



한양대학교  
HANYANG UNIVERSITY

# 평균, 이상치, 극단값

---

박선철

2025-04-18

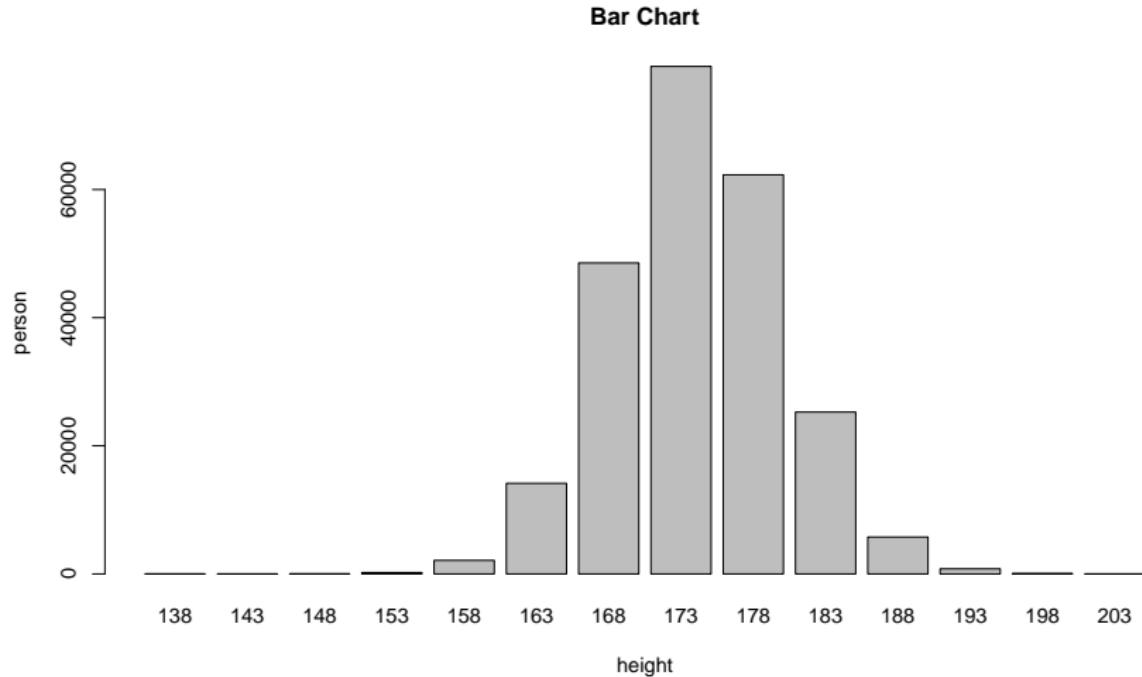
한양대학교 수학과



# 대한민국 성인의 키

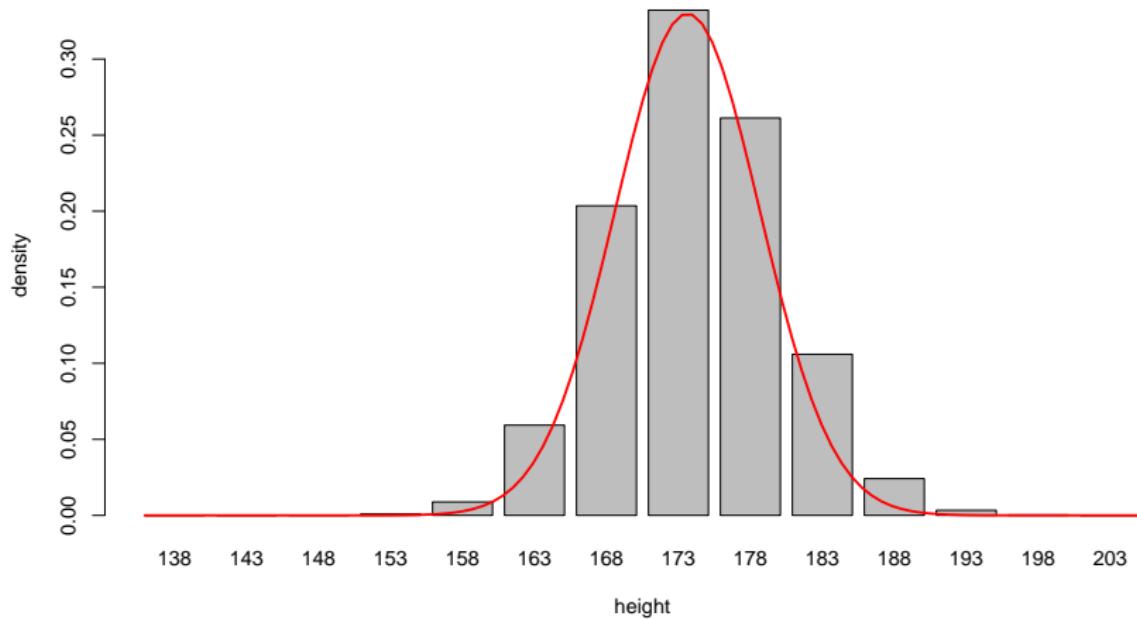
Q. 대한민국 20 세 남자의 평균 키?

- 2023년 신장별 대한민국 병역판정검사 현황



- 바 차트와 확률밀도함수

Bar Chart with Normal Density



**Q. 대한민국 20 세 여자의 평균 키?**

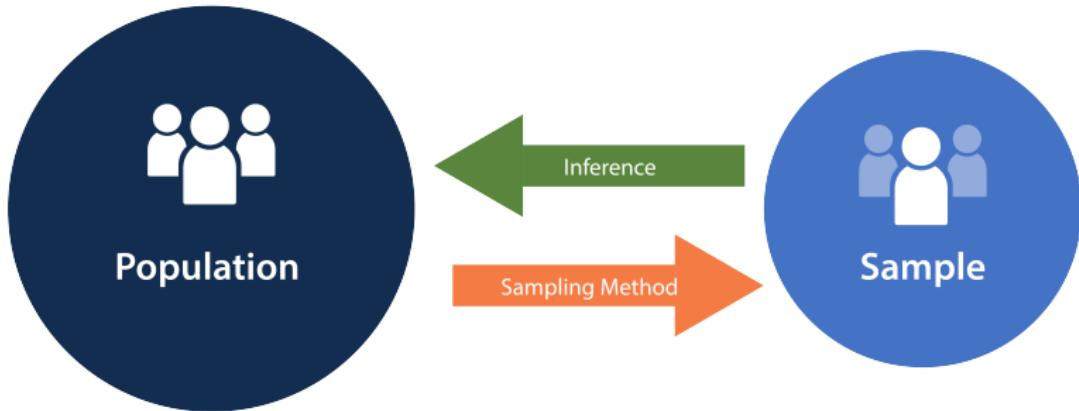
- 8 차 인체치수조사 보고서

**Q. 왜 대한민국 20 세 여자 전체를 대상으로 키를 조사하지 않을까?**



# 모집단과 표본

- 전수조사: 표본오차가 없다는 것이 장점
- 표본조사: 모집단 전체를 조사하는데 시간과 비용이 많이 필요하므로 표본조사를 실시



- 표본을 잘 선택해야 모집단을 잘 대표할 수 있음
  - 충분히 크기가 큰 표본
  - 학력, 연령 등이 치우치지 않은 표본

# 표본평균과 표본분산

- 관찰값이  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  과 같을 때, 이것의 표본평균  $\bar{x}$  와 표본분산  $s_x^2$  는 다음과 같이 정의됨

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

- 표본표준편차: 표본분산의 제곱근
- 표본평균: 모집단의 중심을 추정할 때 좋은 추정량
- 표본분산과 표본표준편차: 모집단의 산포를 추정할 때 좋은 추정량

# 대수의 법칙

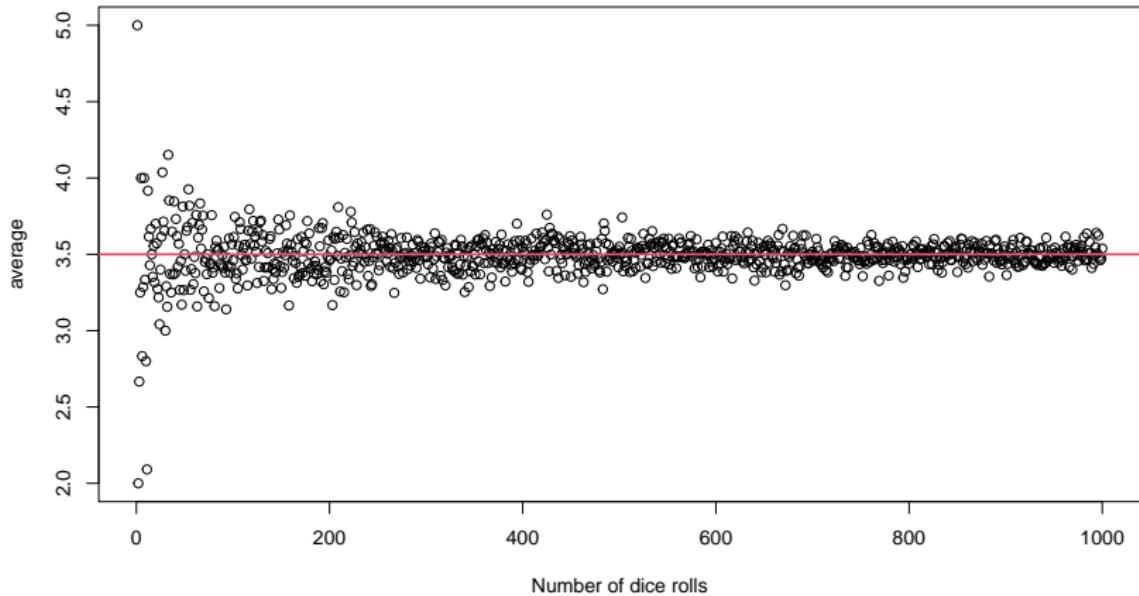
- 주사위 한 개를 던질 때



**Q.** 우리가 얻을 수 있는 주사위 눈금은?

**Q.** 우리가 얻을 수 있는 주사위 눈금의 기댓값은?

- 대수의 법칙 (law of large numbers, LLN): 표본집단의 크기가 커지면 표본평균이 모평균에 가까워짐



**Figure 1:** 주사위를 많이 던지면 표본평균은 3.5 (빨간 선)로 수렴.

# 중심극한정리

- 중심극한정리 (central limit theorem, CLT): 모집단의 분산이 유한하다면 모집단의 형태와 상관없이 같은 모집단에서 독립으로 뽑은 표본들의 표본평균의 분포는 정규분포로 분포수렴함

