한국통계학회 공개강연

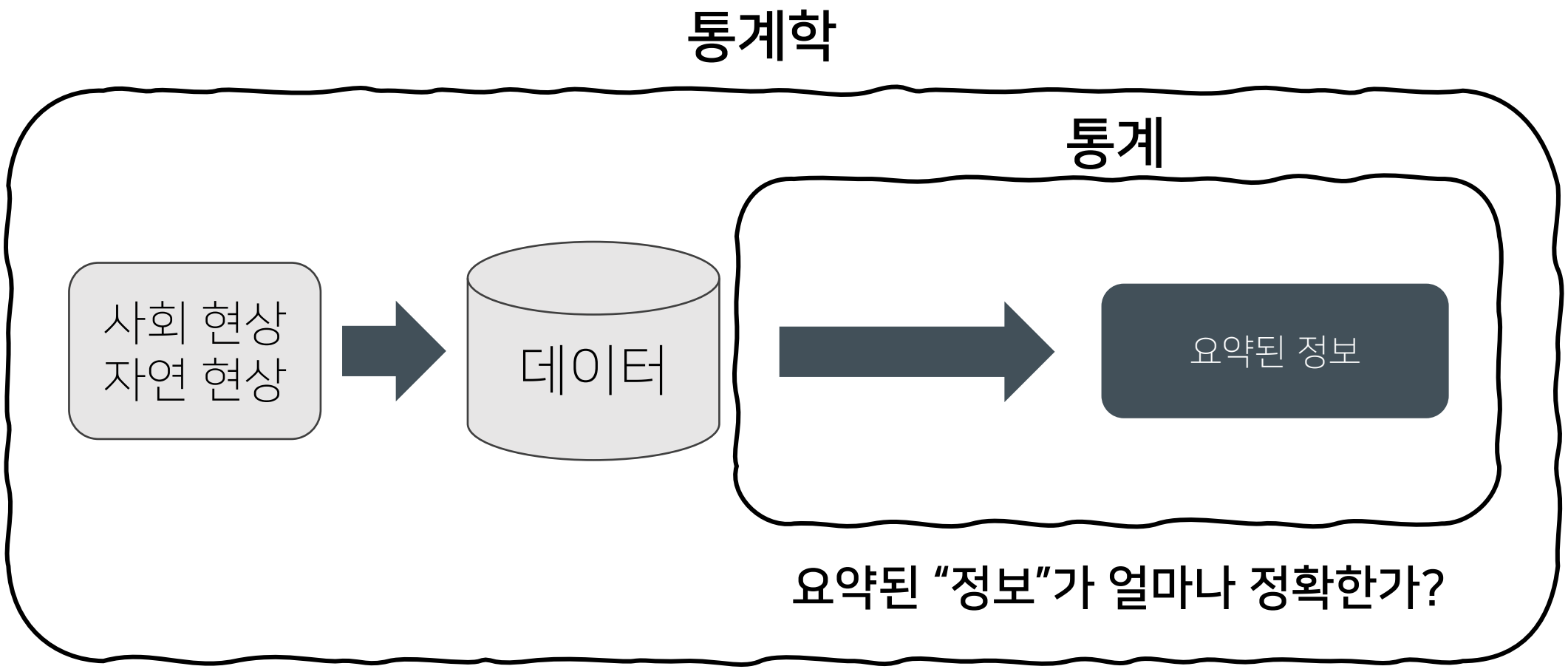
평균 낼 수 없는 것들의 통계

정 성 규
서울대학교 통계학과

Part 1,

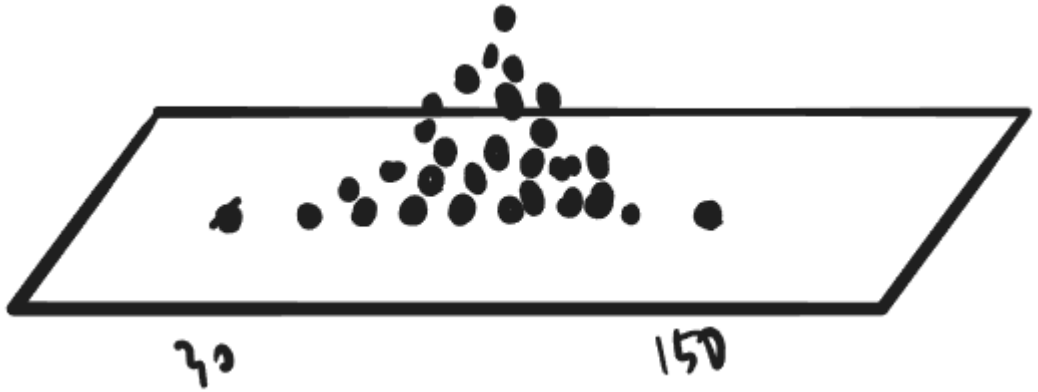
통계와 평균

통계?



평균

90



평균 70

≈ 기준, 대푯값

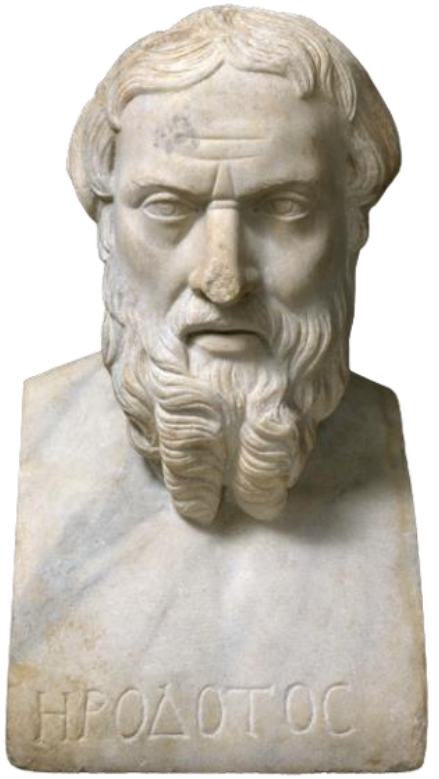
Part 1,

평균과 미지수의 추정

페르미 추정 문제:

Q: 이집트 왕조는 341세대가 있다. 이 왕조는 몇 년이나 지속되었을까?

A: 세 세대당 “평균” 100년 지속된다면, 11340년이다.*



헤로도토스 (BC 484 – BC 425)

* Rubin, E. (1968). STATISTICAL WORLD OF HERODOTUS. *American Statistician*, *22*(1), 31.

평균과 미지수의 추정

미지의 참값 추정 문제:

빛의 속도 측정 실험: 29.6만km/s, 30만km/s, 30.1만km/s

빛의 속도 추정값은? 중앙값 30만km/s?

평균 29.9만km/s?

가우스 “관측오차의 제곱합이 가장 작은 추정값”

평균 29.9 = $\frac{29.6+30+30.1}{3}$

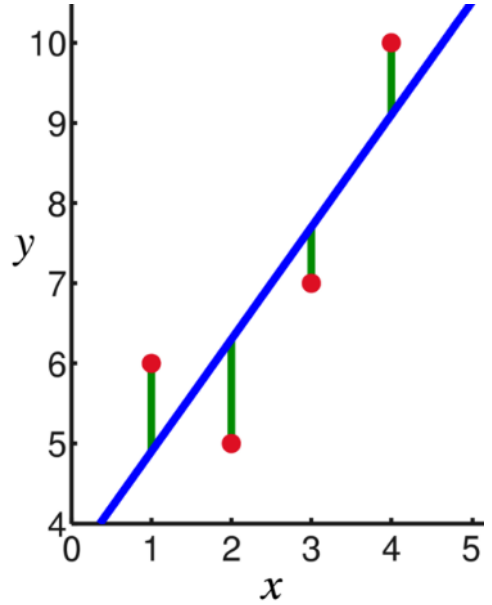
$= argmin_{\mu}\,(29.6-\mu)^2+(30\ -\mu)^2+(30.1-\mu)^2$

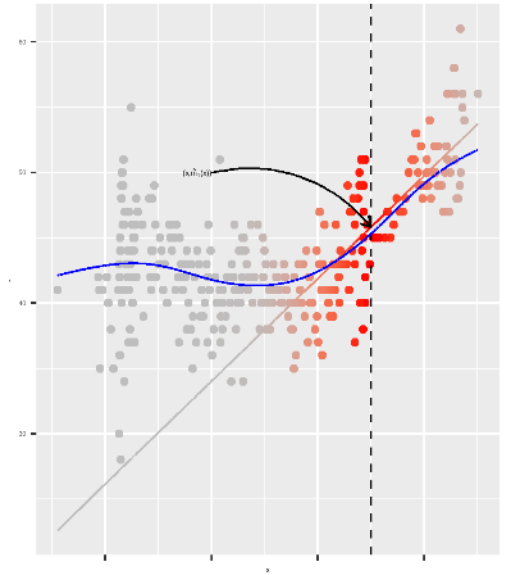
카를 프리드리히 가우스 (1777-1855)

