KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

3. 데이터와 정보 II: 단변량 & 양변량 분포

정치와 데이터분석

박상훈 (sh.park.poli@gmail.com)
강원대학교

오늘의 목표

10:10-11:00

단변량 분포(univariate distribution)의 이해: 중심경향성과 산포도

11:10-12:05

양변량 분포(bivariate distribution)의 이해: 관계의 시각화와 수치적 요약

12:15-12:45

R을 이용한 시각화 실습 및 질의응답

데이터와 정보 II: 단변량 & 양변량 분포

Recap: 측정 수준(levels of measurement)

명목형(Nominal): 순서가 없이 서로 배타적인 범주(예: 종교)

순서형(Ordinal): 순서가 있는 범주(예: 교육 수준, 리커트 척도)

이항형(Binary): 0 혹은 1로 존재/없음만을 가짐(예: 성별)

등간형(Interval): 순서와 등간격을 가짐, 임의의 0점(예: 온도)

비율형(Ratio): 순서와 등간격을 가지며 동시에 절대 0점을 가짐(예: 소득, 나이)

변인의 측정 수준은 우리가 사용할 수 있는 통계 기법을 결정

데이터와 정보 II: 단변량 & 양변량 분포

왜 데이터를 시각화해야 하는가?

데이터 시각화(visualization)를 통해 **패턴**을 발견할 수 있음.

- 데이터에 숨겨진 추세, 비교, 대조, 이상치를 한눈에 파악할 수 있음.

데이터의 시각화는 숫자의 나열보다 강력한 이야기와 메시지를 전달함.

복잡한 정보를 직관적이고 이해하기 쉬운 형태로 변환하는 방법

데이터와 정보 II: 단변량 & 양변량 분포

기술통계(Descriptive Statistics)

수집한 데이터를 요약하고, 패턴을 설명하며, 정보를 간결하게 전달하는 데 사용되는 통계적 방법

- 데이터의 정보를 압축하는 것
- 전형적인(typical) 값은 무엇인가? **중심경향성**
- 오늘 좀 더 살펴볼 내용은 '값들이 얼마나 널리 퍼져 있는가?', **분산(variance)**과 **산포도(scatter plot)**의 내용

Part I. 단변량 분포(Univariate Distribution)

데이터와 정보 II: 단변량 & 양변량 분포

범주형 변인

관측치가 순위를 매길 수 없는 범주 중 하나에 속하는 변인(예: 종교, 혈액형)

- 이러한 범주형 변인을 파악하기 위해서는 '각 범주에 얼마나 많은 관측치가 있는가'를 확인하는 것이 필요
- 주요 도구
  - 빈도표 (Frequency Table)
  - 최빈값 (Mode)
  - 막대 그래프 (Bar Plot)

데이터와 정보 II: 단변량 & 양변량 분포

범주형 변인: 빈도표

R Code | Output

```r
# gapminder 데이터셋에서 2007년 아시아 대륙 국가들의 수를 셉니다.
library(gapminder); library(dplyr)

asia_2007 <- gapminder |>
  dplyr::filter(year == 2007, continent == "Asia")

# sub_region은 예시를 위해 임의로 생성한 변수입니다.
set.seed(123)
asia_2007$sub_region <-
  sample(c("East Asia", "Southeast Asia", "South Asia", "West Asia"),
         size = nrow(asia_2007), replace = TRUE)

# 지역(sub_region)별 국가 빈도표 생성
asia_2007 |> count(sub_region, sort = TRUE)
```

데이터와 정보 II: 단변량 & 양변량 분포

범주형 변인: 최빈값

최빈값은 데이터에서 가장 빈번하게 나타나는 값

범주형 데이터에서 사용할 수 있는 유일한 중심경향성 측도

- 위의 예제에서 최빈값은 "West Asia".
- 11개국으로 가장 많은 빈도

데이터와 정보 II: 단변량 & 양변량 분포

범주형 변인: 막대 그래프

각 범주의 빈도를 막대의 높이로 표현하여 시각적으로 비교할 수 있게 함.

