15단원

R Companion: Multiple Comparisons
http://rcompanion.org/rcompanion/f_01.html

15.1

- 기존의 가설 검증은 1개의 검정통계량으로 판단
- 하지만 microarray 기술 발전으로 개별→수천/수만 개의 유전자를 동시 발현

  마이크로어레이 자료의 분석
  http://www.cdc.go.kr/CDC/cms/content/mobile/56/19956_view.html

- 예시: 전립선암 표지자 연구
  - 인자: 6033개, 샘플: 102명

$$x_{ij} = \text{샘플 } j\text{의 } i\text{번째 유전인자} \quad (15.1)$$

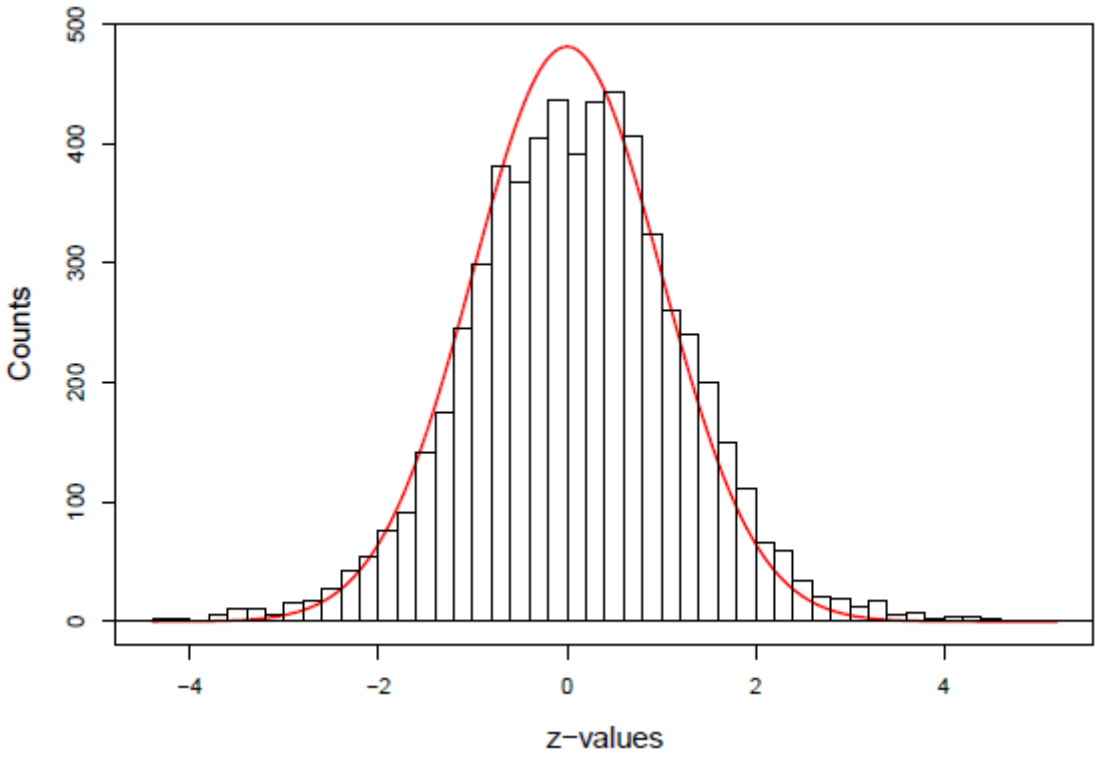
  - 스튜던트 t 분포로 환산 (자유도 = 100)

$$z_i = \Phi^{-1}(F_{100}(t_i)) \quad (15.2)$$

  - 귀무가설

$$H_{0i} : z_i \sim \mathcal{N}(\mu_i, 1), mu_i = 0$$
$$\text{and } \phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \quad (15.3)$$

- 아래 히스토그램과 정규분포 간의 차이를 통해 몇몇 인자의 기각 예상



- 어떤 가설을 기각할지 어떻게 확인하는가?
  - 유의 수준 5%의 가설 검증을 100번 진행할 경우 5번 정도 기각
  - 따라서 몇 가지 방법을 통해 보완 필요
- 단일 검정의 유의 수준