Part 1,

평균과 통계

- 가우스의 최소제곱추정법: 평균 = 최소제곱법의 해.
- 선형회귀분석: 최소제곱추정된 계수 = $(x^T x)^{-1} x^T y$
  = (가중)평균
- 비모수회귀분석 (국소다항회귀, 스플라인 평활):
  최소제곱추정된 회귀곡선 $\hat{m}(x) = \sum_{i=1}^{n} w_i(x)\, y_i$
  = (가중)평균





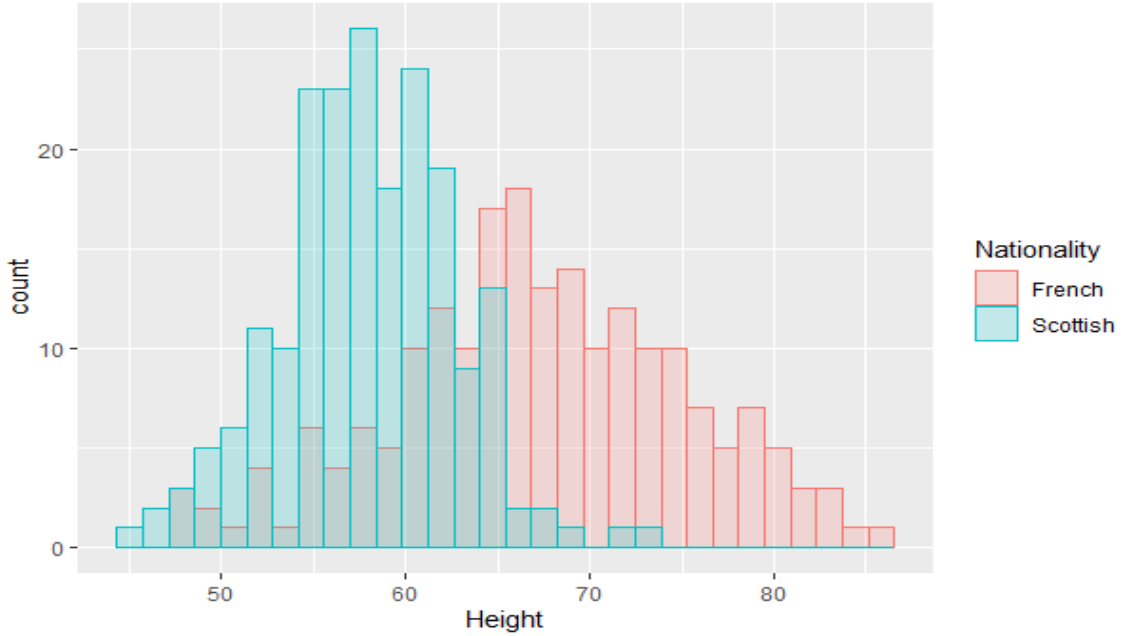
평균 인간

"프랑스군 신병"

Vs

"스코틀랜드군 신병"

인간과 그 재능 발달에 대한 연구 (1842)



아돌프 케틀레
1796-1874

평균 인간

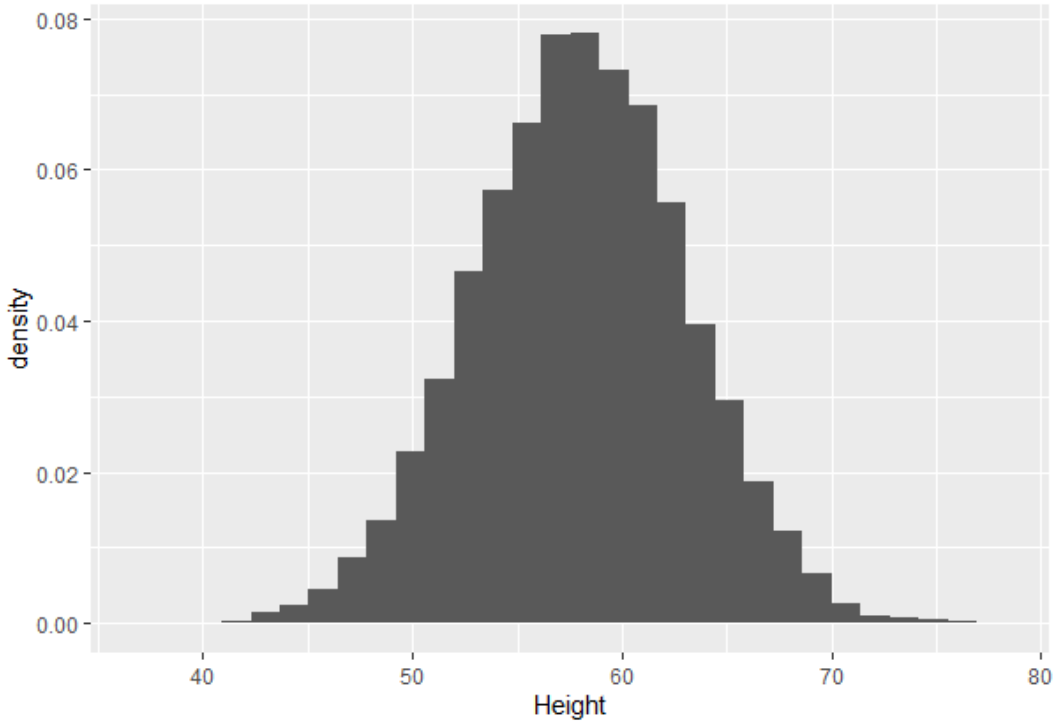
- **“평균 인간”** : 육체적·정신적으로 한 집단의 무게중심에 해당하는 사람,

  즉 집단의 대표 Stereotype이자 신이 의도한 이상

"평균적인 프랑스군 신병" vs "평균적인 스코틀랜드군 신병“

Part 1,

평균 인간은 존재하는가?



- 키가 평균(58-60인치)인 사람의 비율 = 1/5
- 몸무게가 평균인 사람의 비율 = 1/5
- 특성 두 개가 모두 평균인 사람의 비율 = 1/5 x 1/5 = 1/25
- 특성 4개가 평균인 사람의 비율?

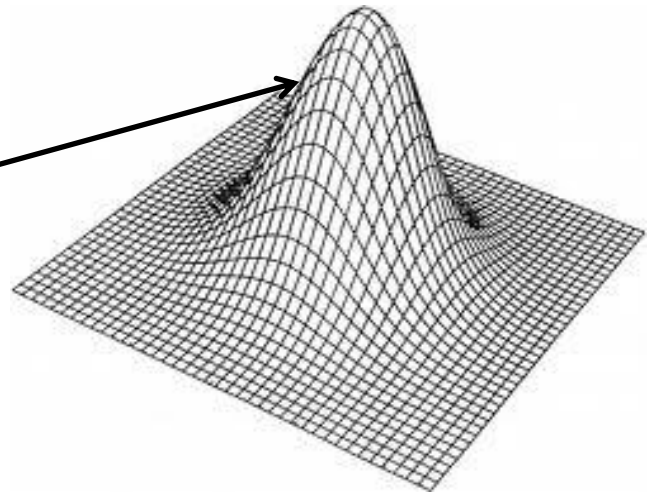
고차원 데이터와 평균

고차원 데이터

$$\underline{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} \sim N_d(\mu, I_d)$$

차원 d가 매우 클 때, 데이터가 어디 있을까?

밀도함수의 꼭대기?



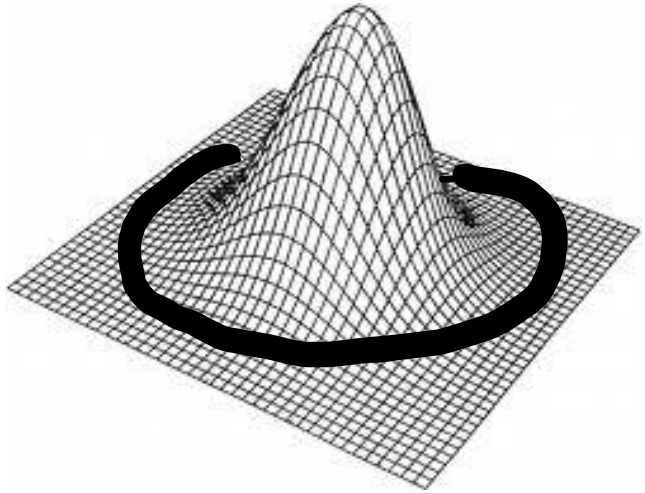
psycnet.apa.org

고차원 데이터의 역설

차원 d가 늘어날 때,

$$\|Z - \mu\| = \sqrt{d} + O_p(1)$$

- 데이터는 (반지름이 $\sqrt{d}$인) 구체의 표면 위에 존재
- 구체의 중심 (“평균”) 은 여전히 가장 높은 밀도!