R Code | Plot

```
# geom_bar()는 데이터의 빈도를 자동으로 계산하여 그려줍니다.
asia_2007 |> ggplot(aes(x = sub_region)) +
  geom_bar(fill = "steelblue") +
  labs(title = "2007년 아시아 국가들의 지역별 분포",
       x = "하위 지역",
       y = "국가 수")
```

데이터와 정보 II: 단변량 & 양변량 분포

범주형 변인: 막대 그래프 꾸미기

reorder() 함수를 사용해 막대를 크기 순으로 정렬하고, geom_text()로 막대 위에 값을 표시할 수 있음.

R Code | Plot

```
# 빈도 계산 후 시각화
asia_counts <- asia_2007 |>
  count(sub_region)

asia_counts |> ggplot(aes(x = reorder(sub_region, n), y = n)) +
  geom_col(fill = "skyblue") + # geom_col()은 계산된 값을 사용
  geom_text(aes(label = n), vjust = -0.3) +
  coord_flip() + # x축과 y축을 바꿈
  labs(title = "2007년 아시아 국가들의 지역별 분포 (정렬)",
       x = "하위 지역", y = "국가 수")
```

데이터와 정보 II: 단변량 & 양변량 분포

연속병 변인

연속형 변인은 값들 사이에 등간격이 존재하는 변인(예: 나이, 소득)

- 데이터의 중심은 어디인가? (Central Tendency)
- 데이터가 얼마나 퍼져 있는가? (Dispersion/Variation)
- 주요 도구
  - 중심경향성: 평균, 중앙값
  - 산포도: 범위, 사분위수, 분산, 표준편차
  - 시각화: 히스토그램, 밀도 플롯, 박스 플롯

데이터와 정보 II: 단변량 & 양변량 분포

중심경향성의 측정

데이터 분포의 **중심** 또는 **전형적인** 값을 나타내는 통계량

- 평균(Mean): 모든 값을 더해 값의 개수로 나눈 산술평균
  - 평균은 분포의 무게중심, 모집단 수준에서는 기대값(expected value)이라고도 불림.
  - 모든 데이터 값을 사용하는 특징, 이상치(outlier)라 불리는 극단적인 값에 매우 민감
- 중앙값(Median): 모든 값을 순서대로 나열했을 때 정중앙에 위치하는 값
  - 데이터를 크기순으로 정렬했을 때 중앙에 위치하는 값, 50번째 백분위수(50th percentile).
  - 이상치에 영향을 받지 않아 분포가 한쪽으로 치우쳐 있을 때(skewed), 평균보다 중심을 더 잘 나타낼 수 있음

데이터와 정보 II: 단변량 & 양변량 분포

중심경향성의 측정: 평균

R Code | Plot

```r
df <- data.frame(value = c(1, 2, 3, 4, 100))
mean_val <- mean(df$value)
ggplot(df, aes(x=value)) +
  geom_dotplot(binwidth = 1.5) +
  geom_vline(aes(xintercept=mean_val), color="red", linetype="dashed") +
  annotate("text", x=mean_val+10, y=0.8, label=paste("평균 =", mean_val), color="red") +
  labs(title="이상치(100)에 의해 평균이 오른쪽으로 이동") +
  theme(axis.text.y=element_blank(), axis.ticks.y=element_blank())
```

데이터와 정보 II: 단변량 & 양변량 분포

중심경향성의 측정: 평균

R Code | Plot

```r
df <- data.frame(value = c(1, 2, 3, 4, 100))
median_val <- median(df$value)
ggplot(df, aes(x=value)) + geom_dotplot(binwidth = 1.5) +
  geom_vline(aes(xintercept=median_val), color="blue", linetype="dashed") +
  annotate("text", x=median_val+15, y=0.8, label=paste("중앙값 =", median_val), color="blue")
  labs(title="이상치(100)에 영향을 받지 않는 중앙값") +
  theme(axis.text.y=element_blank(), axis.ticks.y=element_blank())
```