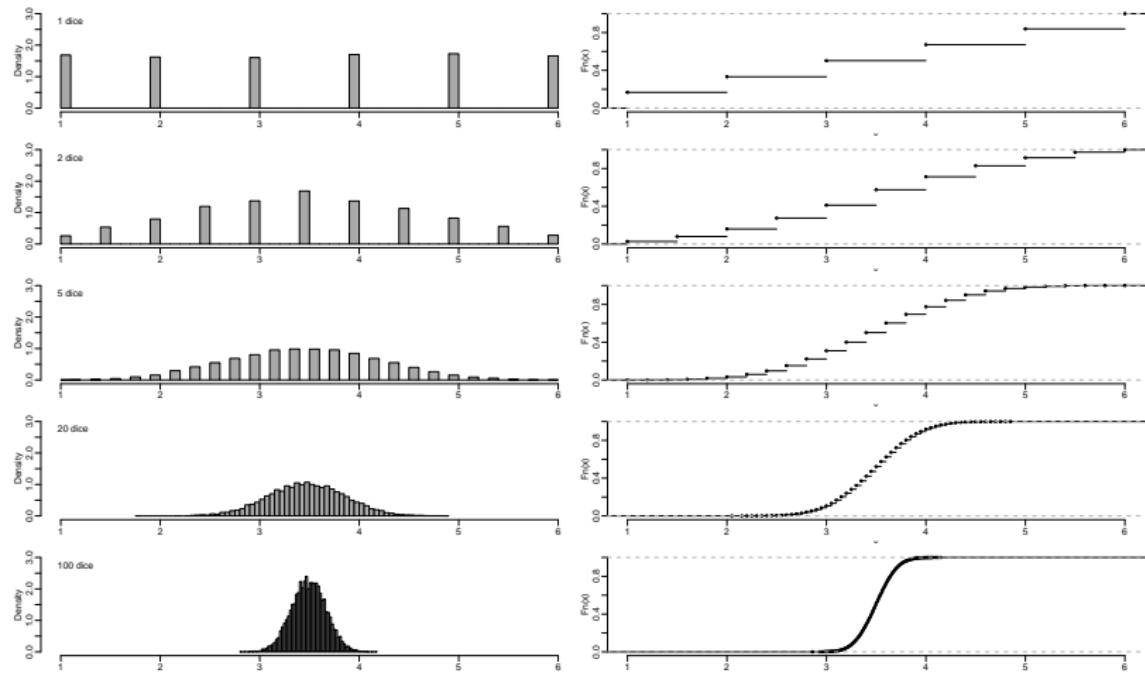
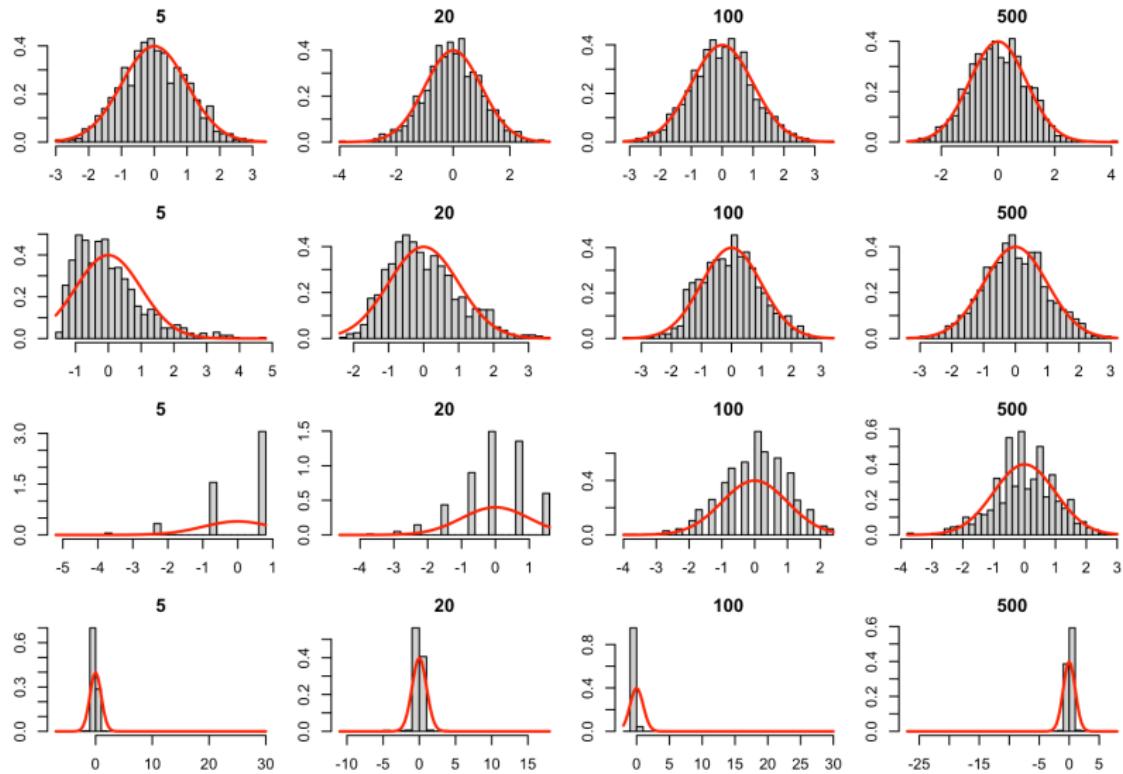


Figure 2: 주사위를 많이 던질수록 얻을 수 있는 표본평균의 확률밀도함수는 정규분포로 수렴.



**Figure 3:** 정규, 감마, 베르누이, 코시 분포와 CLT.

# 정규분포

- 확률변수  $X$  가 평균이  $\mu$  이고 분산이  $\sigma^2$  인 정규분포를 따른다고 할 때,  $X$  의 확률밀도함수는

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

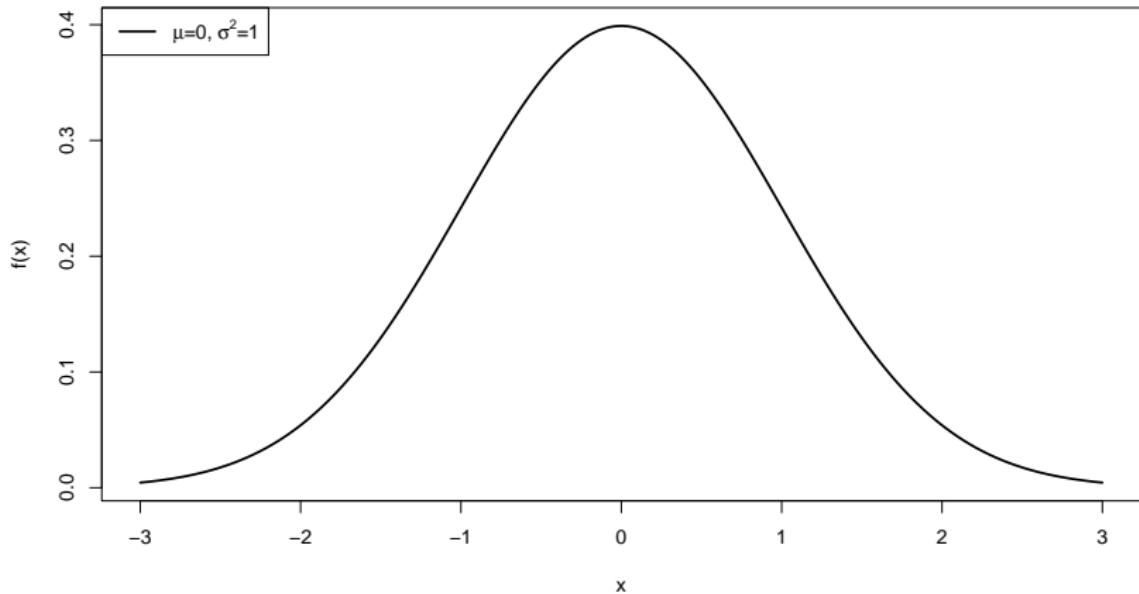
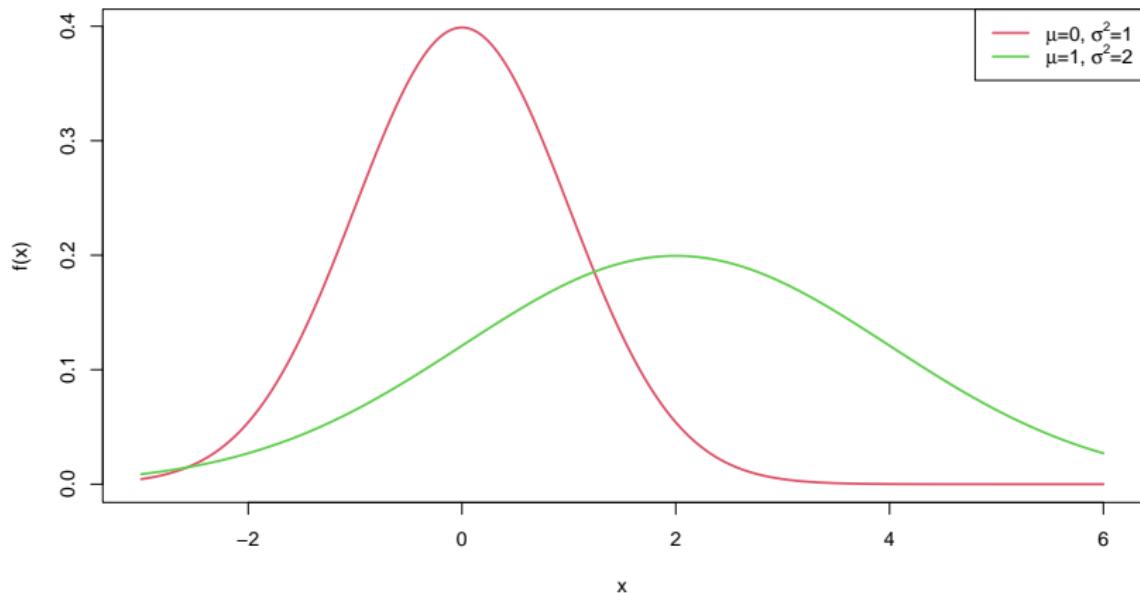


Figure 4: 정규분포의 밀도함수.

# 정규분포에서 평균과 분산



**Figure 5:** 정규분포에서 평균과 분산의 역할.

# 표본중앙값과 표본분위수

- 관찰값이  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  과 같이 있을 때, 이것의 표본중앙값  $t$  는

$$t = \text{Med}(\mathbf{x}), \quad \text{if } \#\{x_i > t\} = \#\{x_i < t\},$$

이때  $\#$  기호는  $\{\}$  안을 만족하는 표본의 개수를 의미함

- Median absolute deviation about the median (MAD):

$$\text{MAD}(\mathbf{x}) = \text{MAD}(x_1, x_2, \dots, x_n) = \text{Med}\{|\mathbf{x} - \text{Med}(\mathbf{x})|\}.$$

- 표본중앙값은 표본평균 대용으로, MAD 는 표본표준편차 대용으로 쓸 수 있음

# 평균과 중앙값

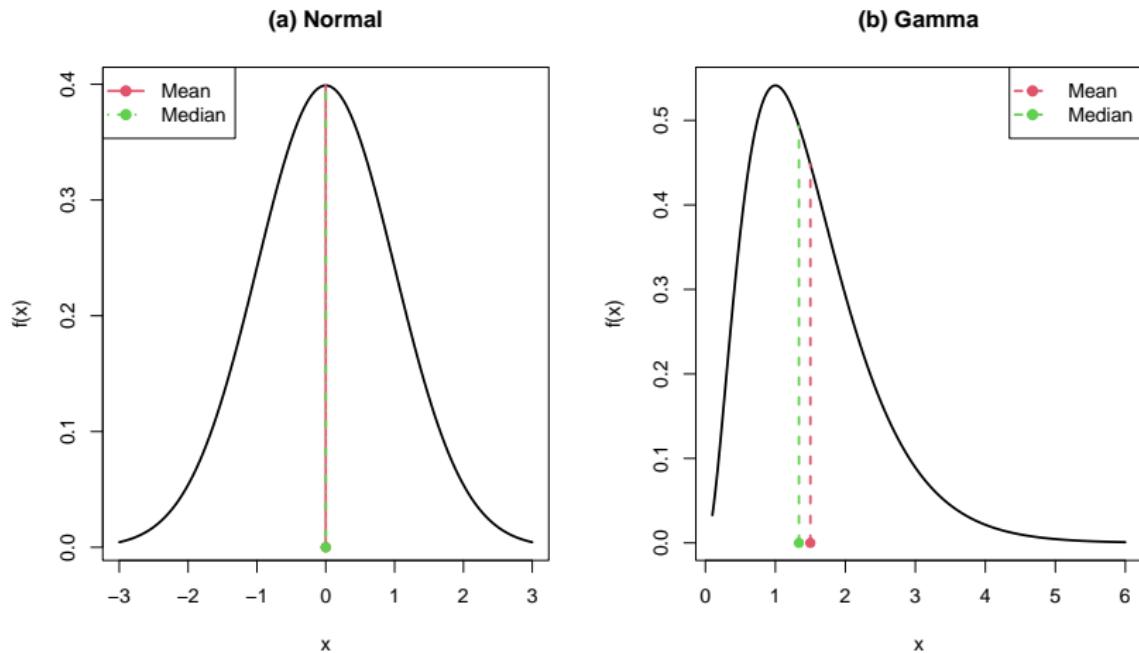


Figure 6: (왼쪽) 정규분포의 평균과 중앙값. (오른쪽) 감마분포의 평균과 중앙값.