- 역설의 해결: 르벡 측도와 비교한 밀도함수

- “평균인간은 없다”의 수학적 증명?

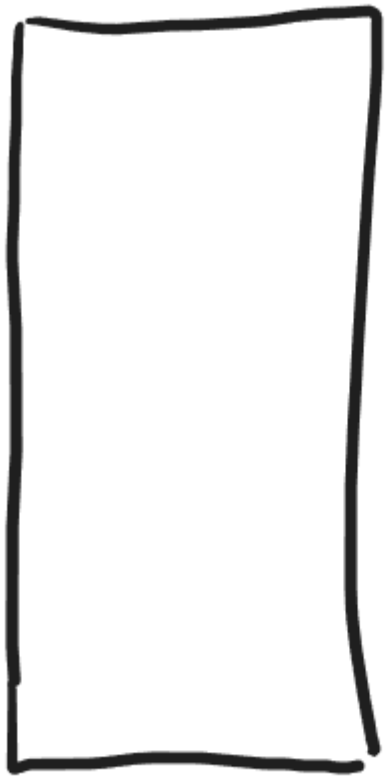
Part 2,

평균낼수없는것들

Part 2,

2007년 여름...

J. S. Marron



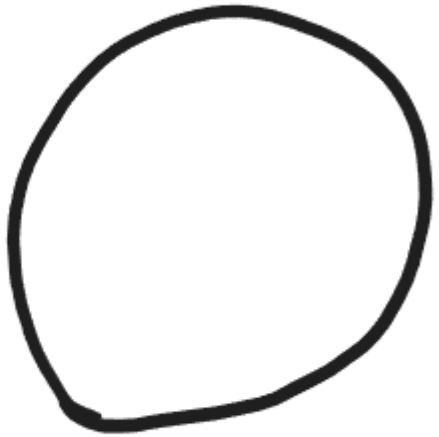
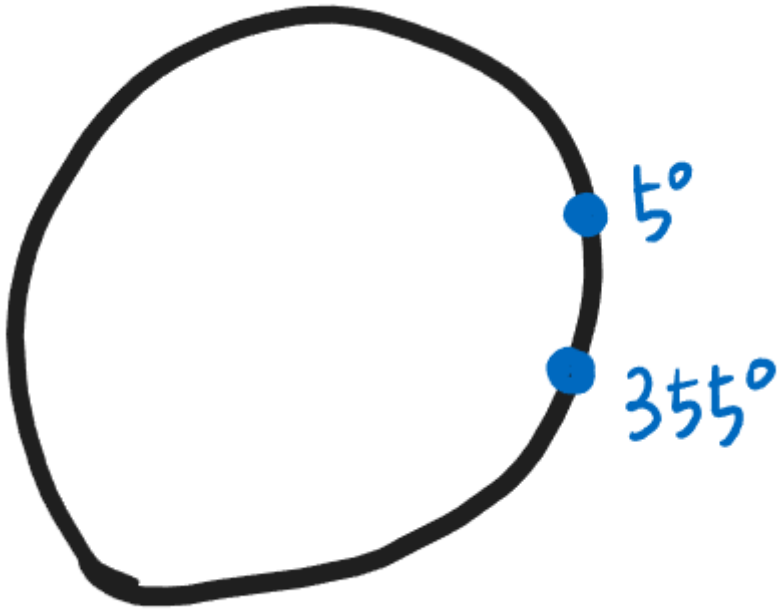


Image: https://marron.web.unc.edu/

Part 2,

방향 데이터와 평균

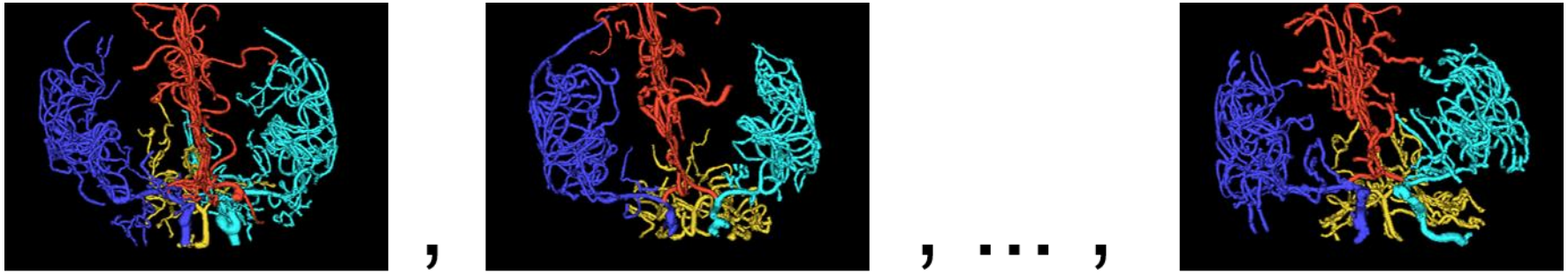


산술평균 ≠ 대표값

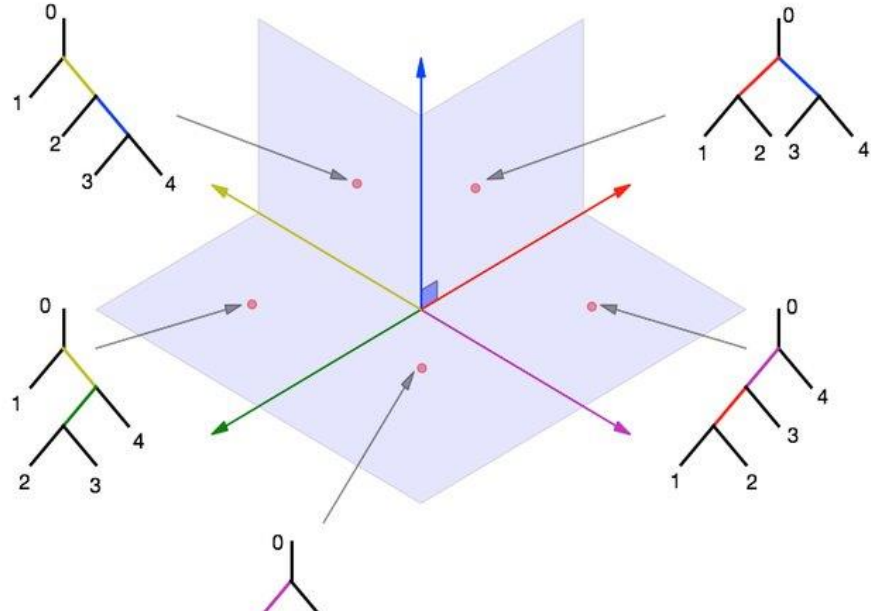
Part 2,

트리 데이터와 평균

- Brain artery trees (Bendich et al. 2016)



- Phylogenetic trees (Dinh et al. 2018)

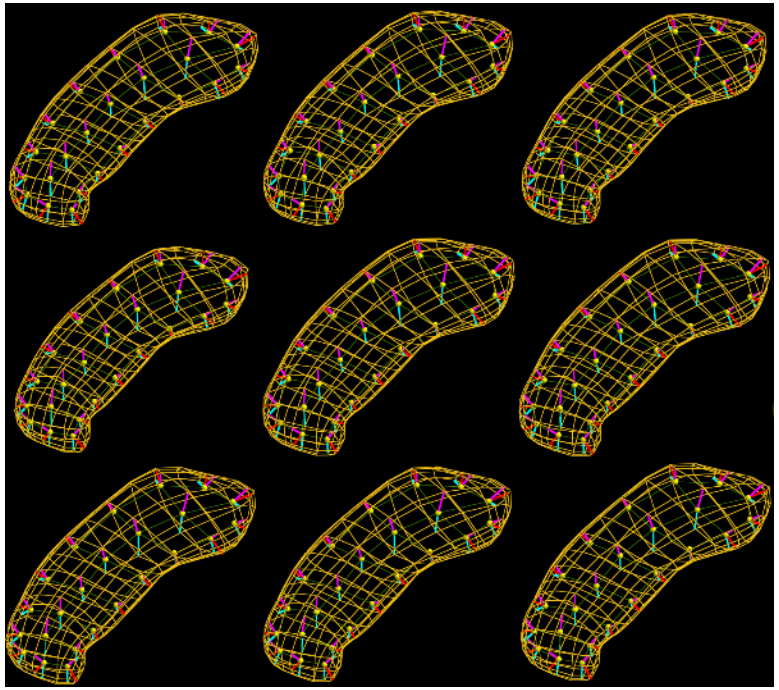


Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016), "Persistent homology analysis of brain artery trees," *Ann. Appl. Stat.*, 10, 198–218.
Dinh, V., Ho, L. S. T., Suchard, M. A., & Matsen IV, F. A. (2018). Consistency and convergence rate of phylogenetic inference via regularization. *Annals of statistics*, *46*(4), 1481.