데이터와 정보 II: 단변량 & 양변량 분포

중심경향성의 측정: 평균 대 중앙값

R Code | Plot

```
gdp_2007 <- gapminder %>% dplyr::filter(year == 2007)
mean_gdp <- mean(gdp_2007$gdpPercap)
median_gdp <- median(gdp_2007$gdpPercap)

gdp_2007 |> ggplot(aes(x = gdpPercap)) +
  geom_histogram(bins = 30, fill = "lightblue", color = "black") +
  geom_vline(aes(xintercept = mean_gdp, color = "평균"), linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median_gdp, color = "중앙값"), linetype = "dotted", size = 1) +
  scale_color_manual(name = "통계량", values = c("평균" = "red", "중앙값" = "blue")) +
  labs(title = "2007년 1인당 GDP 분포", x = "1인당 GDP", y = "빈도") +
  annotate("text", x=mean_gdp+12000, y=20, label=paste("평균:", round(mean_gdp)), color="red"
  annotate("text", x=median_gdp+12000, y=15, label=paste("중앙값:", round(median_gdp)), color=
```

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정(Measures of Dispersion)

데이터가 중심경향성으로부터 얼마나 퍼져 있는지, 즉 변동성(variability)이 얼마나 큰지를 나타냄.

- 범위 (Range): 최대값 - 최소값
- 사분위수 범위 (Interquartile Range, IQR): 3사분위수(75%) - 1사분위수(25%). 데이터의 중간 50%가 포함되는 범위
- 분산 (Variance): 평균으로부터 떨어진 거리의 제곱의 평균. 단위가 원래 단위의 제곱이 되어 해석이 어려움.
- 표준편차 (Standard Deviation): 분산의 제곱근. 원래 데이터와 단위가 같아 해석이 용이하며, 평균적으로 값이 평균에서 얼마나 떨어져 있는지를 나타냄.

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 범위와 사분위수 범위 (Range & IQR)

summary() 함수를 통해 R에서 쉽게 계산할 수 있음.

```
summary(gdp_2007$gdpPercap)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   277.6  1624.8  6124.4 11680.1 18008.8 49357.2
```

- 범위 (Range): 49357.19 - 277.55 = 49079.64
- 사분위수 범위 (IQR): 11981.85 - 1624.84 = 10357.01
    - 이는 1인당 GDP의 중간 50%가 약 $10,357의 범위 안에 분포하고 있음을 의미

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 분산(Variance)과 표준편차(Standard Deviation)

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

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

```
# 2007년 기대수명(lifeExp)의 분산과 표준편차
var(gdp_2007$lifeExp)
```

```
## [1] 145.7578
```

```
sd(gdp_2007$lifeExp)
```

```
## [1] 12.07302
```

2007년 국가들의 평균 기대수명은 약 67세, 표준편차는 약 12년

- 국가들의 기대수명이 평균을 중심으로 약 12년 정도 퍼져있다는 의미

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 히스토그램

연속형 변수의 분포를 시각적으로 표현하는 가장 일반적인 방법

데이터의 범위를 여러 "구간(bin)"으로 나누고 각 구간에 속하는 데이터의 빈도를 막대로 나타냄.

- 구간의 너비(binwidth)에 따라 히스토그램의 모양이 달라질 수 있으므로, 여러 너비를 시도해보는 것이 좋음.