$$\alpha = \Pr\{\text{reject true } H_0\} \quad (15.6)$$

- Bonferroni bound
  - 유의 수준을 검증 횟수로 나눠 가설을 기각할 확률을 감소

$$\Phi^{-1}(0.05) = 1.645 \rightarrow \Phi^{-1}(0.05/N) = 4.31$$

  - 집단별 오류율(family-wise error rate): 제1종 오류

$$\text{FWER} = \Pr\{\text{reject any true } H_{0i}\} \quad (15.7)$$

  - Boole's Inequality: 가설을 한 개 이상 기각할 확률은 각 가설을 기각할 확률의 합보다 작다

$$\begin{aligned}\text{FWER} &= \Pr\left\{\bigcup_{I_0}\left(p_i \le \frac{\alpha}{N}\right)\right\} \le \sum_{I_0}\Pr\left\{p_i \le \frac{\alpha}{N}\right\} \\ &= N_0\frac{\alpha}{N} \le \alpha\end{aligned} \tag{15.8}$$

- Holm's procedure
  - p-값을 오름차순으로 정렬

$$p_{(1)} \le p_{(2)} \le \ldots \le p_{(i)} \le \ldots \le p_{(N)} \tag{15.9}$$

  - (15.10)을 만족하는 최소 i 찾기

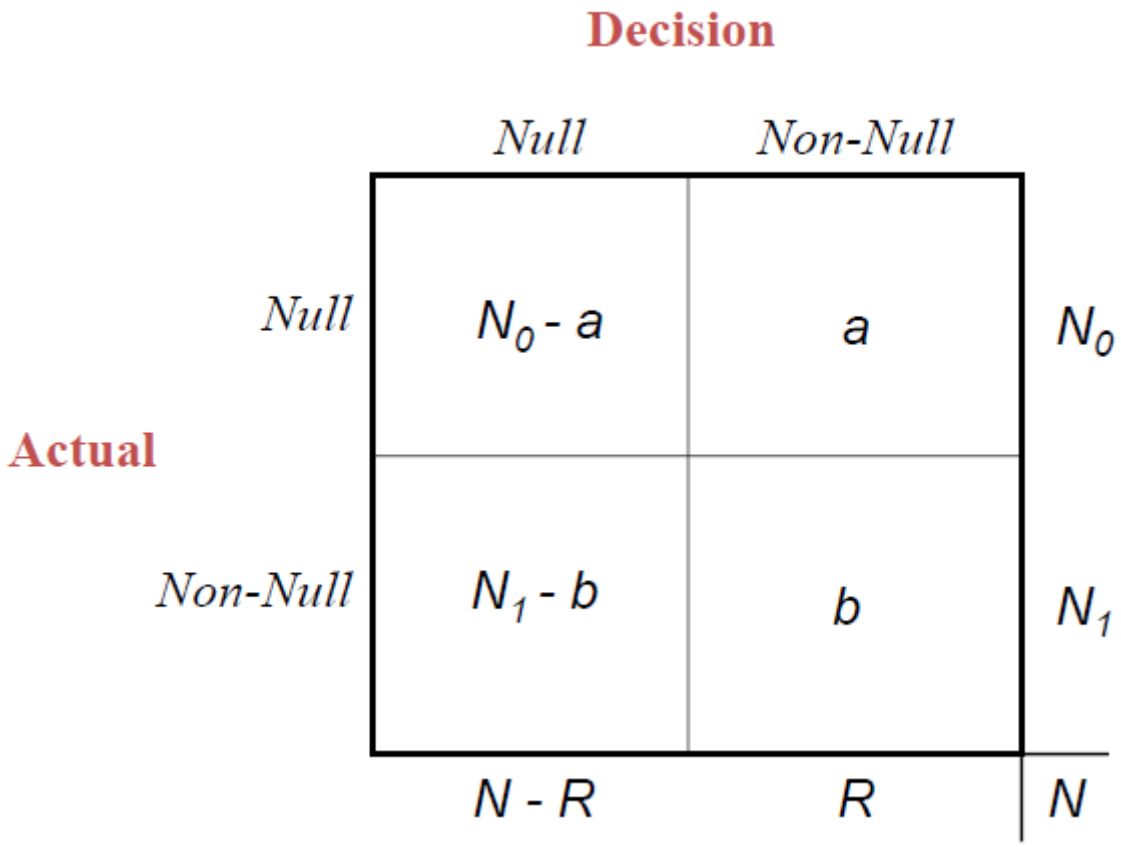
$$p_{(i_0)} > \frac{\alpha}{N-i+1} \tag{15.10}$$

$$\text{Reject all } H_{0(i)} \text{ for } i < i_0, \text{ accept for } i \ge i_0$$