# 이상치와 평균

- Example from Maronna et al. (2019) and Jung (2022)

**i** 다음은 24 개의 밀가루 상품에서 추출한 구리 성분의 함량 (그램 당 백만분의 일 그램 단위) 을 오름차순으로 정렬한 데이터이다.

2.20, 2.20, 2.40, 2.40, 2.50, 2.70,  
2.80, 2.90, 3.03, 3.03, 3.10, 3.37,  
3.40, 3.40, 3.40, 3.50, 3.60, 3.70,  
3.70, 3.70, 3.70, 3.77, 5.28, 28.50

- 중앙값과 MAD 는 이상치가 있어도 모집단의 bulk 에 대한 통계량을 잘 추정할 수 있음 (로버스트 (robust))

```
mean(x) # 이상치 넣었을 때 평균
```

```
[1] 4.261667
```

```
mean(x[-24]) # 이상치 뺐을 때 평균
```

```
[1] 3.207826
```

```
median(x) # 이상치 넣었을 때 중앙값
```

```
[1] 3.385
```

```
median(x[-24]) # 이상치 뺐을 때 중앙값
```

```
[1] 3.37
```

```
sd(x) # 이상치 넣었을 때 표준편차
```

```
[1] 5.206295
```

```
sd(x[-24]) # 이상치 뺐을 때 표준편차
```

```
[1] 0.6871083
```

```
mad(x) # 이상치 넣었을 때 MAD
```

```
[1] 0.526323
```

```
mad(x[-24]) # 이상치 뺐을 때 MAD
```

```
[1] 0.504084
```

# 분위수

- 모집단의 bulk 가 아닌 상위권, 하위권, 더 나아가 꼬리 부분의 특성을 보려면 평균이 아닌 다른 통계량이 필요함
- 그림 출처: [“Quantile loss” blog post]

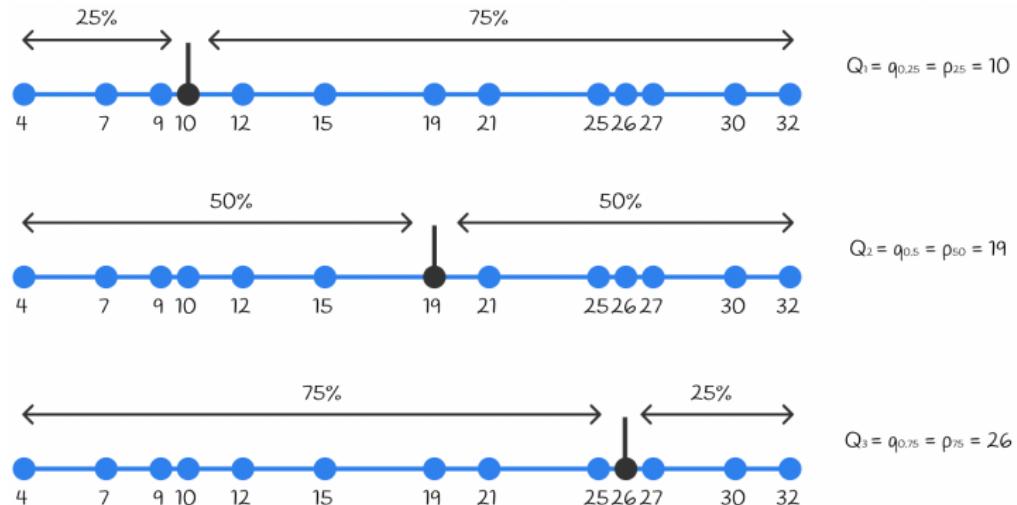


Figure 7: 사분위수 예시.

- 어떤 확률변수  $Y$ 의  $q$ -분위수  $y_{(q)}$  는  $P(Y < y_{(q)}) = q$  인 수

Gamma Distribution

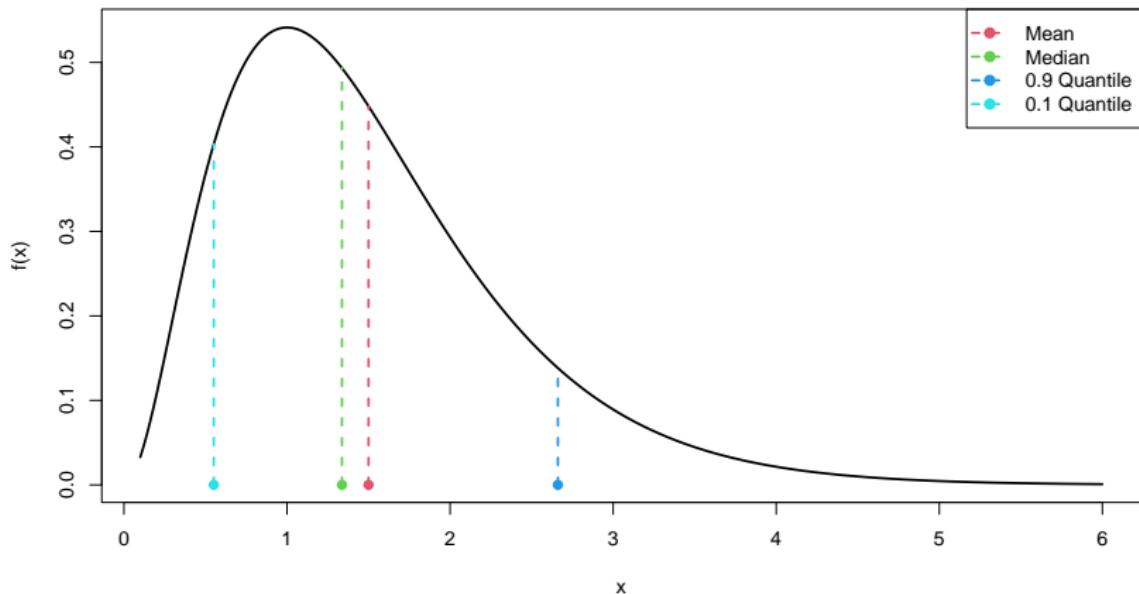


Figure 8: 감마분포의 평균, 중앙값, 분위수.

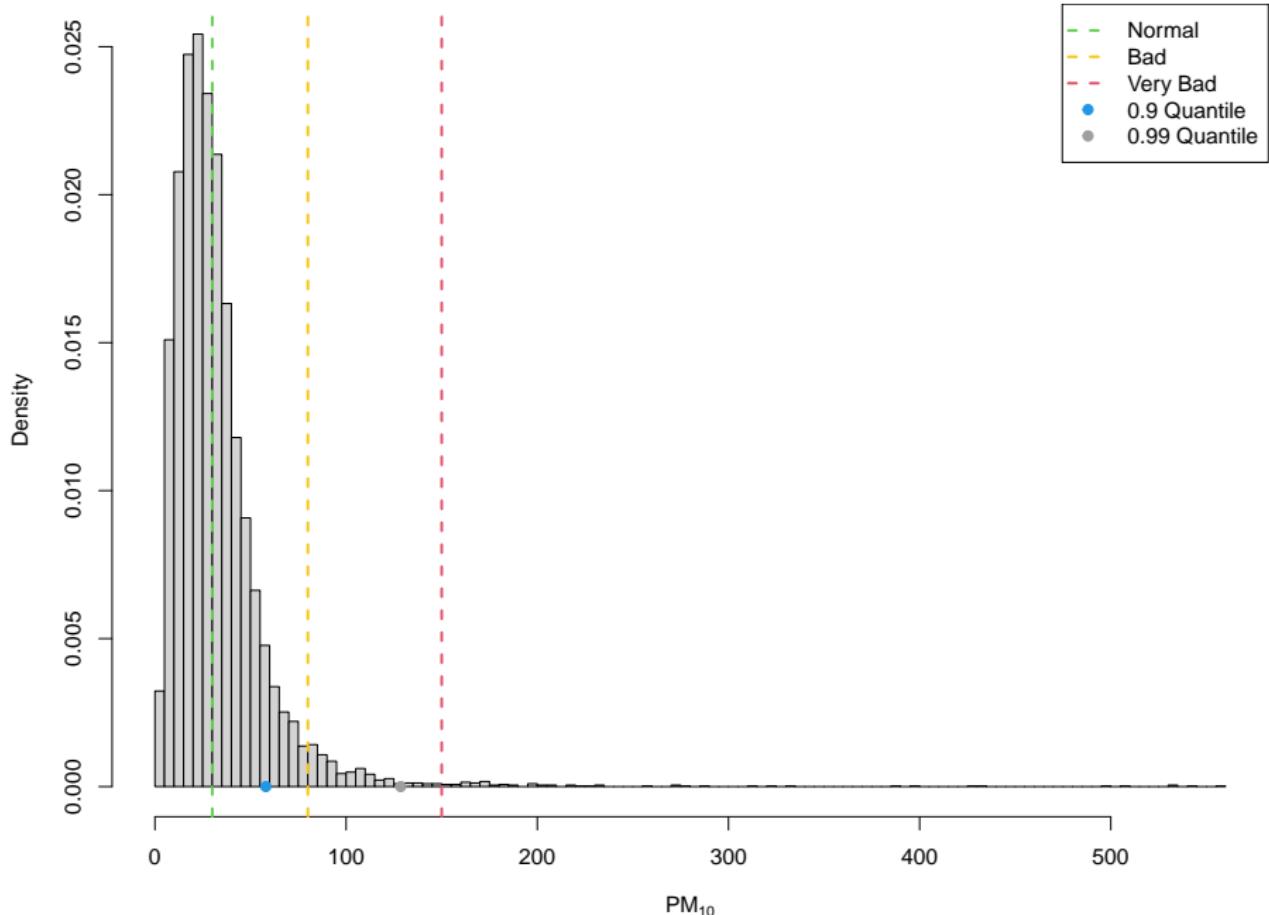
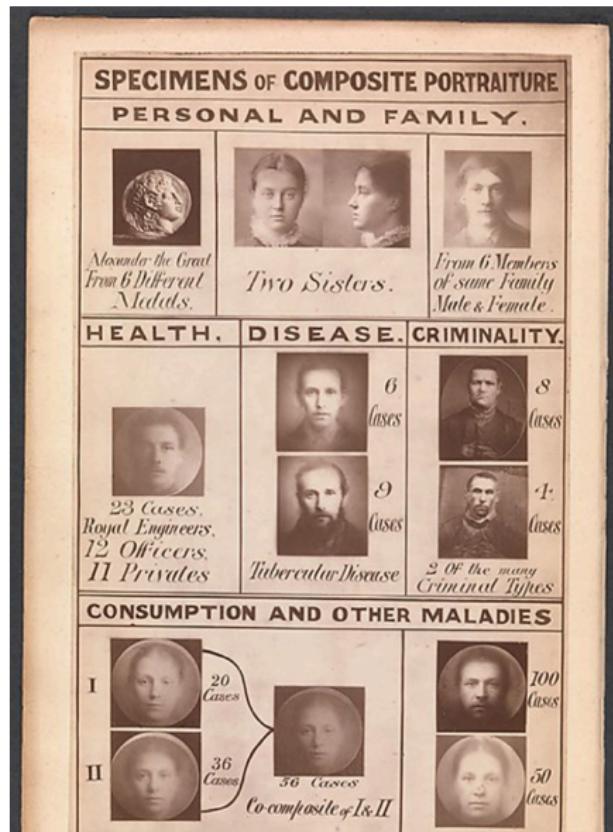


Figure 9: 2024년 서울 기상관측소의 1시간 미세먼지 농도의 히스토그램.