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 히스토그램

R Code | Plot

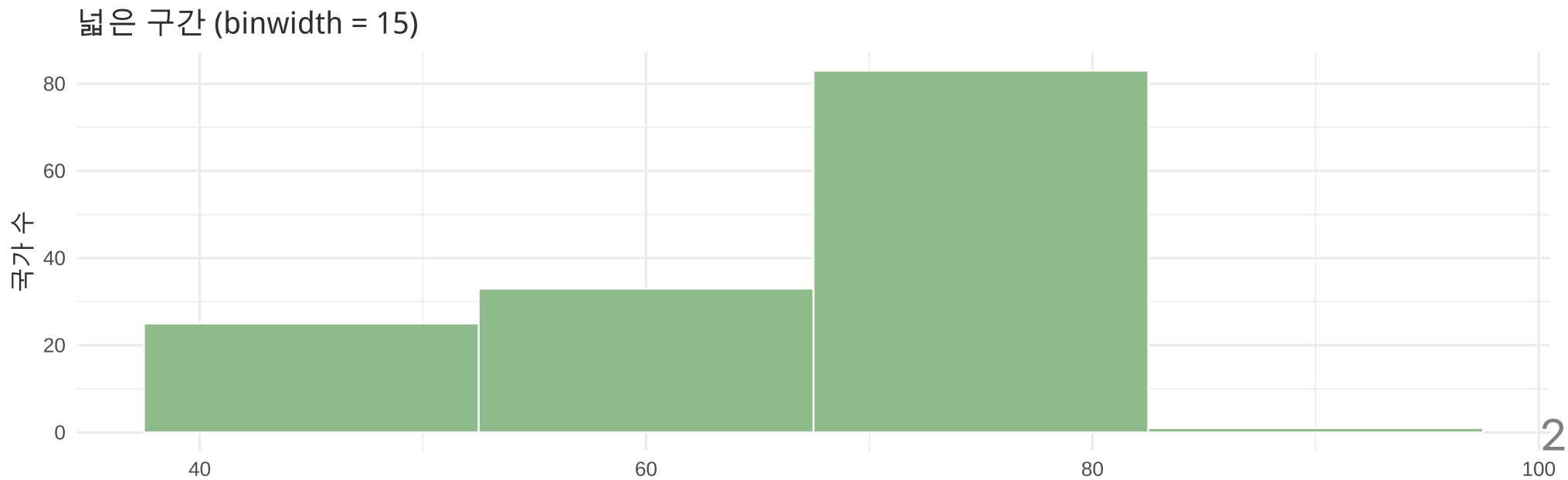
```
# 2007년 기대수명 분포
gdp_2007 |> ggplot(aes(x = lifeExp)) +
  geom_histogram(binwidth = 5, fill = "darkseagreen", color = "white") +
  labs(title = "2007년 전 세계 기대수명 분포",
       x = "기대수명 (세)", y = "국가 수")
```

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 히스토그램 구간 너비 비교

구간 너비가 너무 넓으면 분포의 세부 정보가 사라지고, 너무 좁으면 노이즈가 심해져 전체적인 패턴을 보기 어려움.

넓은 구간 (binwidth = 15) | 좁은 구간 (binwidth = 1)



데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 커널 밀도 플롯(Kernel Density Plot)

히스토그램을 부드러운 곡선 형태로 나타낸 것

분포의 모양을 더 명확하게 파악할 수 있게 해줌.

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 커널 밀도 플롯(Kernel Density Plot)

R Code | Plot

```
# 2007년 기대수명 분포
gdp_2007 |> ggplot(aes(x = lifeExp)) +
  geom_density(fill = "indianred", alpha = 0.5) +
  labs(title = "2007년 전 세계 기대수명 분포 (밀도 플롯)",
       x = "기대수명 (세)",
       y = "밀도")
```

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 커널 밀도 플롯(Kernel Density Plot)

히스토그램을 부드러운 곡선 형태로 나타낸 것

분포의 모양을 더 명확하게 파악할 수 있게 해줌.

- 앞서의 분포는 두 개의 봉우리를 가진 쌍봉형(bimodal) 분포로 보임.
- 기대수명이 낮은 국가 그룹과 높은 국가 그룹, 두 집단으로 나뉠 수 있음을 시사

데이터와 정보 II: 단변량 & 양변량 분포

산포도의 측정: 히스토그램과 밀도 플롯 겹치기

두 그래프를 겹쳐 그리면 분포에 대한 더 풍부한 정보를 얻을 수 있음.

R Code | Plot

```
# 2007년 기대수명 분포
gdp_2007 |> ggplot(aes(x = lifeExp)) +
  geom_histogram(aes(y = ..density..), binwidth = 5, fill = "darkseagreen", color = "white") +
  geom_density(color = "indianred", size = 1) +
  labs(title = "2007년 기대수명 분포 (히스토그램 & 밀도 플롯)",
       x = "기대수명 (세)",
       y = "밀도")
```

데이터와 정보 II: 단변량 & 양변량 분포

단변량 분석 정리

범주형 변인

- 각 범주의 빈도를 파악하고자 하는 목적
- 빈도표, 최빈값, 막대 그래프 이용

연속형 변인

- 중심과 퍼짐의 정도를 파악하고자 하는 목적
- 평균, 중앙값, 표준편차, IQR, 히스토그램, 밀도 플롯 등 이용

Part II. 양변량 분포(Bivariate Distribution)

데이터와 정보 II: 단변량 & 양변량 분포

양변량 분포란?

