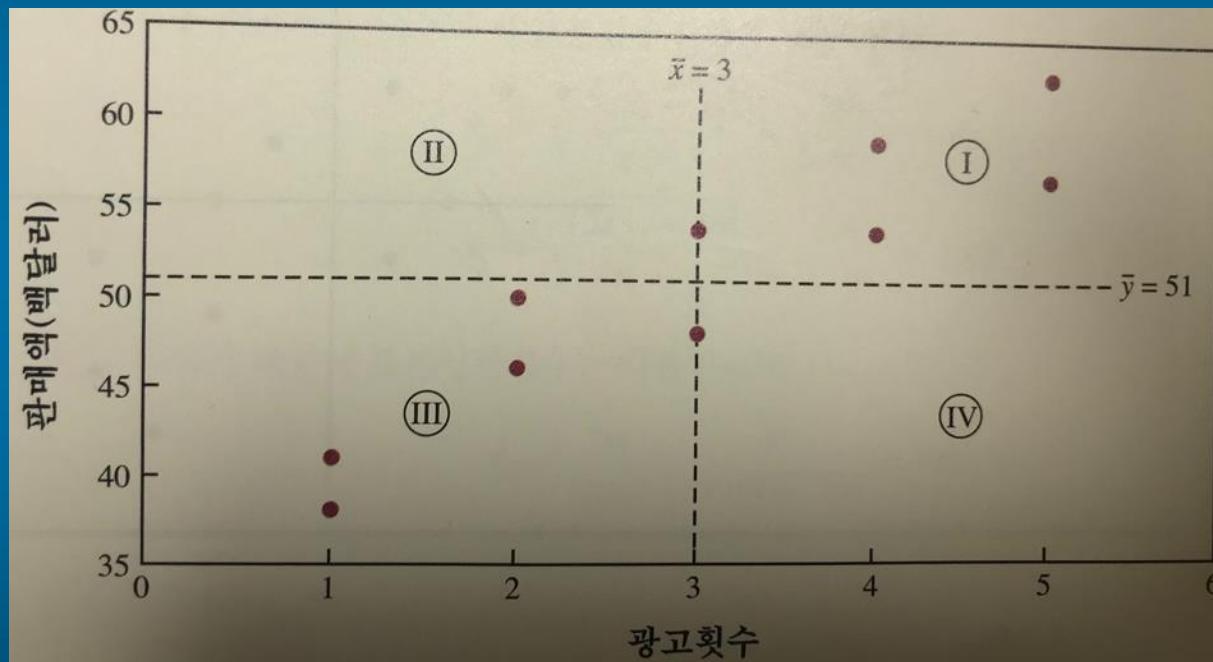
공분산(covariance)

- ▶ 공분산은 두 변수의 선형관계를 측정하는 척도이다.
- ▶ 양의 값은 양의 관계를 나타낸다.
- ▶ 음의 값은 음의 관계를 나타낸다.

공분산(covariance)

$(x_i - \bar{x})(y_i - \bar{y})$ 은 제 1,3 사분면에서 (+) 값을 가지고, 제2,4사분면에서 (-)값.

- 만약 s_{xy} 가 (+) 값이라면 s_{xy} 에 영향을 미치는 점들이 1,3사분면 위치
- 즉 x와 y가 양(+)의 선형관련성을 가진다는 것을 의미: x값 증가시 y값 증가