# 평균 인간?

- 그림 출처: Stephens and Cryle (2017)

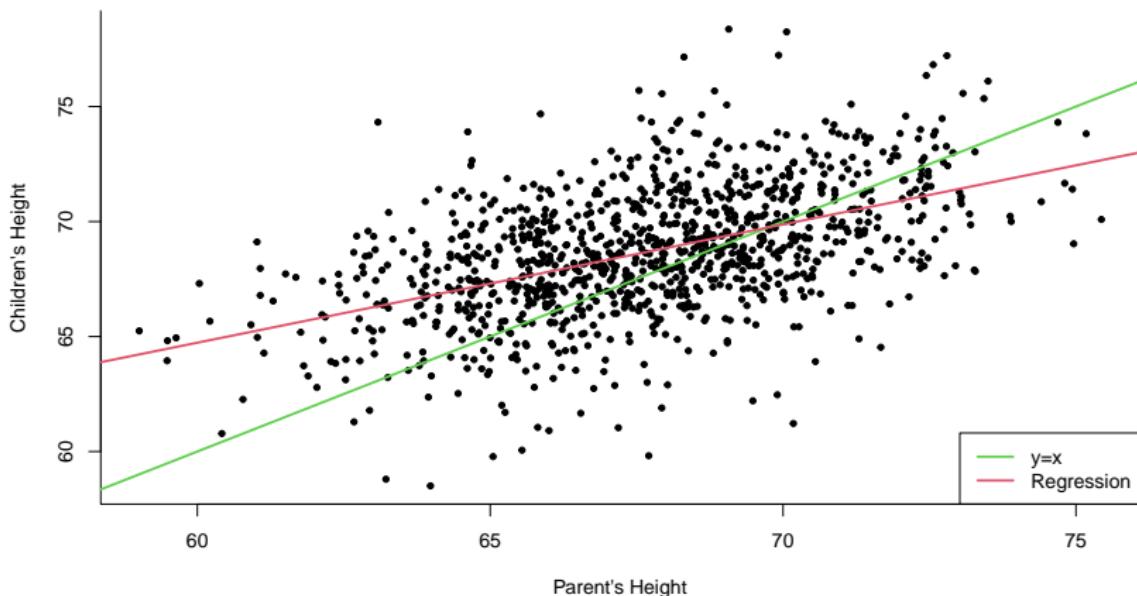


# 회귀분석

- i 1885년 프랜시스 골턴은 **child** ( $y$ , 아들)의 키 (인치로 측정) 와 **parent** ( $x$ , 부모)의 키 (인치로 측정) 과 상관이 있다고 생각하고 928 쌍의 부모와 자식의 키 자료를 모았다. (실제로는 딸의 키에는 1.08을 곱하여 수집하였고, 부모의 키는 (아버지의 키 + 1.08 × 어머니의 키)/2를 하여 자료를 수집했다고 한다.) 사람의 키가 유전자에만 영향을 받는다고 가정하고, 부모의 키 유전자가 자식의 키 유전자로 동일하게 전달된다면, 부모의 키가 크면 자식의 키도 당연히 클 것이다.

- 최소제곱법: 자료와의 오차를 최소로 하는 직선으로 두 변수 사이의 관계를 모델링

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

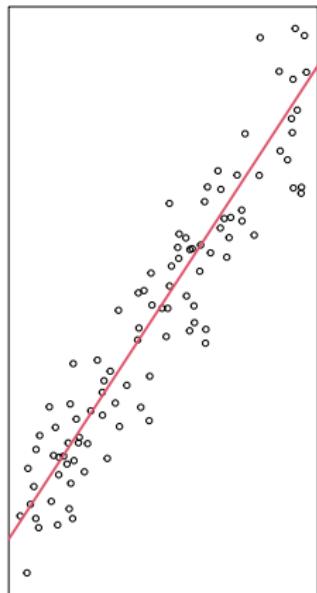


**Figure 11:** 피어슨의 부모와 자식 키 사이의 자료.

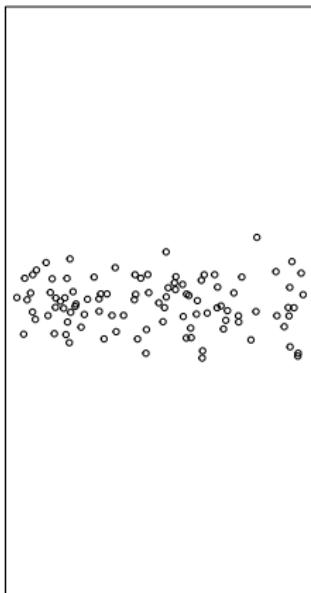
- 부모와 자식의 키가 항상 똑같다면, 기울기가 1인  $y = f(x) = x$  직선 위에 관측값들이 놓여야 한다.
- 적합된 회귀식은  
**Children's height** = 33.8866 + 0.5141 × **Parent's height**.

# 상관분석

(a) Positive Correlation



(b) No Linear Correlation



(c) Negative Correlation

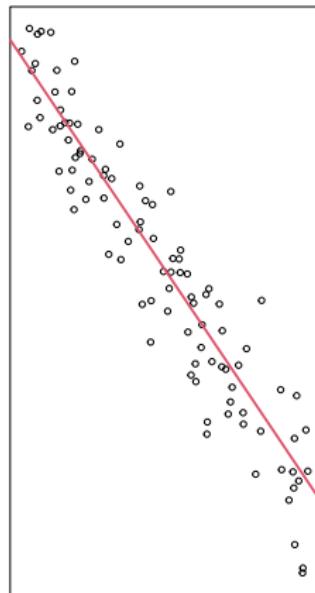
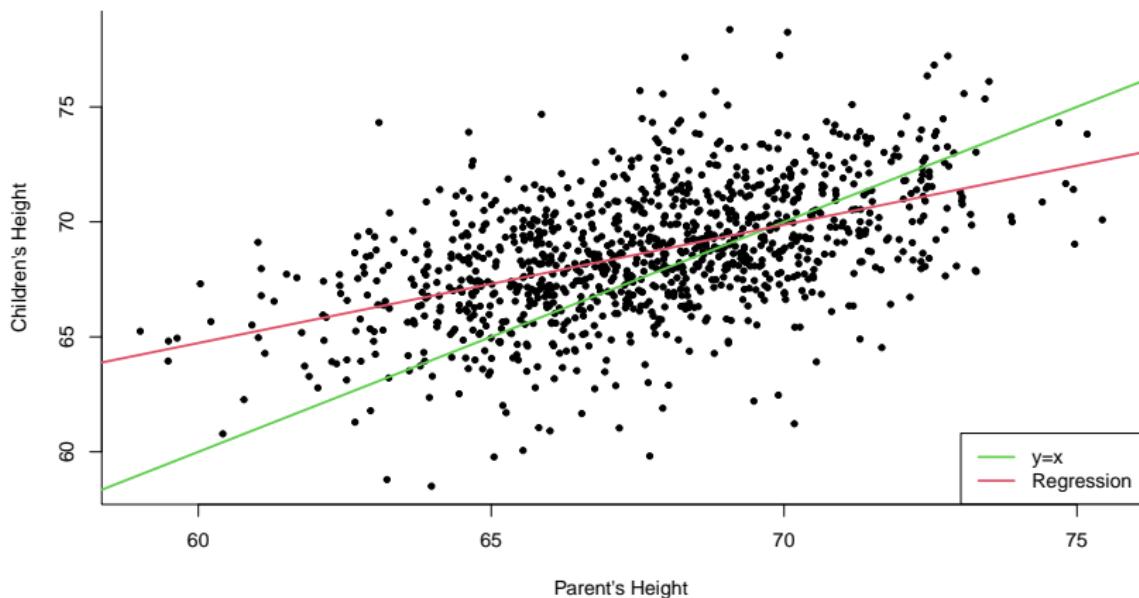


Figure 12: (a) 양의 상관관계, (b) 선형 상관관계 없음, (c) 음의 상관관계에 대한 시뮬레이션 자료.



**Figure 13:** 피어슨의 부모와 자식 키 사이의 자료.

- 이 자료의 상관계수  $r_{xy}$  는 0.5013
- 이 자료의 표본분산은 각각  $s_x = 2.7449$ ,  $s_y = 2.8147$  이고, 앞서 추정된 기울기의 회귀계수 0.5141 은  $0.5041 = 0.5013 \times \frac{2.8147}{2.7449}$  의 관계가 있음

# 상관성과 인과성

Q. 초콜릿을 많이 먹는 나라가 노벨상을 탈 확률이 높다?

- Chocolate Consumption, Cognitive Function, and Nobel Laureates, Messerli (2012)

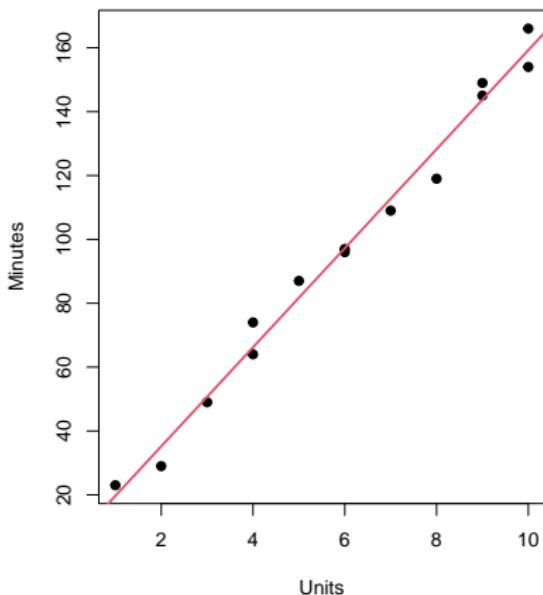
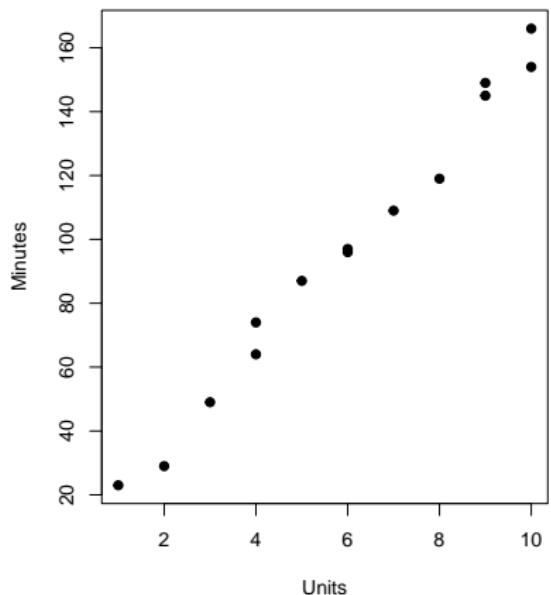


# 회귀분석 예제

- Hadi (2024) 의 예제

**i** 어떤 전자 제품 수리 회사는 서비스 통화를 받고 수리해야 하는 전자 부품의 수 만큼 수리비를 청구한다고 한다. 회사 직원들은 수리 시간이 길수록 수리해야 하는 전자 부품의 수가 많을 것이라고 생각했다. 이에 총  $n=14$  개의 사건을 조사해 수리 시간 (**Minutes**,  $y_i$ ,  $i = 1, \dots, 14$ ) 과 수리해야 하는 전자 부품의 수 (**Units**,  $x_i$ ,  $i = 1, \dots, 14$ ) 의 관계를 조사하였다.