하나의 관측 단위에 대해 두 개의 변인을 동시에 측정한 데이터 집합

- 두 변인 간의 관계(relationship) 또는 연관성(association)을 탐색하는 것
- 한 가지 변인이 변할 때, 다른 변인은 어떻게 변하는가?
- 두 변인은 함께 움직이는가? (정적 관계)? 반대로 움직이는가? (부적 관계) 아니면 관계가 없는가?

데이터와 정보 II: 단변량 & 양변량 분포

관계의 유형

양변량 분석 방법은 두 변인의 측정 수준 조합에 따라 달라짐.

범주형 대 범주형

- 예: 성별과 정당 지지 간의 관계
- 도구: 교차표, 그룹/누적 막대 그래프

범주형 대 연속형

- 예: 대륙과 기대수명의 관계
- 도구: 그룹별 기술통계, 박스플롯, 바이올린 플롯

연속형 대 연속형

- 예: 1인당 GDP와 기대수명의 관계
- 도구: 산점도, 상관계수, 회귀분석

데이터와 정보 II: 단변량 & 양변량 분포

범주형-범주형: 교차표

두 범주형 변수의 빈도를 행과 열로 구성한 표로 두 변수 간의 연관성을 파악

R Code | Output

```r
# 2007년 아시아 국가들의 지역(sub_region)과
# 1인당 GDP 수준(gdp_level) 간의 관계
# gdp_level: 1인당 GDP 중앙값을 기준으로 'High'와 'Low'로 나눔
asia_2007 <- asia_2007 %>%
  mutate(gdp_level = ifelse(gdpPercap > median(gdpPercap), "High", "Low"))

# 교차표 생성
asia_2007 %>%
  count(sub_region, gdp_level) %>%
  pivot_wider(names_from = gdp_level, values_from = n, values_fill = 0)
```

데이터와 정보 II: 단변량 & 양변량 분포

범주형-범주형: 그룹 막대 그래프(Grouped Bar Plot)

교차표의 정보를 시각화하여 그룹 간의 차이를 더 쉽게 비교

R Code | Output

```
# 2007년 아시아 국가들의 지역(sub_region)과
# 1인당 GDP 수준(gdp_level) 간의 관계
# gdp_level: 1인당 GDP 중앙값을 기준으로 'High'와 'Low'로 나눔
asia_2007 |> ggplot(aes(x = sub_region, fill = gdp_level)) +
  geom_bar(position = "dodge") + # "dodge"는 막대를 옆으로 나란히 배치
  labs(title = "아시아 지역별 GDP 수준 분포",
       x = "하위 지역", y = "국가 수", fill = "GDP 수준")
```

데이터와 정보 II: 단변량 & 양변량 분포

범주형-범주형: 비율 누적 막대 그래프

`position = "fill"` 옵션을 사용하면 각 그룹 내에서 차지하는 비율을 비교하기 용이

R Code | Output

```
# 2007년 아시아 국가들의 지역(sub_region)과
# 1인당 GDP 수준(gdp_level) 간의 관계
# gdp_level: 1인당 GDP 중앙값을 기준으로 'High'와 'Low'로 나눔
asia_2007 |> ggplot(aes(x = sub_region, fill = gdp_level)) +
  geom_bar(position = "fill") + # "fill"은 비율로 표시
  scale_y_continuous(labels = scales::percent) +
  labs(title = "아시아 지역별 GDP 수준 비율",
       x = "하위 지역", y = "비율", fill = "GDP 수준")
```

데이터와 정보 II: 단변량 & 양변량 분포

범주형-연속형: 그룹별 박스 플롯

범주에 따라 연속형 변인의 분포가 어떻게 다른지 비교하는데 매우 효과적

R Code | Output

```r
# 2007년 대륙(continent)별 기대수명(lifeExp) 분포 비교
gdp_2007 |> ggplot(aes(x = continent, y = lifeExp, fill = continent)) +
  geom_boxplot() +
  labs(title = "대륙별 기대수명 분포 (2007)",
       x = "대륙", y = "기대수명") +
  theme(legend.position = "none")
```

데이터와 정보 II: 단변량 & 양변량 분포

`redorder()`로 박스 플롯 정렬하기

중앙값 순서로 박스 플롯을 정렬하면 대륙 간의 차이를 더 명확하게 비교할 수 있음.