Part 2,

기하적 특성을 가진 데이터와 평균

- Shapes (Pizer et al. 2013)



Stephen M. Pizer, Sungkyu Jung, Dibyendusekhar Goswami, Xiaojie Zhao, Ritwik Chaudhuri, James N. Damon, Stephan Huckemann, J. S. Marron (2013). "Nested Sphere Statistics of Skeletal Models", in *Innovations for Shape Analysis: Models and Algorithms*. M. Breus, Bruckstein and Maragos (Eds), Springer, Berlin, 93-115.

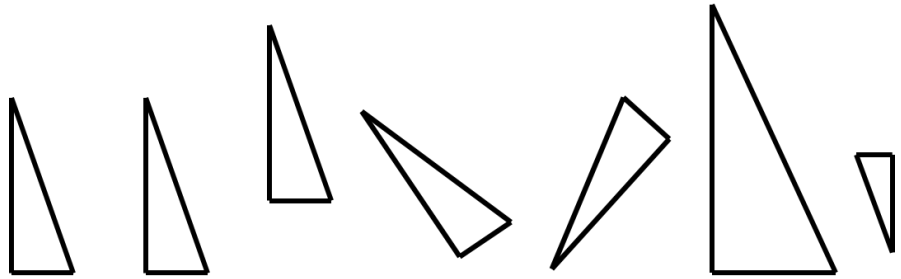
- Diffusions (Groisser et al. 2021)

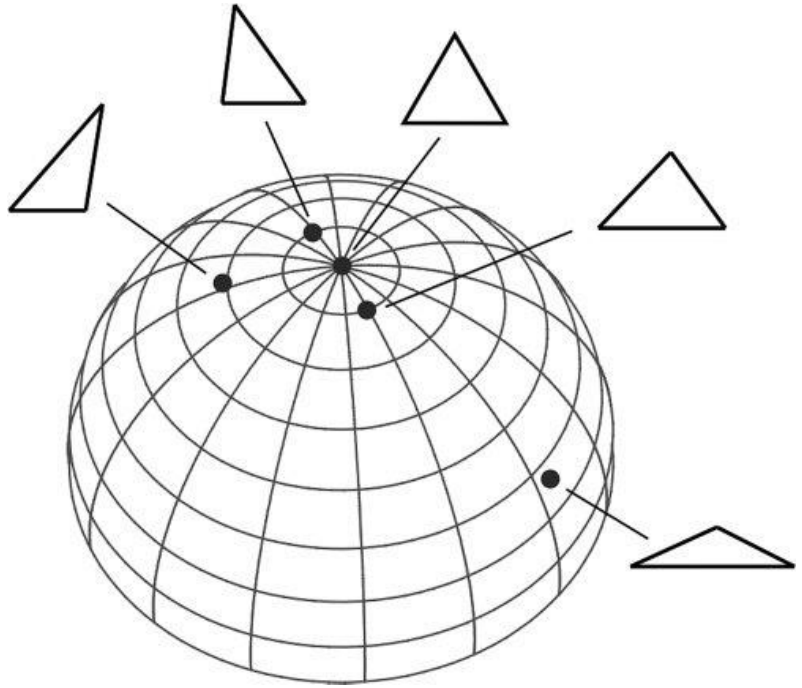
David Groisser, Sungkyu Jung and Armin Schwartzman (2021). "Uniqueness questions in a scaling-rotation geometry on the space of symmetric positive-definite matrices." *Differential Geometry and its Applications*. Volume 79, December, 101798

매니폴드 위의 데이터와 평균

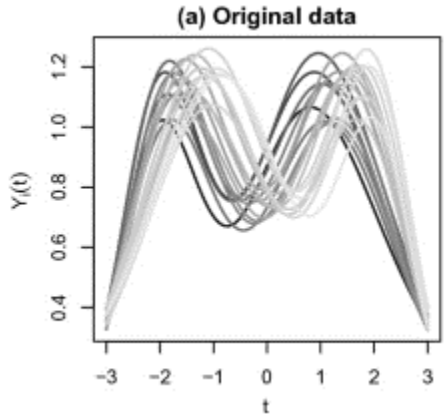
- 표본 공간 = 매니폴드 (다양체)
  (분석 대상의 특성을 반영하는 기하적인 제약이 있는 공간)

Shapes





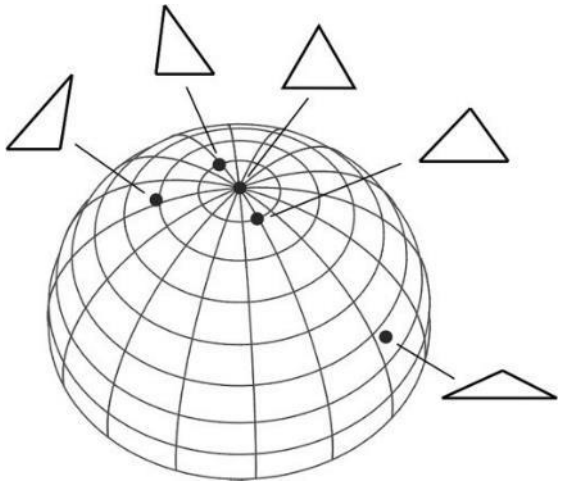
Distributions



Part 2,

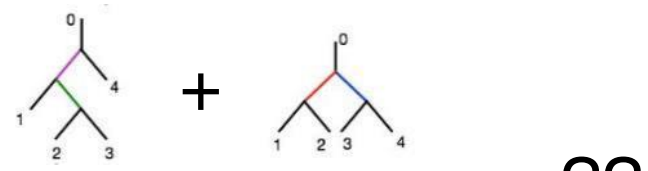
평균 낼 수 없는 것들

산술평균 ∉ 표본공간

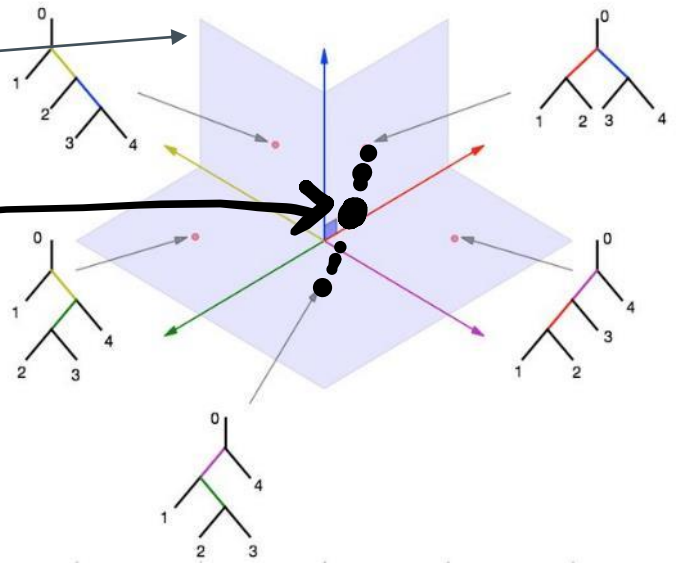


- 삼각형들 = 길이가 1인 벡터들 ∈ S^2
  평균의 길이 ≠ 1

- 트리들 ∈ BHV tree space*



$$\frac{\quad + \quad}{2} = ??$$



* Billera L, Holmes S, Vogtman K. Geometry of the space of phylogenetic trees. *Adv Appl Math*. 2001;27:733–767.

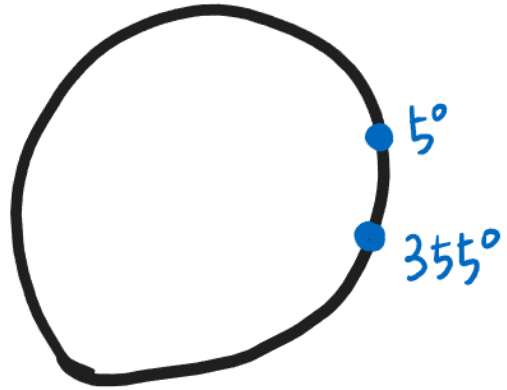
Part 2,

평균 낼 수 없는 것들

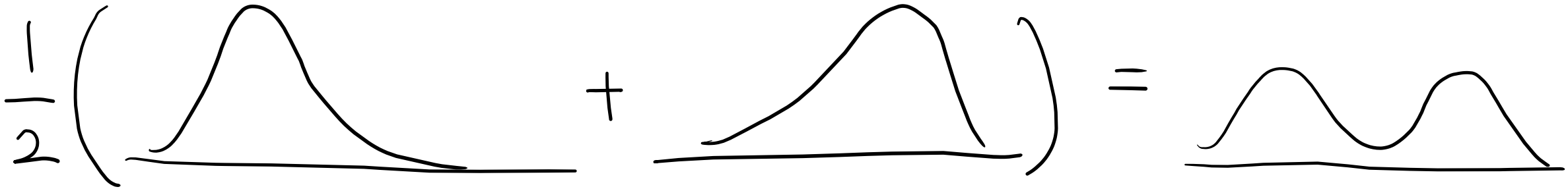
산술평균 ≠ 대표값

- 각도들 $\in [0, 360)$

$$\text{평균} = \frac{5^\circ + 355^\circ}{2} = 180^\circ \neq \text{대표값 } (0^\circ)$$