15.2

http://www.kdiss.org/journal/download_pdf.php?spage=971&volume=25&number=5

- FWER은 small-scale testing을 위해 구상되어 large-scale에 적용하기에는 지나치게 제한적
- 따라서 large-scale에는 덜 보수적인 위발견율(False-Discovery Rate) 사용



- 위 그림에 따르면 false-discovery proportion은 다음과 같다
  - $R = 0$일 경우 Fdp = 0

$$\text{Fdp}(\mathcal{D}) = \frac{a}{R} \quad (15.11)$$

- 하지만, 일반적인 경우라면 Fdp는 관측 불가
  - 몇 가지 가정만 있다면 Fdp의 기대치는 통제 가능
  - FDR을 (15.12)와 같이 Fdp의 기대치로 정의할 경우, (15.13)을 만족하는 D는 FDR을 q에 맞게 통제

$$\text{FDR}(\mathcal{D}) = E\{\text{Fdp}(\mathcal{D})\} \quad (15.12)$$

$$\text{FDR}(\mathcal{D}) \leq q, \;\; 0 \leq q \leq 1 \quad (15.13)$$

- 위 Holm's procedure와는 반대로 D를 찾을 수 있다
  - i를 오름차순으로 정렬했을 때 (15.14)를 만족하는 최대값을 찾고
  - 이보다 작거나 같은 i에 대해서는 귀무 가설을 기각할 규칙 D를 찾는다

$$p_{(i_{\max})} \leq \frac{i_{\max}}{N} q \quad (15.14)$$

- Benjamini-Hochberg FDR Control (각 가설의 p-값은 독립적)

$$\mathrm{FDR}(\mathcal{D}_q) = \pi_0 q \leq q, \;\; \pi_0 = \frac{N_0}{N} \quad (15.15)$$

- 그런데 우리는 몇 개의 가설을 기각할지 미리 알 수 없기 때문에 단순히 q에 맞게 통제한다고 얘기함

$$\pi_0 \to 1$$

  - q는 보통 0.1로 설정함
- FDR을 선호하는 이유?
  - FWER(Holm's)