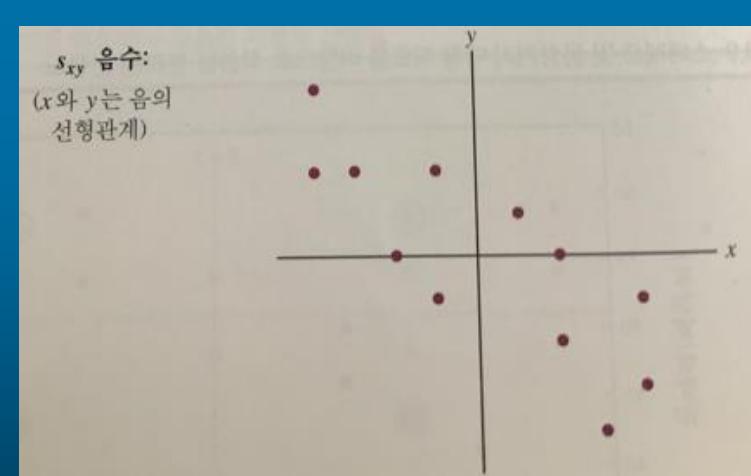
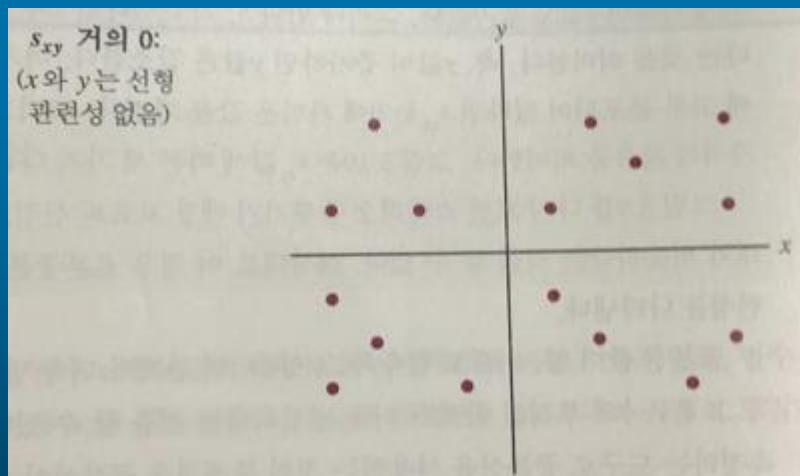


$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \text{ 표본}$$

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N} \text{ 모집단}$$

공분산(covariance)

- 선형연관성의 정도를 측정하는 도구로서 공분산의 문제점:
- 공분산이 x 와 y 를 측정하는 단위에 따라 결과가 달라질 수 있음.
예) 키(x)와 몸무게(y)의 관계를 알고 싶을 경우:
키를 cm로 측정하는 것은 m로 측정했을 때보다, 측정단위로 인하여,
 $(x_i - \bar{x})$ 값이 더 큰 수치를 가지게 되고, 공분산의 값도 커지게 됨(
- 측정단위에 영향을 받지 않으면서 두 변수의 관계를 측정하는 방법
→ 상관계수(correlation coefficient)



상관계수(correlation coefficient)

(피어슨)상관계수는 아래와 같이 계산된다:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

표본의 경우

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

모집단의 경우

(공분산)

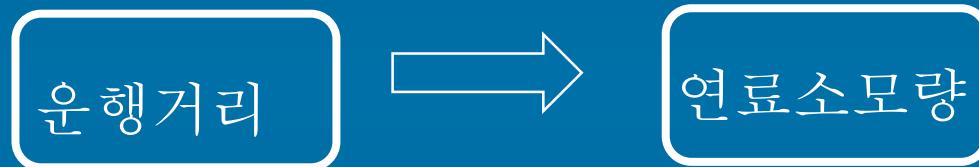
(x표준편차)*(y표준편차)

- ▶ 상관계수는 -1에서 +1사이의 값을 갖는다.
- ▶ 0에 가까워질수록, x와 y사이에 선형관계가 약해짐의 미
- ▶ 1값에 가까울수록 강한 선형관계를 나타낸다.

상관계수와 인과관계

상관관계는 변수들 간의 선형관계를 측정하는 것이지 반드시 인과관계를 측정하는 것은 아니다.

예1: 자동차의 경우, 운행거리와 연료소모량 간에는 양의 상관관계가 있으며 또한 운행거리 → 연료소모량의 인과관계가 존재함.



예2: 차량 배기량과 주택평수의 경우, 보유차량의 배기량과 보유아파트 평수 간에는 양의 상관관계가 존재하지만, 양자 간에는 인과관계가 존재하지 않는다. 이 둘 변수의 각각을 인과하는 것은 제3의 변수인 소득이라고 볼 수 있음.