- 실수 값을 가지는 분포들 = 밀도함수들 $\in F$



Part 2,

프레쉐 평균 (Fréchet mean)

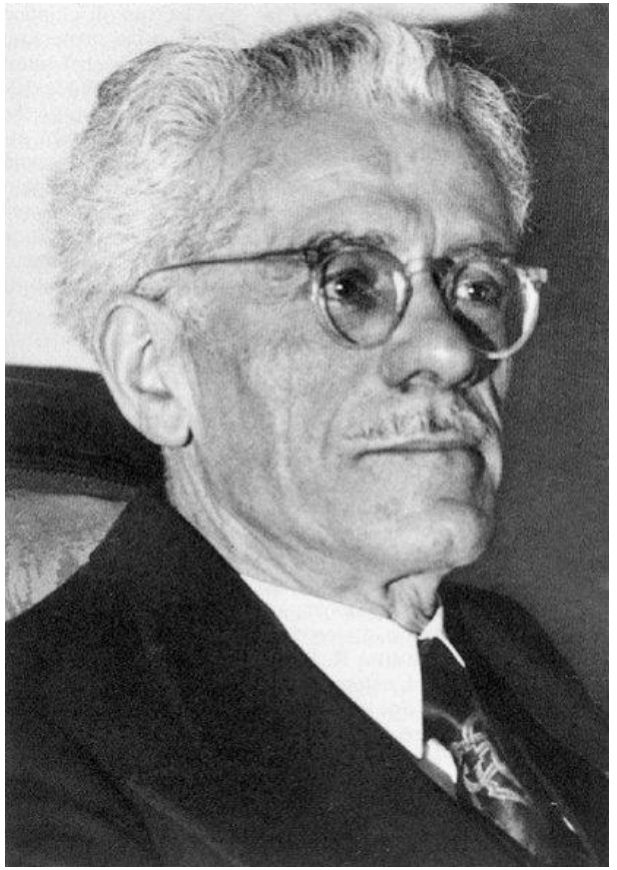
- 표본공간 M
  - M = R^p, S^p , BHV tree space, 등등…
- 표본공간 위의 거리 d(·,·): M x M →[0, ∞)

"가우스의 최소제곱법"

만약 M이 벡터공간 (R) 이라면, $x_1, \cdots, x_n \in R$ 의 평균: $\bar{x} = \arg\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2$

- $x_1, \cdots, x_n \in (M, d)$ 의 프레쉐 평균:

$$\bar{x} \equiv \arg\min_{\mu \in M} \sum_{i=1}^{n} d(x_i, \mu)^2$$

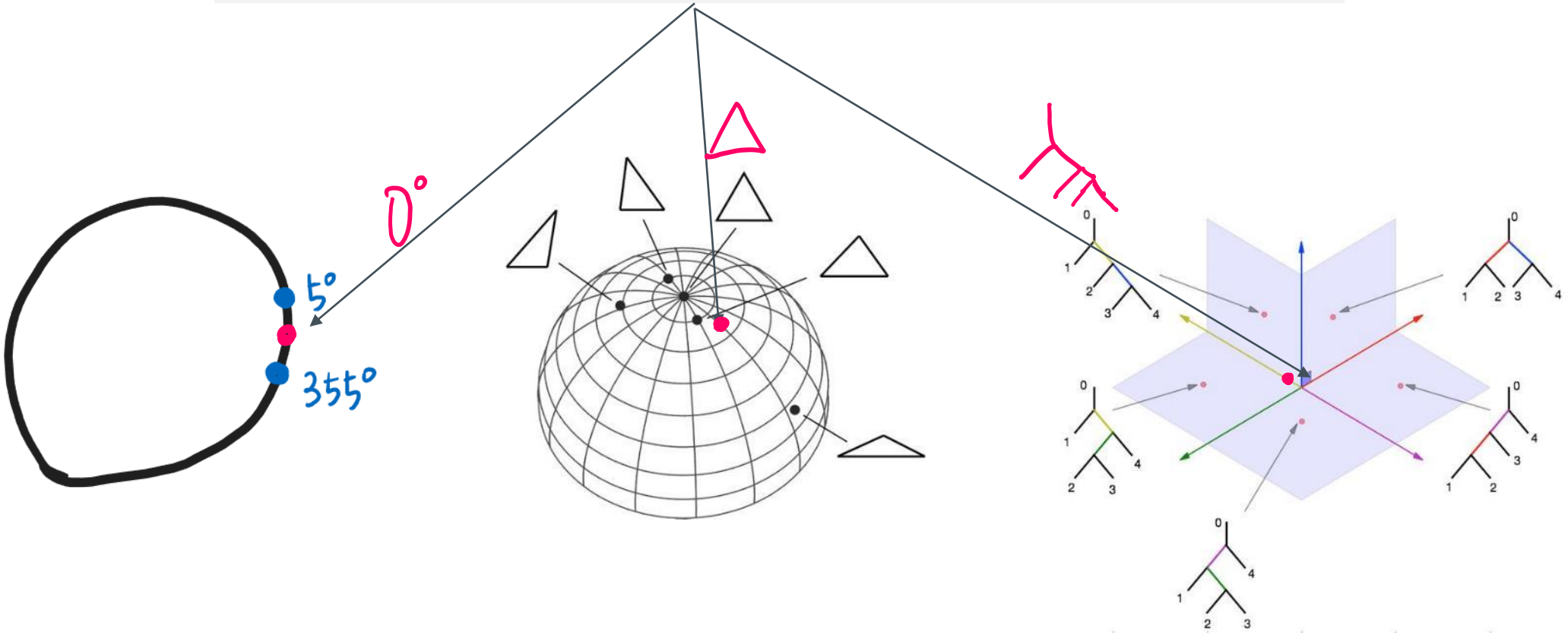


Maurice René Fréchet (1878-1973)

Part 2,

프레쉐 평균

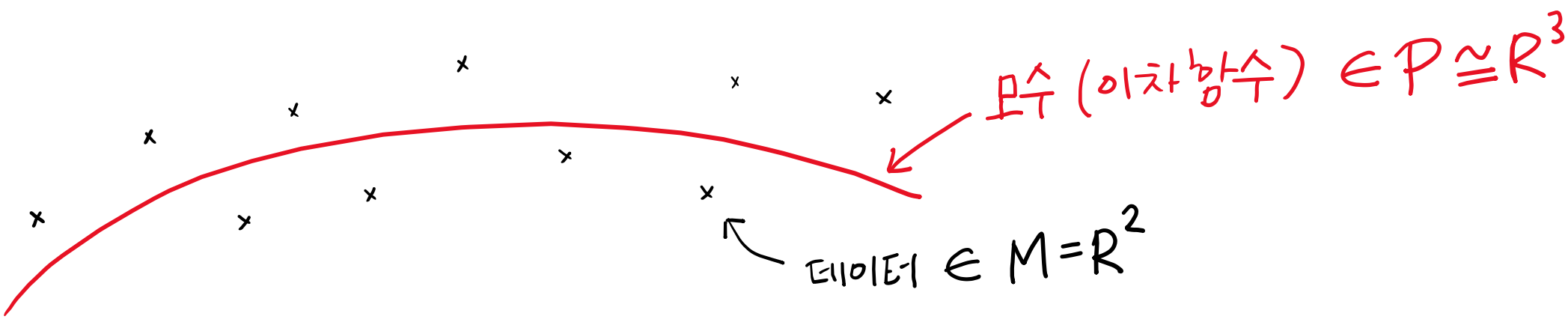
프레쉐 평균 = 오차제곱합을 최소로 하는 점



Part 2,

일반화된 프레쉐 평균*

- 표본공간 M
  - M = R^p, S^p , BHV tree space, 등등…
- 모수공간 P "모수 = 분포를 기술하는 값"
  - 모수가 프레쉐 평균: P = M
  - 모수가 이차회귀곡선: 표본공간 M = R^2 = {(x,y)}
    모수공간 P = {이차함수들} = {(a,b,c) ∈ R^3}



* HUCKEMANN, S. (2011) Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *Ann. Statist*. 39 1098-1124.