R Code | Output

```
# 2007년 대륙(continent)별 기대수명(lifeExp) 분포 비교
gdp_2007 |> ggplot(aes(x = reorder(continent, lifeExp, FUN = median), y = lifeExp, fill = co
  geom_boxplot() +
  labs(title = "대륙별 기대수명 분포 (중앙값 순 정렬)",
       x = "대륙", y = "기대수명") +
  theme(legend.position = "none")
```

데이터와 정보 II: 단변량 & 양변량 분포

연속형-연속형: 산점도(Scatter Plot)

두 연속형 변수 간의 관계를 시각화하는 가장 기본적인 방법. 각 관측치를 x축과 y축의 값에 해당하는 점으로 표시

R Code | Output

```r
# 2007년 1인당 GDP와 기대수명의 관계
gdp_2007 |> ggplot(aes(x = gdpPercap, y = lifeExp)) +
  geom_point() +
  labs(title = "1인당 GDP와 기대수명 (2007)",
       x = "1인당 GDP", y = "기대수명")
```

데이터와 정보 II: 단변량 & 양변량 분포

연속형-연속형: 산점도(Scatter Plot)

두 연속형 변수 간의 관계를 시각화하는 가장 기본적인 방법. 각 관측치를 x축과 y축의 값에 해당하는 점으로 표시

1인당 GDP가 높은 국가일수록 기대수명도 높은 경향이 보임(정적 관계).

하지만 x축의 분포가 매우 치우쳐 있어 관계를 자세히 보기 어려움.

데이터와 정보 II: 단변량 & 양변량 분포

연속형-연속형: 축 변환 (Axis Transformation)

한쪽으로 치우친 변수는 로그 변환(log transformation)을 통해 분포를 대칭에 가깝게 만들고, 변수 간의 선형 관계를 더 명확하게 볼 수 있음.

R Code | Output

```
gdp_2007 |> ggplot(aes(x = gdpPercap, y = lifeExp)) +
  geom_point() +
  scale_x_log10() + # x축을 로그 스케일로 변환
  labs(title = "1인당 GDP(로그)와 기대수명 (2007)",
       x = "1인당 GDP (로그 스케일)", y = "기대수명")
```

데이터와 정보 II: 단변량 & 양변량 분포

연속형-연속형: 축 변환 (Axis Transformation)

한쪽으로 치우친 변수는 로그 변환(log transformation)을 통해 분포를 대칭에 가깝게 만들고, 변수 간의 선형 관계를 더 명확하게 볼 수 있음.

로그 변환 후, 두 변수 간의 양의 선형 관계가 훨씬 뚜렷하게 나타남.

데이터와 정보 II: 단변량 & 양변량 분포

연속형-연속형: 산점도 꾸미기 Aesthetics 추가

aes() 안에 color, size, shape 등 다른 변수를 추가하여 더 많은 정보를 한 번에 표현할 수 있음.

R Code | Output

```
gdp_2007 |> ggplot(aes(x = gdpPercap, y = lifeExp, color = continent, size = pop)) +
  geom_point(alpha = 0.7) + # 점의 투명도 조절
  scale_x_log10(labels = scales::dollar) +
  scale_size_continuous(range = c(1, 10), labels = scales::comma) + # 점 크기 범위 조절
  labs(title = "1인당 GDP와 기대수명 (2007)",
       x = "1인당 GDP (로그 스케일)", y = "기대수명",
       color = "대륙", size = "인구")
```

데이터와 정보 II: 단변량 & 양변량 분포

연속형-연속형: 산점도 꾸미기 Aesthetics 추가

`aes()` 안에 `color`, `size`, `shape` 등 다른 변수를 추가하여 더 많은 정보를 한 번에 표현할 수 있음.