	minutes	units
1	23	1
2	29	2
3	49	3
4	64	4
5	74	4
6	87	5



**Figure 15:** 최소제곱법으로 구한 회귀직선.

# 외삽의 위험성: 100m 세계기록 데이터

- 기사 출처: 2156년에는 여자가 더 빨라진다, 문화일보 노성열 기자, 2004년 9월 30일

## “2156년엔 여자가 더 빨라진다”

문화일보 | 입력 2004-09-30 11:40

🖨️ 프린트



노성열



댓글



폰트

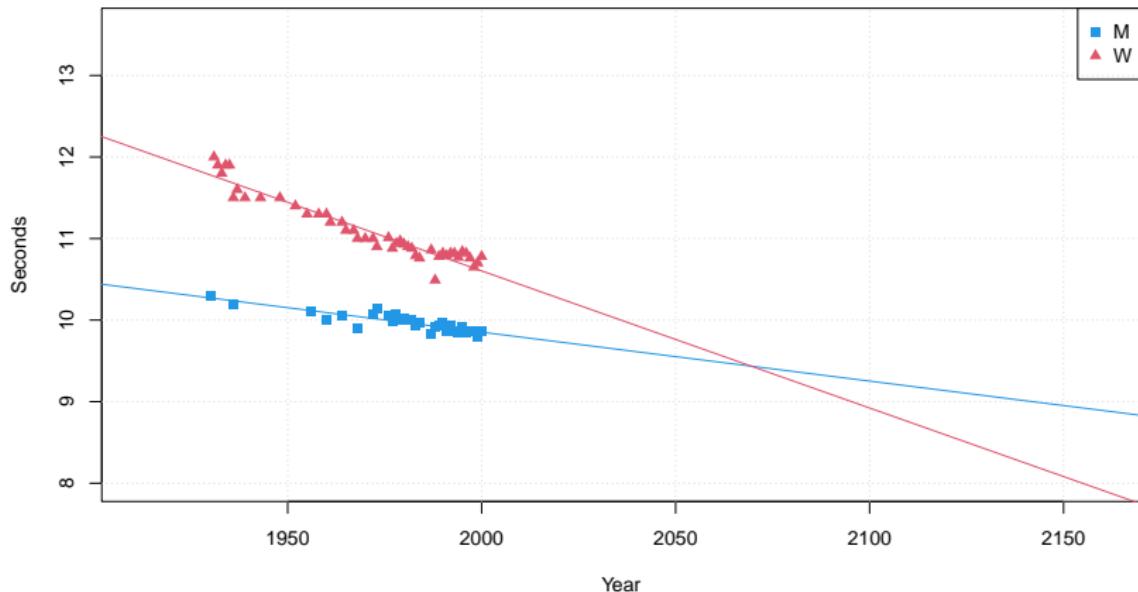


공유

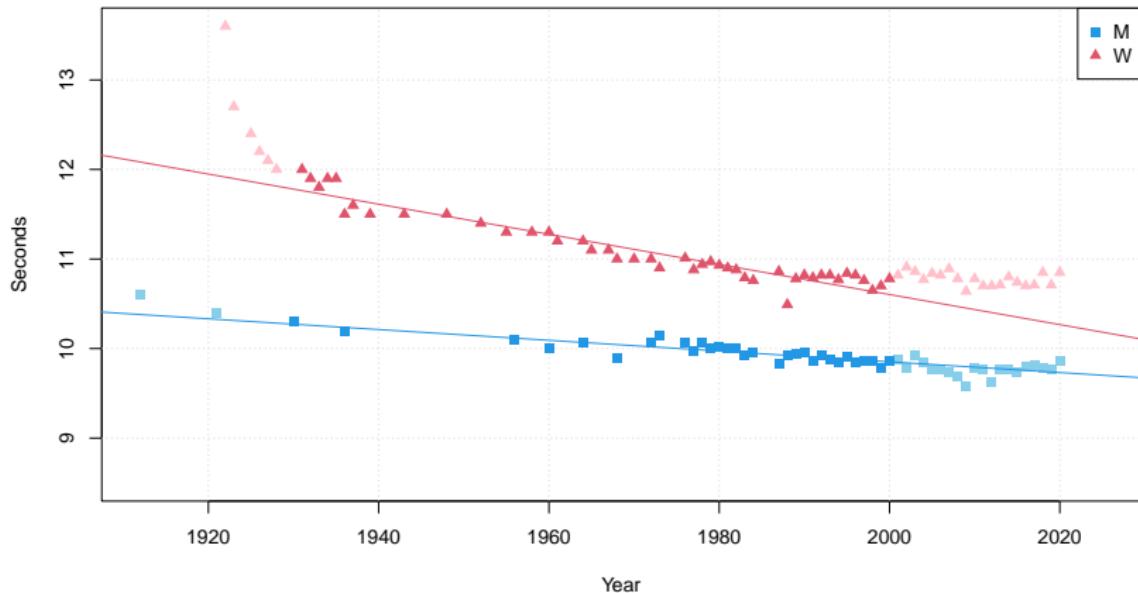


“2156년 올림픽 육상 100m에서 여자부 기록이 사상 최초로 남자부 기록을 추월하게 된다.”

- 자료 출처: <https://biostatistics.letgen.org/>



**Figure 16:** 1930~2000 년까지의 자료만을 이용하여 성별로 선형회귀직선을 적합하였다.



**Figure 17:** 1910~2020년까지의 자료를 이용하여 선형회귀직선을 적합하였을 때는 결과가 조금 달라진다.

# Anscombe's quartet

- Anscombe (1973) 의 논문

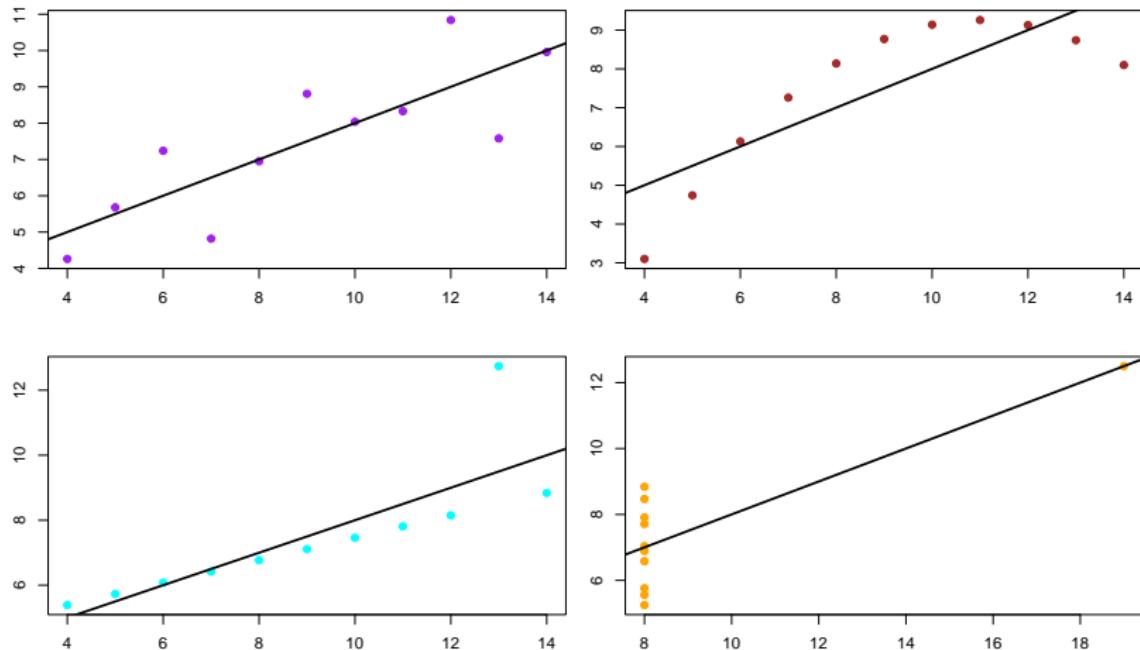
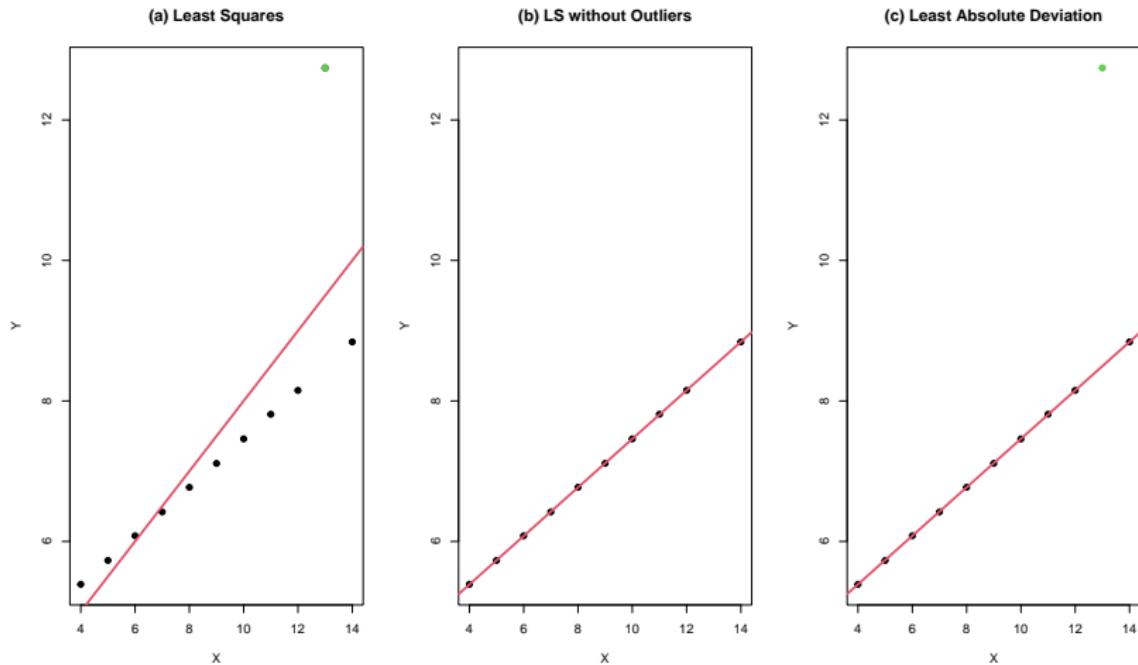


Figure 18: 분포나 그라프는 매우 다르지만 회귀직선은 동일한 네 개의 자료.

# 이상치와 회귀분석



**Figure 19:** (a) 최소제곱법으로 구한 회귀직선, (b) 이상치를 제거하고 최소제곱법으로 구한 회귀직선. (c) 최소절대편차법으로 구한 회귀직선.

# 이상치와 극단값

- 이상치: 자료의 bulk에 있는 다른 관찰값보다 많이 떨어져 있는 점
  - 제거해야 할 대상: 측정 및 기록의 오류, 아예 다른 분포에서 왔을 때
  - 같은 분포에서 왔을 때에는?
- 극단값  $\approx$ : 특정 블록 (한 달, 일년 등)의 관찰값의 최대/최소, 또는 특정 임계값 (180mm/12 시간, 호우경보) 이상

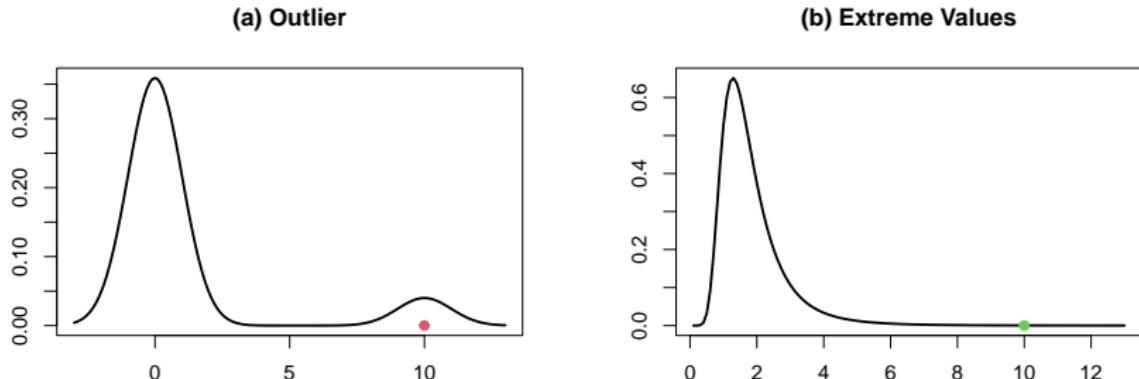
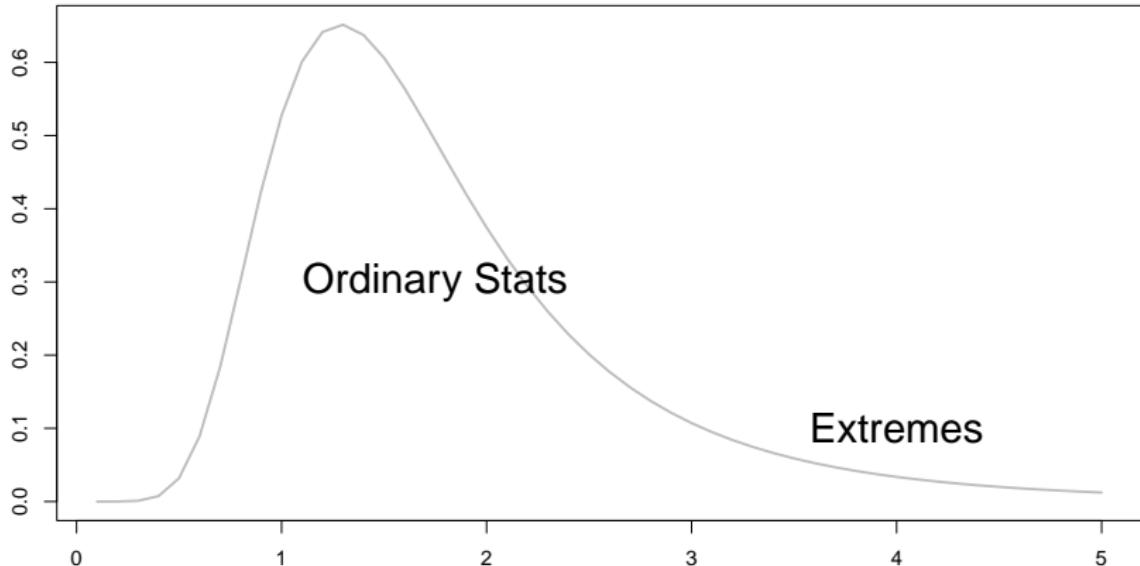


Figure 20: 이상치와 극단값의 묘사.