Part 2,

일반화된 프레쉐 평균*

- 표본공간 M
  - M = R^p, S^p , BHV tree space, 등등…
- 모수공간 P "모수 = 분포를 기술하는 값"
  - 모수가 프레쉐 평균: P = M
  - 모수가 이차회귀곡선: 표본공간 M = R^2 = {(x,y)}
    모수공간 P = {이차함수들} = {(a,b,c) ∈ R^3}
- 모수가 분포를 얼마나 잘 기술하는지에 대한 측도

$$\rho(\cdot,\cdot): P \times M \rightarrow [0, \infty)$$

예: 이차회귀곡선의 최소제곱추정: $\rho((a,b,c), (x,y)) = |y - (a+bx+cx^2)|$

$$\arg\min_{a,b,c} \sum_{i=1}^{n} \rho\big((a,b,c),(x_i,y_i)\big)^2$$

* HUCKEMANN, S. (2011) Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *Ann. Statist*. 39 1098-1124.

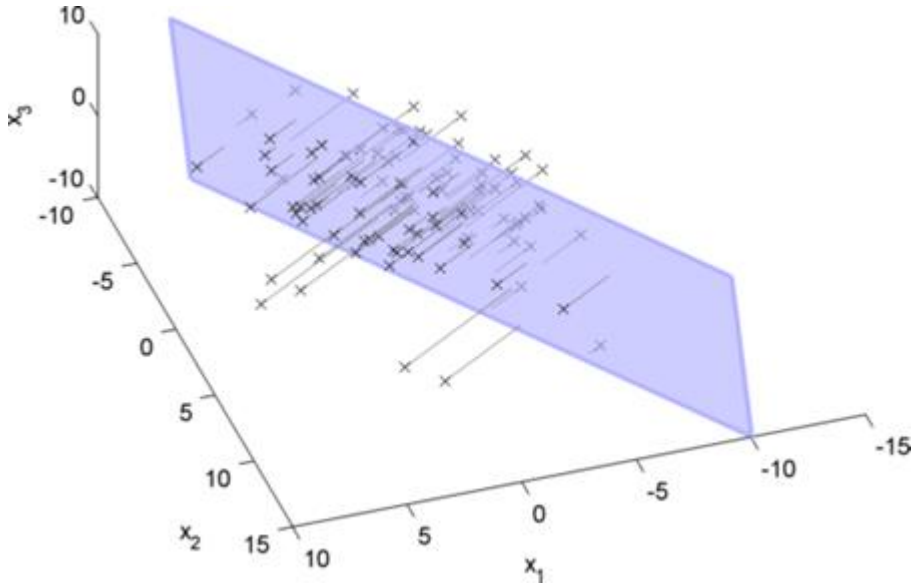
Part 2,

일반화된 프레쉐 평균

- 표본공간 M
  - M = R^p, S^p 등의 매니폴드…
- 모수공간 P "모수 = 분포를 기술하는 값"
  - 모수가 (차원축소된) "submanifold": 모수공간 P = {$M_q \subset$ M: 차원이 q}
- ρ : 작은 차원의 M_q 가 주어진 분포를 얼마나 잘 기술하는지에 대한 측도

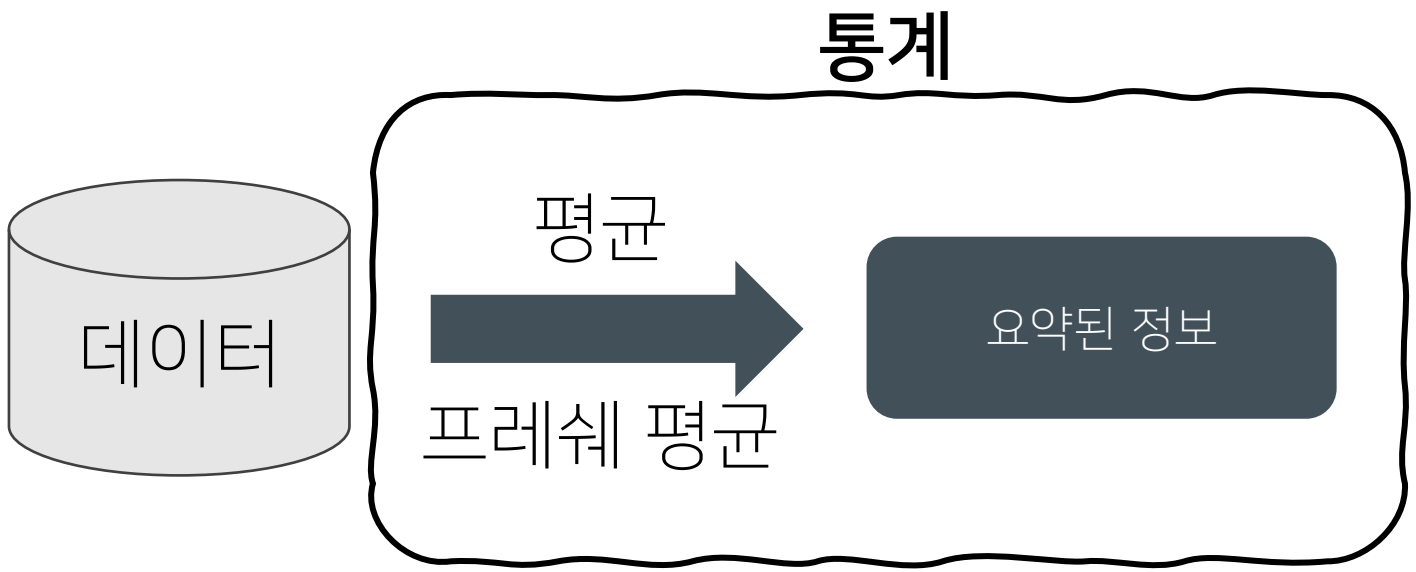
예: 주성분분석

매니폴드 데이터의 차원축소



Part 2,

평균이 얼마나 정확한가?



요약된 “정보”가 얼마나 정확한가?

Part 2,

평균의 불확실성과 중심극한정리

- $x_1, x_2, \ldots, x_n \sim_{iid} (\mu, \sigma^2)$
  평균 $\bar{x}_n$이 얼마나 달라질 수 있을까?

"중심극한정리"

- 점근분포 $\bar{x}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
  "관측값이 많을 때, 평균의 불확실성 패턴이 정규분포로 기술됨"
  "평균의 표준편차가 $1/\sqrt{n}$ 의 크기 (제곱근의 법칙)"

- 모수 μ $(= \arg\min_m E\, d(x,m)^2)$ 의 추정량 $\bar{x}_n$ 의 타당성

Part 2,

프레쉐 평균의 불확실성?

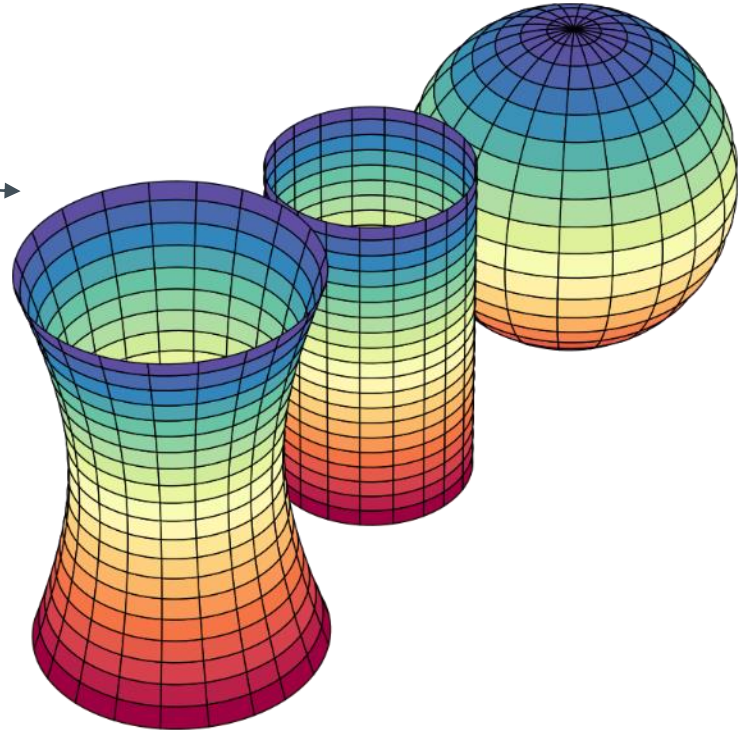
중심극한정리 on Riemannian manifold (리만 다양체)

- 표본공간 M, $x_1, x_2, \ldots, x_n \sim_{iid} F$
- 모수공간 P (리만 다양체)
- 일반화 프레쉐 표본평균 $m_n \in P$
- 일반화 프레쉐 모평균 $m \in P$

  평균 m_n이 얼마나 달라질 수 있을까?

  m_n 의 (점근)분포가 정규분포?