$$p_{(i)} \leq \frac{\alpha}{N - i + 1} \quad (15.16)$$

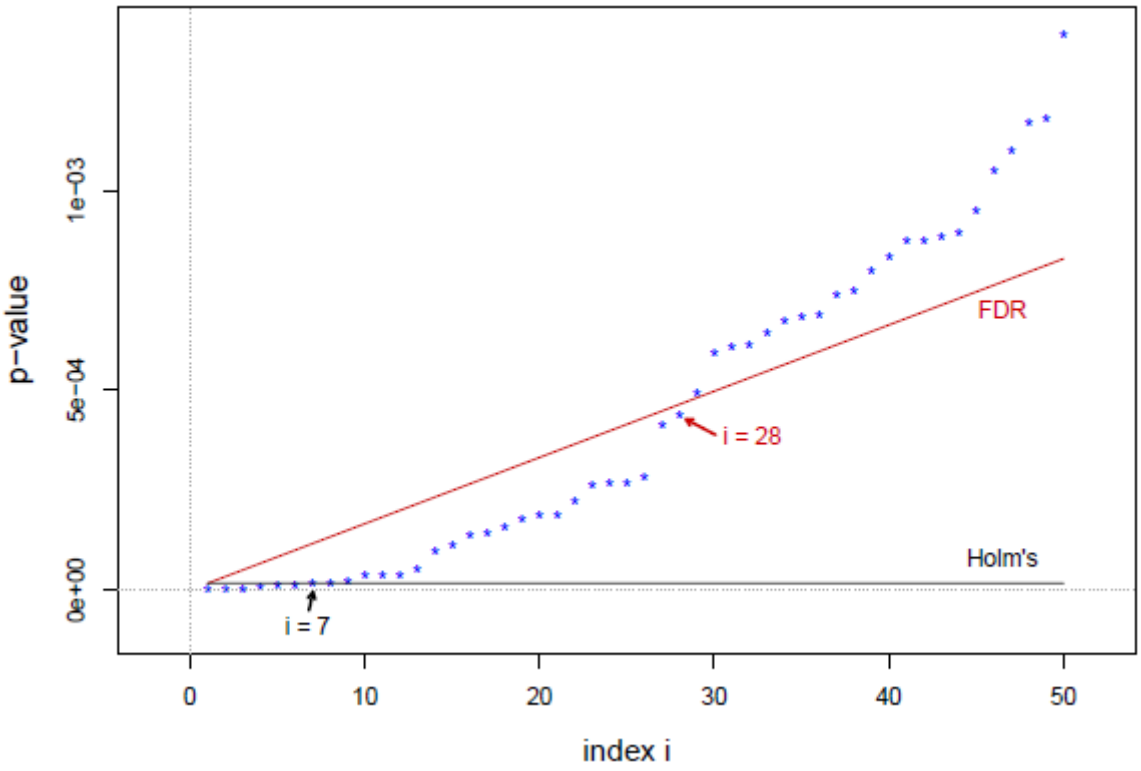
  - FDR

$$p_{(i)} \leq \frac{q}{N} i \quad (15.17)$$

  - 둘 간의 비율은 i와 선형

$$\frac{\mathrm{Threshold}(\mathcal{D}_q)}{\mathrm{Threshold}(\mathrm{Holm's})} = \frac{q}{\alpha} \left(1 - \frac{i-1}{N}\right) i \quad (15.18)$$

  - 따라서 귀무 가설을 기각하는 데에 있어 좀 더 느슨한 기준을 갖게 된다



- 그럼 FDR은 문제가 없는가?
  1. 비율(FDR)을 제어하는 게 제1종 오류의 확률 자체를 제어하는 것만큼 유의미한가?
  2. q의 값은 어떻게 설정하는가?
  3. 각 p-값은 과연 독립적일 수 있는가?
  4. 특정 i의 FDR의 유의미함은 이전 i의 영향을 받게 되는데 합리적인가?

15.3

- 위에 나열한 질문에 대한 답을 찾기 위해 베이지안으로 재정의

$$\begin{aligned} \pi_0 &= \Pr\{\text{null}\} \qquad f_0(z) \text{ density if null,} \\ \pi_1 &= \Pr\{\text{non-null}\} \quad f_1(z) \text{ density if non-null.} \end{aligned} \tag{15.19}$$

$$\pi_0 \approx 1 \text{ and } \phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \text{ from (15.3)}$$

$$\pi_1 \text{ unknown}$$

  - 생존 곡선의 경우,

$$S_0(z) = 1 - F_0(z) \text{ and } S_1(z) = 1 - F_1(z) \quad (15.20)$$
$$\Rightarrow S(z) = \pi_0 S_0(z) + \pi_1 S_1(z) \quad (15.21)$$
$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z) \quad (15.22)$$
$$\Rightarrow S(z_0) = \int_{z_0}^{\infty} f(z)dz \quad (15.23)$$

- Bayes false-discovery rate는 (15.24)와 같이 정의

$$g(\mu|x) = \frac{g(\mu)f_\mu(x)}{f(x)}, \mu \in \Omega \quad (3.5)$$
$$\begin{aligned} \text{Fdr}(z_0) &\equiv \Pr\{\text{case } i \text{ is null} | z_i \geq z_0\} \\ &= \frac{\pi_0 S_0(z_0)}{S(z_0)} \end{aligned} \quad (15.24)$$
$$\Rightarrow \pi_0 = g(\mu),\ S_0(z_0) = f_\mu(x),\ S(z_0) = f(x)$$

  - 통상적으로 분자는 아래의 값으로 알고 있다고 가정하지만,

$$S_0(z_0) = 1 - \Phi(z_0) \text{ and } \pi_0 \to 1$$

  - 분모의 경우 추정값만 유추 가능

$$\hat{S}(z_0) = \frac{N(z_0)}{N}, \text{ where } N(z_0) = \#\{z_i \geq z_0\} \quad (15.25)$$

  - (15.24)에 (15.25)를 대입하여 Fdr의 추정값 유도

$$\widehat{\text{Fdr}}(z_0) = \frac{\pi_0 S_0(z_0)}{\hat{S}(z_0)} \quad (15.26)$$

  - (15.5)와 (15.20)으로 부터 아래의 관계를 유추할 수 있고

$$p = 1 - F_0(z) \quad (15.5)$$
$$S_0(z) = 1 - F_0(z) \quad (15.20)$$
$$\Rightarrow p_i = S_0(z_i)$$