# 일반적인 통계 vs 극단값 통계

- 일반적인 통계: 확률분포의 주요한 부분 (bulk)에 관심
- 극단값 통계: 확률분포의 꼬리 부분 (tail)에 관심



**Figure 21:** 일반적인 통계는 확률분포의 bulk에 관심이 있으나, 극단값 통계는 확률분포의 tail에 관심이 있음.

# 극단값을 왜 생각할까?

- 출처: 이 땅, 통계학의 오늘 (1), 고려대학교 최종후 교수님

**i** 네덜란드 영토의 40%는 해수면보다도 낮은데 이는 제방 둑으로 보호되고 있다. 그러나 겨울철 불어오는 폭풍우는 해수면을 밀어 올리고 해변가에 위치한 제방 둑은 이를 견뎌 내야만 한다.

이를 위해 네덜란드 정부는 경비와 안정성을 모두 고려하여 연중 최대 해수면이 제방 둑을 넘칠 확률이 0.0001이 되도록 제방 둑의 높이를 정하고자 한다. 이때 사용할 수 있는 해수면 자료는 100년 남짓이라고 한다. 이 자료를 이용하여 해수면이 10,000년에 한번 정도 넘어설 정도의 제방 둑의 높이를 추정 할 수 있을까?

- 최대 100년 남짓 자료를 이용하여 10,000년에 한 번 정도 일어날 현상에 대한 모델링을 해야 함
- 주된 전략: 꼬리가 두꺼운 확률분포의 선택, 외삽 (extrapolation)

# 일반화 극단값 분포

- 확률변수  $Y_1, Y_2, \dots$  가  $Y_1, Y_2, \dots, \stackrel{\text{i.i.d.}}{\sim} F$  일 때 부분 최댓값 (**partial maximum**)  $M_n$  을 고려하자.

$$M_n = \max(Y_1, \dots, Y_n).$$

- 만약 어떤 상수열  $a_n > 0, b_n$  이 존재해  $n \rightarrow \infty$  일 때

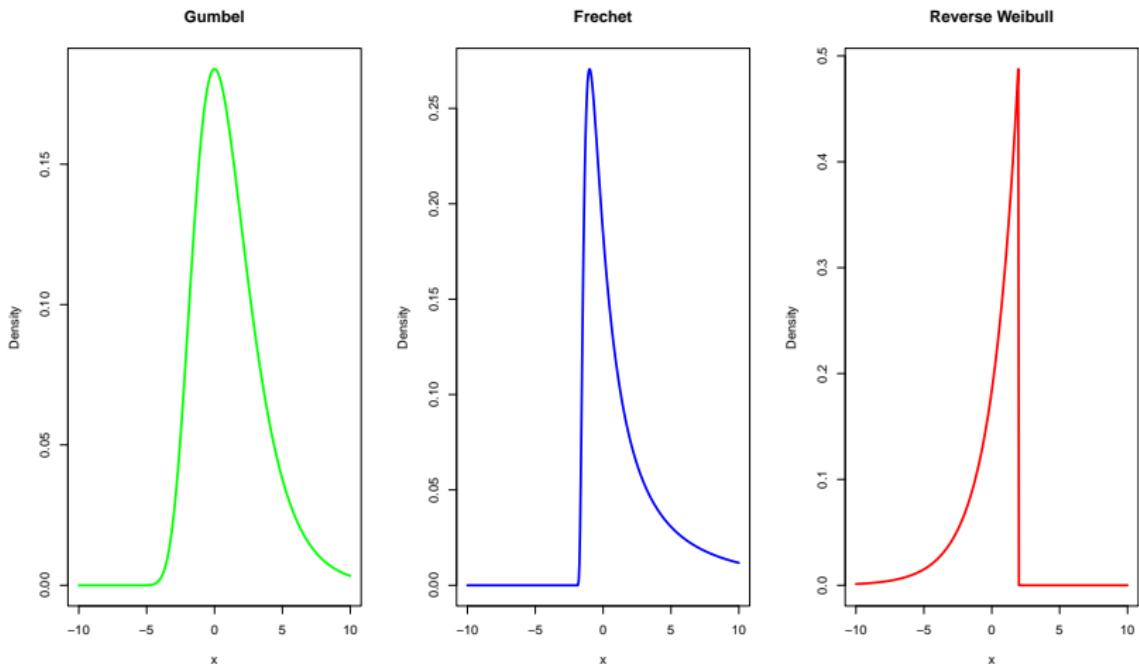
$$\frac{M_n - b_n}{a_n} \xrightarrow{D} Z \sim G,$$

이라고 하자. 이때  $G$  가 퇴화분포 (degenerate distribution) 가 아니라면,  $G$  의 형태는 다음과 같다.

$$G(y) = \begin{cases} \exp \left\{ - \left( 1 + \xi \frac{y-\mu}{\sigma} \right)_+^{-1/\xi} \right\}, & \xi \neq 0, \\ \exp \left\{ - \exp \left( - \frac{y-\mu}{\sigma} \right) \right\}, & \xi = 0 \end{cases},$$

with  $\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$ , defined in  $\{y : 1 + \xi(y - \mu)/\sigma > 0\}$ .

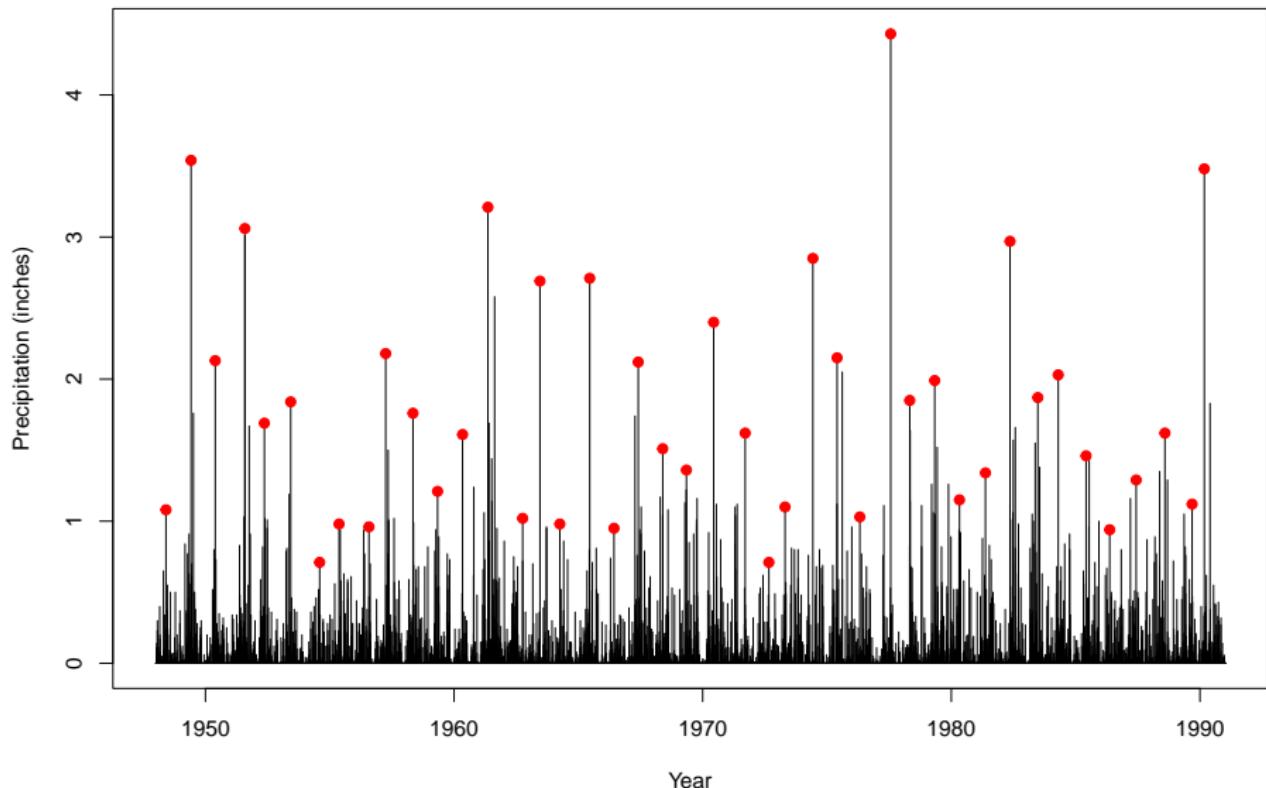
- 여기서  $\xi$ 는 극단값지수 (extreme value index) 라 하고, 꼬리의 두꺼움을 결정한다:
  - $\xi = 0$  (**Gumbel**, 얇은 꼬리)
  - $\xi > 0$  (**Frechet**, 두꺼운 꼬리)
  - $\xi < 0$  (**Weibull**, 유한한 값만을 갖는 꼬리)



**Figure 22:** 일반화 극단값 분포의 세 가지 타입.

# 예제: 포트 콜린스 일 강수량 자료 (R `extRemes` 패키지)

Precipitation at Fort Collins



- 출처: Cooley 의 2013 년 Joint Statistical Meeting 발표자료

**i** 1997 년 어느 날, 미국의 포트 콜린스에서는 4.63 인치 (약 118mm) 의 강수량이 관측되었다. 그런데, 1948 년부터 1990 년까지 포트 콜린스에서 관측된 최대 일 강수량은 4.43 인치 (약 113mm) 였다. 우리는 1948 년부터 1990 년 까지의 강수량 자료만 가지고 1997 년에 관측된 강수량이 이 지역에 대략 몇 년에 한 번 꼴로 발생할 강수량인지를 말하고 싶다. 이를 위해 다음 두 가지 방법을 고려하였다.

1. 자료를 꼬리가 얇은 확률분포인 감마분포로 모델링한 후, 추정된 분포를 바탕으로 4.63 인치를 넘어갈 확률을 구한다.
2. 자료를 꼬리가 두꺼운 확률분포인 일반화 극단값 (GEV) 분포로 모델링 한 후, 추정된 분포를 바탕으로 4.63 인치를 넘어갈 확률을 구한다.

## (1) 감마분포를 이용한 모델링

- $X_t$  를 포트 콜린스에서의 여름철 (4 월 ~10 월) 시간  $t$  에서의 일 강수량이라고 하자.
- $X_t$  는 감마확률변수를 따른다고 가정하자.  $X_t > 0$  이어야 하므로 비가 내리지 않은 날 (강수량이 0 인 날) 은 제외하고, 나머지 자료만 모아 감마분포로 모델링한다.
- 비가 올 확률을  $p$  라고 할 때,  $p = 0.2629$  였다.
- $X_t|X_t > 0$  가 감마확률변수를 따름을 가정하고, 즉  $[X_t|X_t > 0] \sim \text{Gamma}(\alpha, \beta)$  을 가정하고 최대가능도법으로  $\alpha, \beta$  를 추정하면  $\hat{\alpha} = 0.656, \hat{\beta} = 3.20$  이다.

- 그러면  $X_t > 4.63$  일 확률은

$$\begin{aligned}
 P(X_t > 4.63) &= P(X_t > 4.63 | X_t > 0)P(X_t > 0) \\
 &= (1 - F_X(4.63)) \times 0.263 \\
 &= (1 - 0.9999999) \times 0.263 = 3.01 \times 10^{-8}
 \end{aligned}$$

$$\begin{aligned}
 P(\text{연 최대 강수량} > 4.63) &= 1 - P(\text{모든 연 최대 강수량} < 4.63) \\
 &= 1 - (1 - P(\text{일 강수량} > 4.63))^{214} \\
 &= 1 - (1 - 3.01 \times 10^{-8})^{214} \\
 &= 6.441 \times 10^{-6}
 \end{aligned}$$

(일 강수량 관찰값이 모두 독립임을 가정하고, 1년 중 여름철에는 총 214일이 있다고 생각하고 문제를 푼다.)