  벡터 공간이 아닌 P에서의 정규분포?



https://en.wikipedia.org/wiki/Curvature_of_Riemannian_manifolds#/media/File:Gaussian_curvature.svg

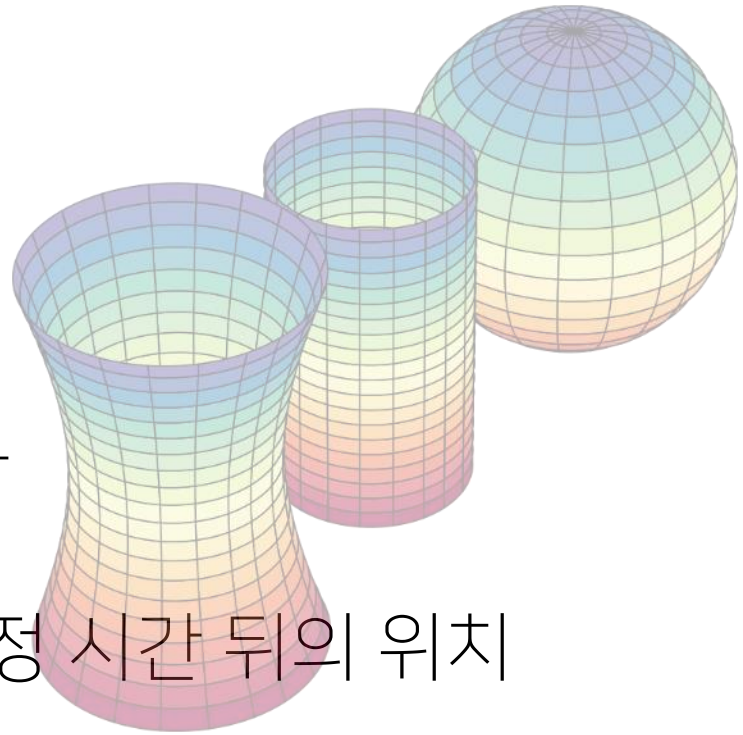
Part 2,

리만 다양체에서의 정규분포

- 모수공간 P (리만 다양체)에서의 정규분포?

- 정규분포?

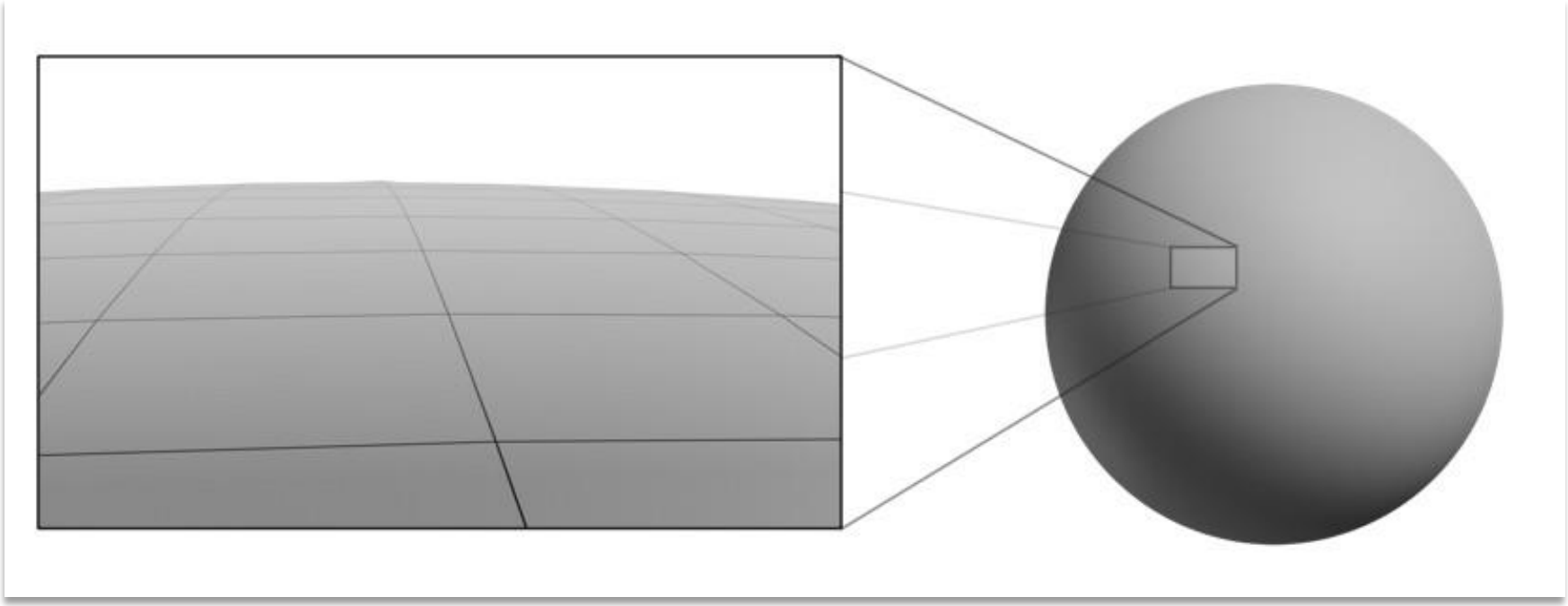
  1. 밀도함수 ∝ exp[- c (x-μ)^2]
  2. 변동이 같은 모든 분포 중 엔트로피가 가장 큼
  3. 열확산 방정식에 따라 확산하는 한 입자의 일정 시간 뒤의 위치



https://en.wikipedia.org/wiki/Curvature_of_Riemannian_manifolds#/media/File:Gaussian_curvature.svg

Part 2,

리만 다양체에서의 정규분포



3. 열확산 방정식에 따라 확산하는 한 입자의 일정 시간 뒤의 위치

- 국소적으로 (매우 작은 평면에서) 정의되는 정규분포

https://sinestesia.co/blog/tutorials/non-manifold-meshes-and-how-to-fix-them/

Part 2,

프레쉐 평균의 불확실성?

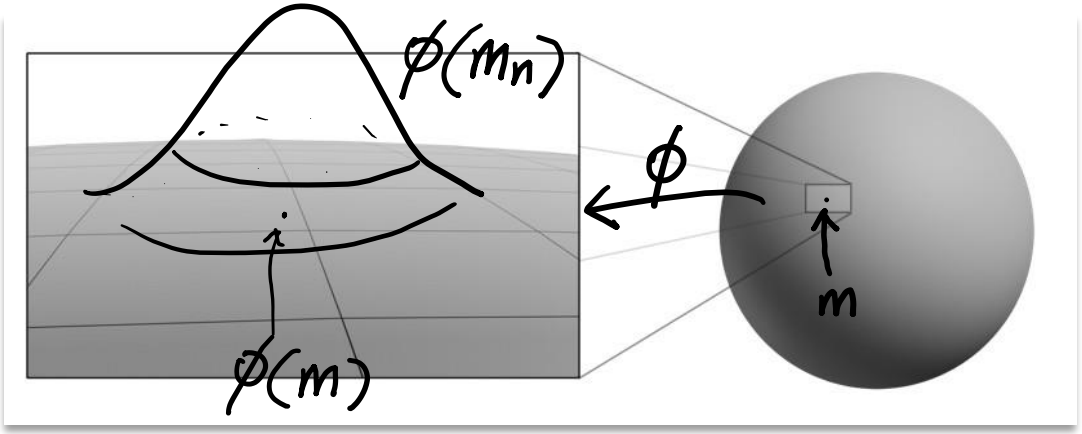
중심극한정리 on Riemannian manifold (리만 다양체)

- 리만 다양체: 국소적으로 평평함.
- 큰 수의 법칙*: 프레쉐 평균 $m_n \rightarrow m$
- m 주변의 한없이 작고 평평한 "평면"에서의 정규분포 필요
- 중심극한정리:

  m_n 의 분포가 m 에 한없이 가까워지므로

  m 주변의 매우 작은 평면에서의 분포가

  정규분포: $\Phi(m_n) - \Phi(m) \sim N(0, n^{-1}\Sigma)$



* BHATTACHARYA, R. AND PATRANGENARU, V. (2003) Large sample theory of intrinsic and extrinsic sample means on manifolds I. Ann. Statist. 31 1-29
BHATTACHARYA, R. AND PATRANGENARU, V. (2005) Large sample theory of intrinsic and extrinsic sample means on manifolds II. Ann. Statist. 33 1225-1259.

Part 2,

프레쉐 평균의 불확실성?