대륙별로 색을, 인구별로 점의 크기를 다르게 하여 다차원적인 정보를 효과적으로 전달

데이터와 정보 II: 단변량 & 양변량 분포

관계의 수치적 요약

시각화는 관계의 전반적인 형태를 보여주지만, 관계의 방향과 강도를 정량적으로 요약하기 위해서는 수치적 측도가 필요

- 공분산(Covariance)
- 상관계수(Correlation Coefficient)

데이터와 정보 II: 단변량 & 양변량 분포

관계의 수치적 요약: 공분산

두 변수가 함께 변화하는 정도

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$Cov > 0$: 두 변수가 같은 방향으로 움직임 (정적 관계)

$Cov < 0$: 두 변수가 반대 방향으로 움직임 (부적 관계)

단, 변수의 단위에 따라 공분산의 값은 크게 달라지므로, 관계의 강도를 직접 비교하기는 어려움.

데이터와 정보 II: 단변량 & 양변량 분포

관계의 수치적 요약: 공분산

두 변수가 함께 변화하는 정도

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$Cov > 0$: 두 변수가 같은 방향으로 움직임 (정적 관계)

$Cov < 0$: 두 변수가 반대 방향으로 움직임 (부적 관계)

```
# 2007년 1인당 GDP와 기대수명의 공분산
cov(gdp_2007$gdpPercap, gdp_2007$lifeExp)
```

```
## [1] 105368
```

데이터와 정보 II: 단변량 & 양변량 분포

관계의 수치적 요약: 상관계수

공분산을 각 변수의 표준편차로 나누어 표준화한 값으로, -1과 1 사이의 값을 가짐. **피어슨 상관계수(r)**가 가장 널리 쓰임.

$$r = \frac{Cov(x, y)}{s_x s_y}$$

- -1: 완벽한 음의 선형 관계
- 0: 선형 관계 없음.
- +1: 완벽한 양의 선형 관계

상관관계는 단위의 영향을 받지 않아 관계의 방향과 강도 모두를 비교 가능

```
cor(gdp_2007$gdpPercap, gdp_2007$lifeExp)
```

```
## [1] 0.6786624
```

1인당 GDP와 기대수명 간에는 0.68의 강한 양의 선형 관계

데이터와 정보 II: 단변량 & 양변량 분포

관계의 수치적 요약: 상관계수

상관계수는 선형(linear) 관계의 강도만을 측정

- 비선형 관계가 강하더라도 상관계수는 0에 가까울 수 있음.

상관관계는 인과관계가 아님!!!!

- 상관계수는 두 변인이 함께 움직이는 경향을 보여줄 뿐, 하나의 변인이 다른 변인의 원인이 됨을 의미하지 않음.
- **제3의 변인(confounding variable)**이나 우연에 의해 높은 상관관계가 나타날 수 있음.

양변량 분석의 방법 중 하나로 단순회귀분석이 있으나, 이는 나중에 다룰 예정

데이터와 정보 II: 단변량 & 양변량 분포

양변량 분석 정리

범주형-범주형

- 그룹/누적 막대 그래프
- 교차표

범주형-연속형

- 박스 플롯, 바이올린 플롯
- 그룹별 통계(평균, 중앙값)

연속형-연속형

- 산점도
- 공분산, 상관계수, 회귀계수, R^2

데이터와 정보 II: 단변량 & 양변량 분포

단변량 분석: 데이터의 기초적인 특성(중심, 산포, 분포)을 파악하는 과정

양변량 분석: 두 변수 간의 관계(방향, 강도, 형태)를 시각적, 수치적으로 탐색하는 과정

기술통계는 "어떤 관계가 있는가"를 보여줄 뿐, "왜 그런 관계가 있는가"(인과관계)를 설명하지는 않음!

R을 이용한 시각화 실습 및 질의응답

KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

감사합니다!

궁금한 것이 있으면 언제든 연락하세요.

강사 연락처

연락처	박상훈
	sh.park.poli@gmail.com
	sanghoon-park.com/
	영상바이오관 405