- 복귀수준:  $(6.441 \times 10^{-6})^{-1} = 155,255$ 년마다 한 번 꼴로 일어날 것으로 기대되는 사건

## (2) GEV 분포를 이용한 모델링

- $M_n = \max_{t=1,\dots,n}(X_t)$  라고 하고 이것이  $M_n \sim \text{GEV}(\mu, \sigma, \xi)$  라고 가정하자.

$$F_{M_n}(x) = P(M_n \leq x) = \exp \left\{ - \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}$$

- 최대가능도로 추정된 모수는:  $\hat{\mu} = 1.384$ ,  $\hat{\sigma} = 0.574$ ,  $\hat{\xi} = 0.188$ .
- $P(\text{연 최대 강수량} > 4.63) = 1 - F_{M_n}(4.63) = 0.021$
- 복귀수준:  $(0.021)^{-1} = 47.6$  년마다 한 번 꼴로 일어날 것으로 기대되는 사건

## 질문: 포트 콜린스의 월 최대 강수량 패턴은 변했을까?

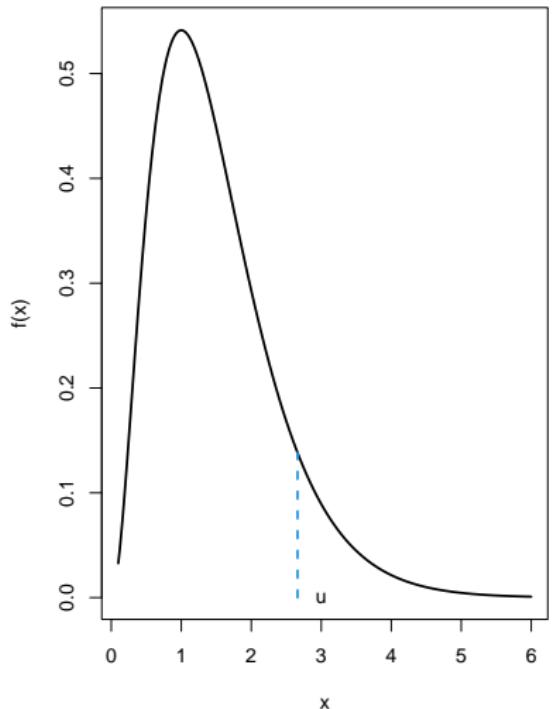
- 2024년까지의 포트 콜린스의 월별 일 최대 강수량 자료를 National Oceanic and Atmospheric Administration에서 추가로 수집
- 똑같이 43년간의 자료 (1982-2024년)를 이용해 GEV 분포 모델링
- 최대가능도로 추정된 모수는:  $\hat{\mu} = 0.285$ ,  $\hat{\sigma} = 0.261$ ,  $\hat{\xi} = 0.421$ .
- $P(\text{연 최대 강수량} > 4.63) = 1 - F_{M_n}(4.63) = 0.007$
- 복귀수준: 141.1년마다 한 번 꼴로 일어날 것으로 기대되는 사건
- 이를 통해 포트 콜린스의 월 최대 강수량의 강도가 점점 줄고 있음을 짐작 가능

• 출처: National Oceanic and Atmospheric Administration

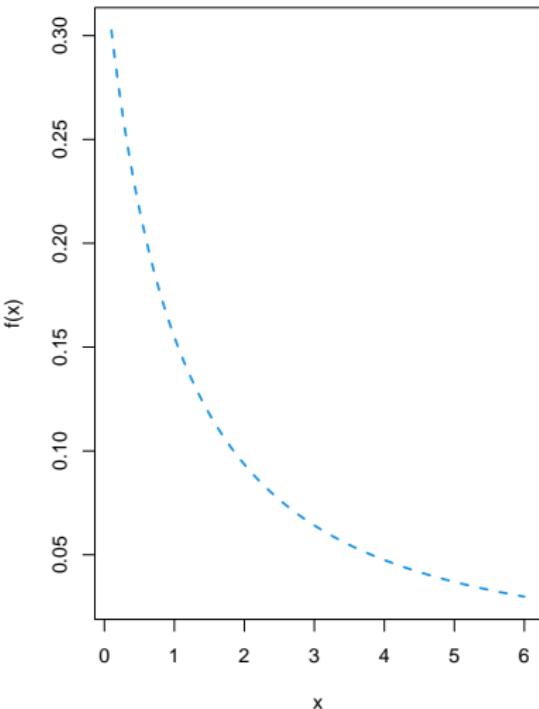
NOWData - NOAA Online Weather Data														<a href="#">Enlarge results</a>	<a href="#">Print</a>	
2002	0.24	0.16	0.53	0.13	1.24	0.40	0.05	0.29	0.51	0.37	0.22	0.01	1.24			
2003	0.03	0.23	3.30	1.06	1.63	0.22	0.33	2.49	0.15	0.08	0.16	0.27	3.30			
2004	0.16	0.28	0.28	0.58	0.39	0.85	0.66	1.06	0.81	0.41	0.72	0.05	1.06			
2005	0.21	0.20	0.59	0.55	1.50	0.97	0.18	0.42	0.12	1.18	0.05	0.10	1.50			
2006	0.06	0.29	0.69	0.08	0.29	0.05	0.44	0.23	0.48	0.91	0.29	1.30	1.30			
2007	0.22	0.07	1.04	0.54	0.38	0.19	0.38	2.19	0.77	0.83	0.37	0.42	2.19			
2008	0.03	0.16	0.27	0.38	0.45	1.75	0.54	1.55	0.78	0.49	0.01	0.34	1.75			
2009	0.13	0.09	0.77	1.70	0.62	1.07	2.10	0.10	0.26	1.20	0.52	0.51	2.10			
2010	0.08	0.17	0.78	1.14	0.92	0.98	0.52	0.54	0.04	0.29	0.34	0.17	1.14			
2011	0.21	0.26	0.12	0.90	1.18	1.72	0.67	0.06	1.03	1.07	0.75	0.32	1.72			
2012	0.04	0.51	T	0.11	0.55	0.24	0.93	0.02	1.92	0.17	0.14	0.22	1.92			
2013	0.09	0.31	0.39	0.79	1.51	0.48	0.56	0.51	1.99	0.49	0.21	0.34	1.99			
2014	0.64	0.12	0.53	0.55	1.97	0.49	0.96	0.13	0.36	0.59	0.57	0.33	1.97			
2015	0.06	0.38	0.08	1.62	1.55	0.33	0.42	0.25	0.13	0.72	0.64	0.50	1.62			
2016	0.45	0.49	1.44	0.73	0.67	0.02	0.27	0.19	0.12	0.13	0.22	0.20	1.44			
2017	0.35	0.43	0.50	0.65	2.77	0.09	0.33	0.74	0.89	0.98	0.38	0.09	2.77			
2018	0.38	0.25	0.31	0.42	1.74	0.43	0.74	0.10	0.41	0.51	0.27	0.02	1.74			
2019	0.27	0.12	1.03	0.66	0.71	0.74	0.30	0.13	0.18	0.23	1.00	0.23	1.03			
2020	T	0.16	0.90	1.30	0.74	0.57	0.06	0.30	1.19	1.03	0.37	0.19	1.30			
2021	0.14	0.24	1.78	0.55	1.45	0.29	0.22	0.34	0.29	0.11	0.15	0.49	1.78			
2022	0.26	0.48	0.31	0.09	0.70	0.37	1.97	0.22	0.52	0.21	0.29	0.10	1.97			
2023	0.47	0.33	0.52	0.39	0.91	1.98	1.21	2.45	0.31	0.30	0.26	0.13	2.45			
2024	0.14	1.63	0.93	1.11	0.27	0.14	0.36	0.90	0.27	0.26	0.53	T	1.63			
<b>Mean</b>	0.21	0.28	0.73	0.77	0.97	0.71	0.78	0.72	0.58	0.53	0.36	0.22	1.79			
<b>Max</b>	0.64	1.63	3.30	2.41	2.77	2.49	4.63	2.49	1.99	1.20	1.00	1.30	4.63			
	2014	2024	2003	1999	2017	1992	1997	2003	2013	2009	2019	2006	1997			
<b>Min</b>	T	T	T	0.08	0.27	0.02	0.05	0.02	0.01	0.08	0.01	T	0.82			
	2020	1992	2012	2006	2024	2016	2002	2012	1992	2003	2008	2024	2000			

# Peak-over threshold 방법

(a) Probability Distribution F



(b) Extreme Value Distribution (GPD)



# Generalized Pareto distribution (GPD)

- $y > 0$ 에서 high threshold exceedances는 일반화 파레토분포 (generalized Pareto distribution)로 묘사 가능 (Coles 2001),

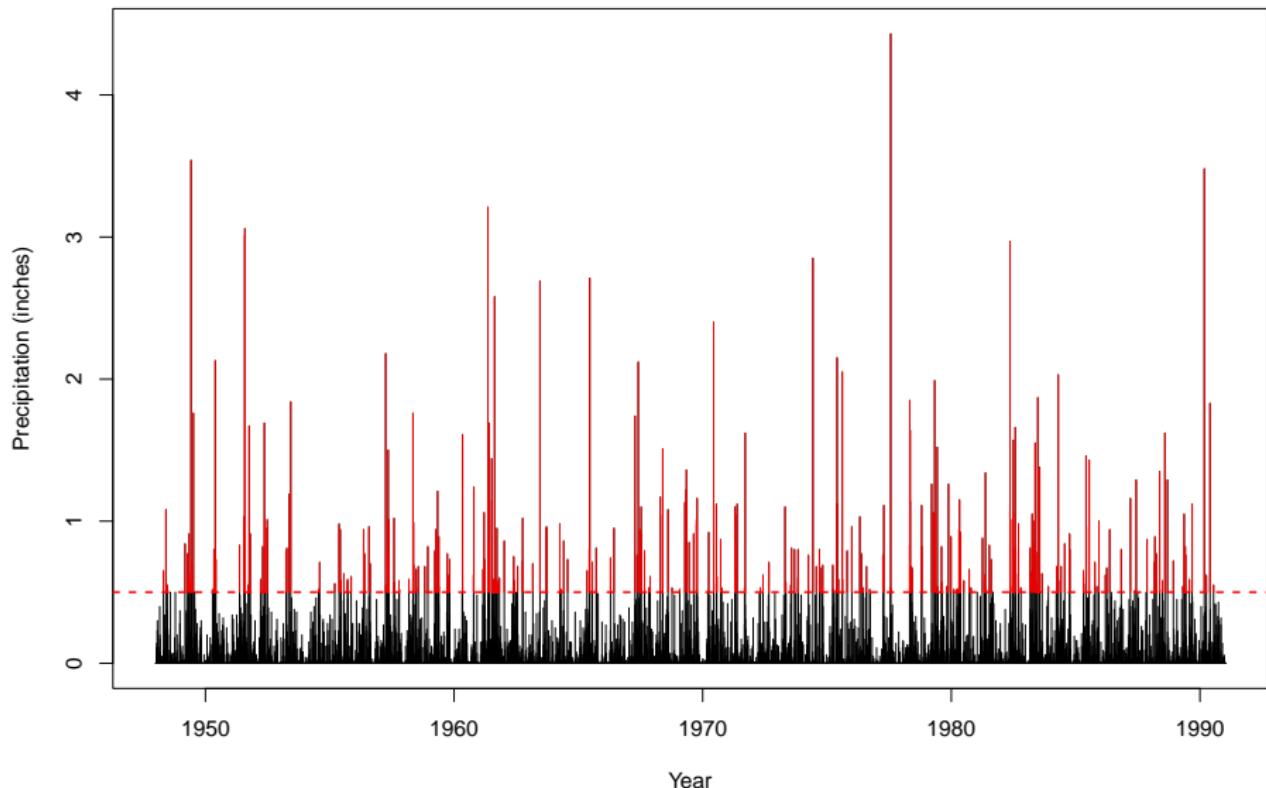
$$\begin{aligned} P\left(\frac{Y - b_n}{a_n} > u + y \mid \frac{Y - b_n}{a_n} > u\right) &= \frac{P\{(Y - b_n)/a_n > u + y\}}{P\{(Y - b_n)/a_n > u\}} \\ &\approx \frac{\{1 + \xi(y + u - \mu)/\sigma\}_+^{-1/\xi}}{\{1 + \xi(u - \mu)/\sigma\}_+^{-1/\xi}} \\ &= (1 + \xi y / \tilde{\sigma})_+^{-1/\xi}, \end{aligned}$$

이때  $\tilde{\sigma} = \sigma + \xi(u - \mu) > 0$ 이다.

- 1에서 우변의 값을 뺀 값, 즉  $H(y) = 1 - (1 + \xi y / \tilde{\sigma})_+^{-1/\xi}$ 는 일반화 파레토분포의 누적분포함수이다.