중심극한정리 on Riemannian manifold (리만 다양체)

- 프레쉐 평균 $m_n \rightarrow m$
- 중심극한정리: $\Phi(m_n) - \Phi(m) \sim N(0, n^{-1}\Sigma)$

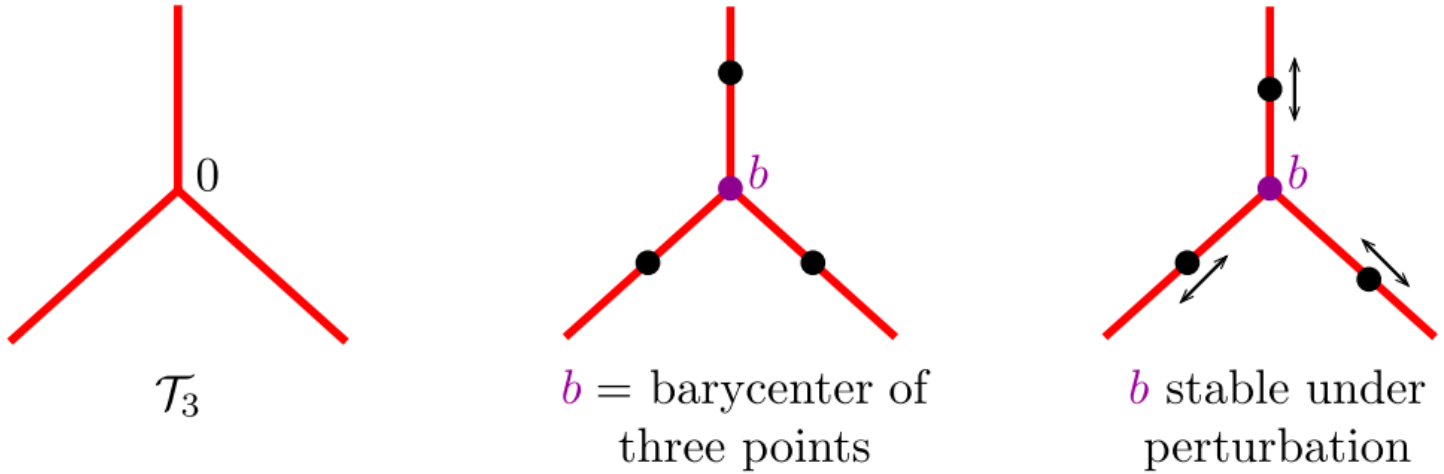
"관측값이 많을 때, **프레쉐 평균**의 불확실성 패턴이 정규분포로 기술됨"

"평균의 표준편차가 $1/\sqrt{n}$ 의 크기 (제곱근의 법칙)"

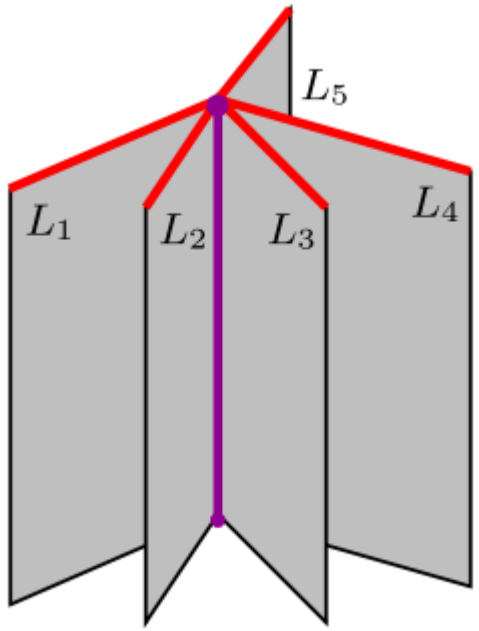
끈적한 평균(Sticky mean) 현상

- 층화된 모수공간 P에서 $m_n \to m$ (더 빠르게, $1/\sqrt{n}$ 보다 작은 변동)
- P = 거미/열린 책 (BHV tree space, Shape spaces, Space of covariance matrices)

거미 (1층: 점, 2층: 세 다각의 다리)



열린 책 (1층: 책등, 2층: 책장)

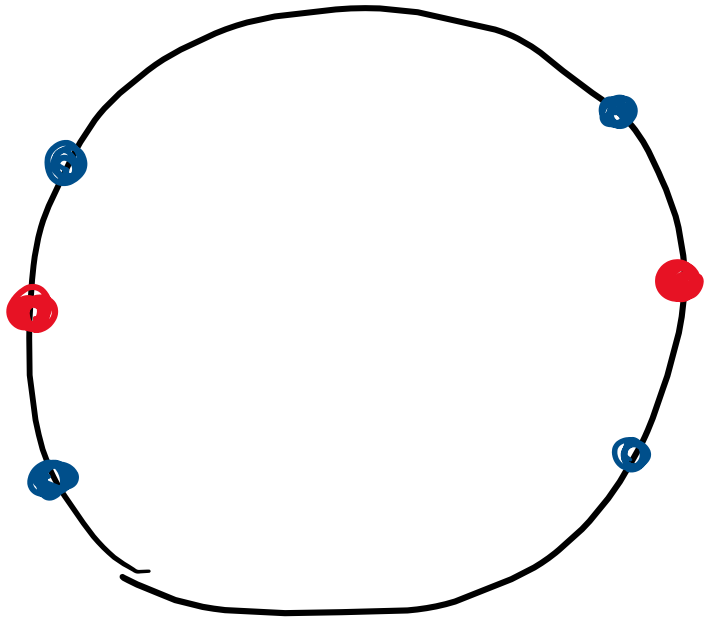


HOTZ, Thomas, et al. Sticky central limit theorems on open books. *The Annals of Applied Probability*, 2013, 23.6: 2238-2258.

Part 2,

얼룩진 평균(Smeary mean) 현상

- 프레쉐 평균의 **비유일성**:
  분포의 범위(support)가 넓고 매니폴드의 curvature > 0 일 때 종종 발생



{20°, 160°, -20°, -160°}의 프레쉐 평균?

m_n = 0°, 180° 두 개의 프레쉐 평균!

Part 2,

얼룩진 평균(Smeary mean) 현상

- 프레쉐 평균의 비유일성:
  분포의 범위(support)가 넓고 매니폴드의 curvature > 0 일 때 종종 발생

- 프레쉐 평균이 유일하지만 여전히 분포의 범위가 넓을 때 "얼룩진 평균":

$m_n \to m$ (더 느리게, 더 큰 변동)

표준편차가 $1/\sqrt{n}$ 보다 더 큼.

Eltzner, B., & Huckemann, S. F. (2019). A smeary central limit theorem for manifolds with application to high-dimensional spheres. *The Annals of Statistics*, *47*(6), 3360-3381.

Part 2,

평균 낼 수 없는 것들: 열린 문제들

- 프레쉐 평균의 불확실성: 표본 공간 M과 모수 공간 P의 특성에 따라 다름
- Active한 연구 분야!
- "매니폴드 가설"과 빅데이터의 잠재 공간*
- 잠재 공간이 매니폴드일 때, Manifold-specific distance/distribution?

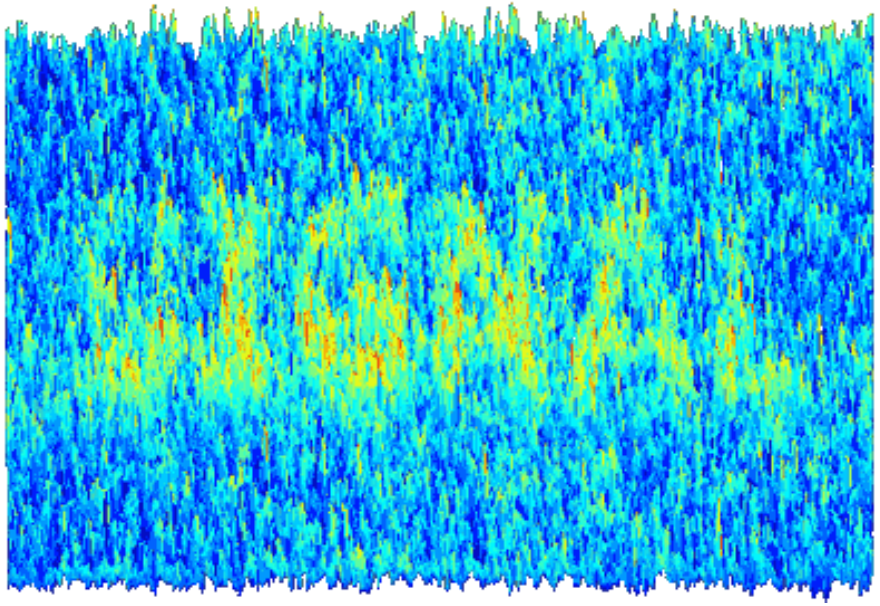
* Kühnel, L., Fletcher, T., Joshi, S., & Sommer, S. (2021, January). Latent Space Geometric Statistics. In *International Conference on Pattern Recognition* (pp. 163-178). Springer, Cham.

정리

Part 2,

현대 통계학의 도전과 성취

- 데이터 = 신호와 소음
- 통계: 계산과 요약으로 "신호" 파악
- 통계학: 이 "신호"가 진짜 신호인가?



- 대표적인 요약 "평균", 산술평균 낼 수 없는 것들의 "프레쉐 평균"
- 통계학의 새로운 도전: Big Data? Complex Data?

https://www.business2community.com/social-media-articles/5-ways-boost-signal-noise-ratio-social-networks-0850990

End of Slide

감사합니다