가중평균과 그룹화 자료 (The weighted mean and working with grouped data)

- 가중평균
- 그룹화 자료의 평균
- 그룹화 자료의 분산
- 그룹화 자료의 표준편차

가중평균(weighted mean)

- ▶ 관찰값의 중요도를 반영한 가중치를 각각의 자료값에 부여하여 평균을 계산할 때, 이러한 평균을 ‘가중 평균’이라고 한다.

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

여기서:

x_i = i 번째 관찰값

w_i = i 번째 관찰값의 가중치

- ▶ 관찰값이 중요도에 있어서 서로 다를 때, 분석자는 각 관찰값이 가지는 중요도를 가장 잘 반영할 수 있는 가중치를 선택하여야 한다.

예) 어느 금융자산 포트폴리오가 50%는 주식, 40%는 채권, 10%는 현금으로 구성되어 있다. 주식 수익률은 12%이며 채권 수익률은 7%고 하자. 포트폴리오의 수익률은?

$$R_w = 0.5 \times 0.12 + 0.4 \times 0.07 + 0.1 \times 0 = 0.088 \quad (8.8\%) \quad \text{Slide 43}$$

그룹화된 자료(grouped data)

- ▶ 가중평균 계산법이 그룹화된 자료의 평균, 분산, 표준편차의 대략적인 값을 구하는데 사용된다.
- ▶ 가중평균을 계산하기 위해, 각 계급의 **중간점**을 그 계급의 평균처럼 가정하여 사용한다.
- ▶ 계급의 도수를 가중치로 사용하여 계급 중간점들의 가중평균을 계산한다.
- ▶ 분산과 표준편차를 계산할 때도 유사한 방법으로 계급의 도수를 가중치로 사용한다.

그룹화 자료의 평균

표본평균

$$\bar{x} = \frac{\sum f_i M_i}{n}$$

모집단 평균

$$\mu = \frac{\sum f_i M_i}{N}$$

여기서:

f_i = i 계급의 (빈)도수

M_i = i 계급의 중간점

그룹화 자료의 표본평균



앞선 예에서 본 70채의 아파트 표본 월세자료가 아래와 같이 도수분포 형식으로 그룹화되어 있다.

Rent (\$)	Frequency
420-439	8
440-459	17
460-479	12
480-499	8
500-519	7
520-539	4
540-559	2
560-579	4
580-599	2
600-619	6

그룹화 자료의 표본 평균



Rent (\$)	f_i	M_i	$f_i M_i$
420-439	8	429.5	3436.0
440-459	17	449.5	7641.5
460-479	12	469.5	5634.0
480-499	8	489.5	3916.0
500-519	7	509.5	3566.5
520-539	4	529.5	2118.0
540-559	2	549.5	1099.0
560-579	4	569.5	2278.0
580-599	2	589.5	1179.0
600-619	6	609.5	3657.0
Total	70		34525.0

$$\bar{x} = \frac{34,525}{70} = 493.21$$

이런 근사값은 실제 평균인 \$490.80과는 \$2.41 정도 차이가 있다.

그룹화된 자료의 분산

▶ 표본의 경우

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n-1}$$

▶ 모집단의 경우

$$\sigma^2 = \frac{\sum f_i (M_i - \mu)^2}{N}$$

그룹화 자료에서 표본분산



Rent (\$)	f_i	M_i	$M_i - \bar{x}$	$(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
420-439	8	429.5	-63.7	4058.96	32471.71
440-459	17	449.5	-43.7	1910.56	32479.59
460-479	12	469.5	-23.7	562.16	6745.97
480-499	8	489.5	-3.7	13.76	110.11
500-519	7	509.5	16.3	265.36	1857.55
520-539	4	529.5	36.3	1316.96	5267.86
540-559	2	549.5	56.3	3168.56	6337.13
560-579	4	569.5	76.3	5820.16	23280.66
580-599	2	589.5	96.3	9271.76	18543.53
600-619	6	609.5	116.3	13523.36	81140.18
Total	70				208234.29

계속 →

그룹화 자료에서 표본 분산



▶ 표본 분산

$$s^2 = 208,234.29 / (70 - 1) = 3,017.89$$

▶ 표본 표준편차

$$s = \sqrt{3,017.89} = 54.94$$

이러한 근사값은 실제 표준편차인 \$54.74와는
겨우 \$.20 정도 차이가 난다.