# 예제: 포트 콜린스 일 강수량 자료

Precipitation at Fort Collins



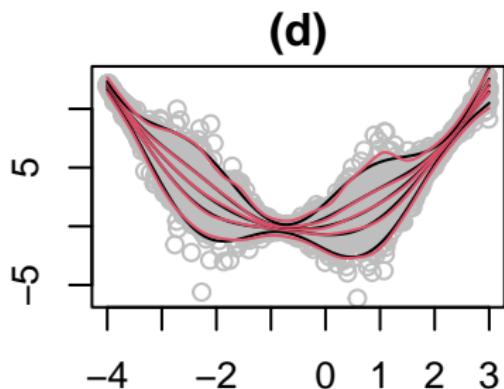
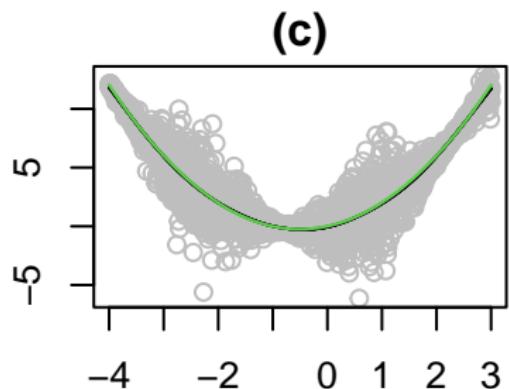
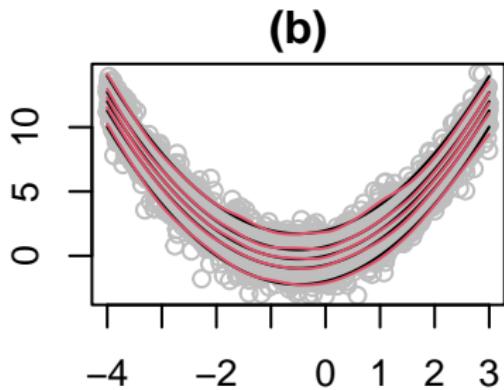
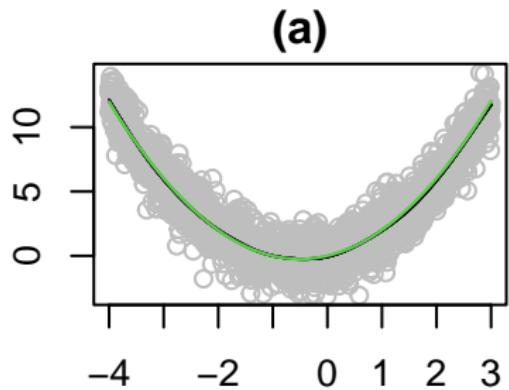
# GEV vs GPD

GEV	GPD
$n = 43$	$u = 0.5$
$\hat{\mu} = 1.38(0.011)$	$n = 309$
$\hat{\sigma} = 0.57(0.007)$	$\hat{\beta} = 0.33(0.001)$
$\hat{\xi} = 0.19(0.029)$	$\hat{\xi} = 0.33(0.005)$
$\hat{r}_{100} = 5.58 \text{ inches}$	$\hat{r}_{100} = 5.80 \text{ inches}$
$95\% \text{ CI} = (2.13, 9.03)$	$95\% \text{ CI} = (3.30, 8.30)$

- GPD 가 GEV 보다  $\hat{\xi}$  의 표준오차가 작고, 복귀수준의 신뢰구간도 더 좁음

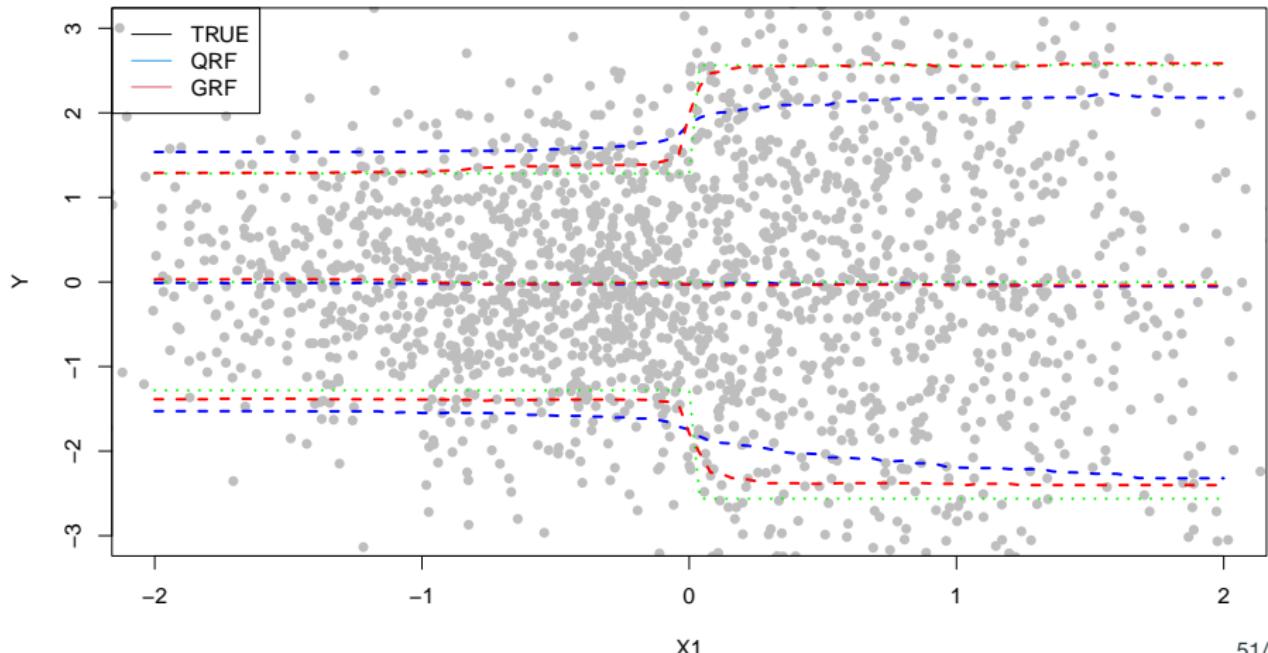
# 분위수 가법모형

- 분위수 가법모형: (Fasiolo et al. 2020)



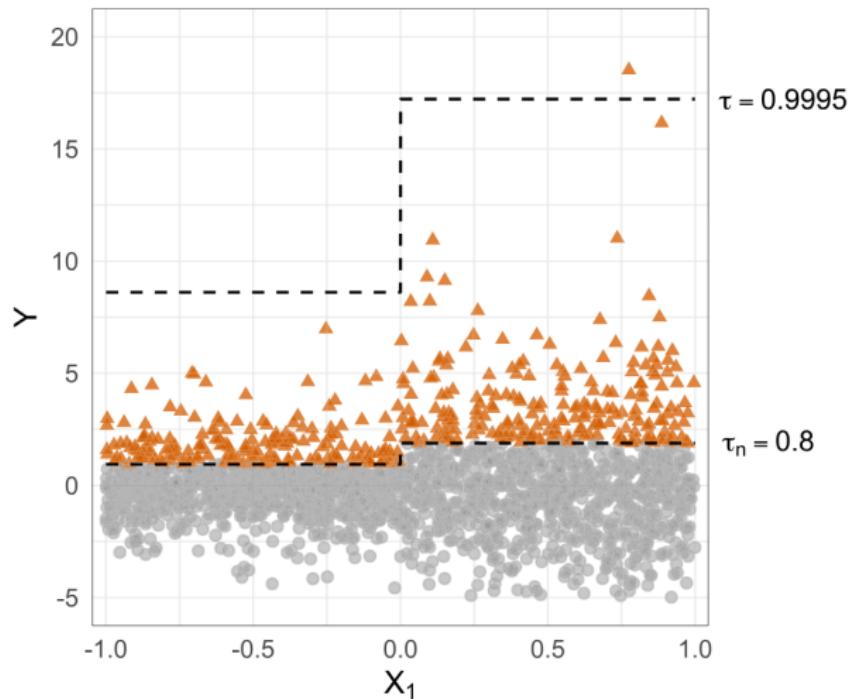
# 분위수, 일반화 회귀 포레스트

- 분위수 회귀 포레스트: (Meinshausen 2006)
- 분위수 회귀 포레스트의 응용: (Park et al. 2018)
- 일반화 회귀 포레스트: (Athey, Tibshirani, and Wager 2019)

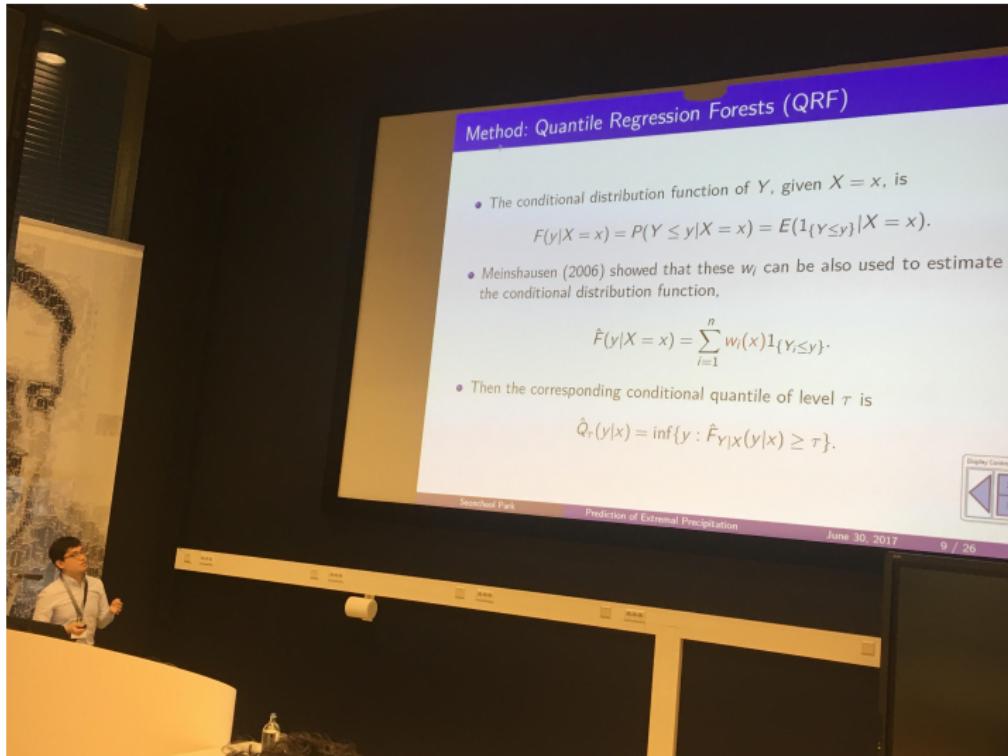


# 극단 랜덤 포레스트

- 극단 랜덤 포레스트 (Gnecco, Terefe, and Engelke 2024)



# Extreme value analysis (EVA) conference



## 참고문헌

- Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician*, **27(1)**, 17–21.
- Athey, S., Tibshirani, J., and Wagner, S. (2019). Generalized Random Forests. *The Annals of Statistics*, **47(2)**, 1148–1178.
- Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London.
- Cooley, D. (2013). *Introduction to Analysis of Extremes: Univariate and Multivariate Cases*. Joint Statistical Meetings, Montreal, Canada; August 2, 2013. (Full-day short course)
- Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Yanning, G. (2020). Fast Calibrated Additive Quantile Regression. *Journal of the American Statistical Association*, **116(535)**, 1402–1412.
- Gnecco, N., Terefe, E. M., and Engelke, S. (2024). Extremal Random Forests. *Journal of the American Statistical Association*, **119(548)**, 3059–3072.

## 참고문헌

- Hadi, A. S., and Chatterjee, S. (2023). *Regression Analysis By Example Using R, 6th Edition*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.
- Jung, S. (2022). *Nonparametric Statistics with R*. Freeacademy.
- Marrona, R. A, Martin, R. D., Yohai, V. J, and Salibián-Barrera, M. (2018). *Robust Statistics: Theory and Methods (with R)*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.
- Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research*, **7(35)**, 983–999.
- Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *The New England Journal of Medicine*, **367(16)**, 1562–1564.
- Park, S., Kwon, J., Kim, J., and Oh, H.-S. (2018). Prediction of Extremal Precipitation by Quantile Regression Forests: From SNU Multiscale Team. *Extremes*, **21(3)**, 463–476.
- Stephens, E., and Cryle, P. (2017). Eugenics and the Normal Body: The Role of Visual Images and Intelligence Testing in Framing the Treatment of People with Disabilities in the Early Twentieth Century. *Continuum*, **31(3)**, 365–376.