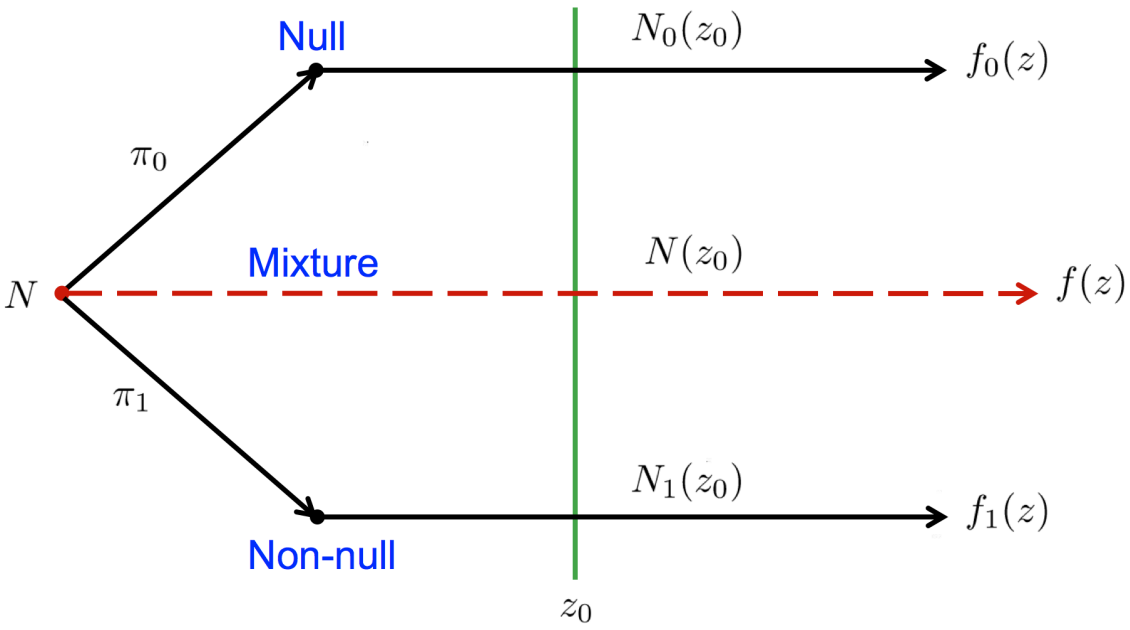
  - (15.25)와 함께 (15.14)를 재정의하면

$$
\begin{aligned}
p_{(i)} &\le \frac{i}{N} q \\
&\Rightarrow S_0(z_{(i)}) \le \hat{S}(z_{(i)}) q \quad (15.27) \\
&\Rightarrow \widehat{\mathrm{Fdr}}(z_{(i)}) \le \pi_0 q \quad (15.28)
\end{aligned}
$$

- 다시 말하면, 제1종 오류의 사후 확률이 낮은 케이스를 기각

1. 비율(FDR)을 제어하는 게 제1종 오류의 확률 자체를 제어하는 것만큼 유의미한가?
   FDR은 제1종 오류의 사후 확률과 관련이 있다
2. q의 값은 어떻게 설정하는가?
   베이즈 리스크의 허용 오차 최대값
3. 각 p-값은 과연 독립적일 수 있는가?
   상관성은 있으나 S 및 Fdr의 추정값은 무편향. 다만 상관성에 따른 비용(분산의 증가) 발생.
4. 특정 i의 FDR의 유의미함은 이전 i의 영향을 받게 되는데 합리적인가?
   15.4에서 좀 더 깊게 다룰 예정

- 아래와 같은 상황이 주어진다고 가정



$$
\begin{aligned}
N(z_0) = N_0(z_0) + N_1(z_0) = R \quad (15.30) \\
\mathrm{Fdp} = \frac{N_0(z_0)}{N(z_0)} \quad (15.31)
\end{aligned}
$$

- 우리는 분모만 관측 가능하기 때문에 분자에 추정값 (15.32)를 대입하여 (15.33) 유도

$$E\{N_0(z_0)\} = N\pi_0 S_0(z_0) \quad (15.32)$$

$$\widehat{\text{Fdp}} = \frac{N\pi_0 S_0(z_0)}{N(z_0)} = \frac{\pi_0 S_0(z_0)}{\hat{S}(z_0)} = \widehat{\text{Fdr}}(z_0) \quad (15.33)$$

- 결국 우리는 Fdr과 Fdp을 경험적 베이즈로 추정 가능

15.4

- 베이지안 관점에서는 부등식(≥)이 아니라 등식(=)에 더 초점을 맞추게 된다
  - 특정값이 관측됐을 때 귀무 가설을 기각할 확률
- 따라서 지역 오발견율을 (15.34)와 같이 새로 정의 ↔ (15.24)

$$\text{fdr}(z_0) = \Pr\{\text{case } i \text{ is null} | z_i = z_0\} \quad (15.34)$$

- 기각 구간을 (15.35)와 같이 정의하고

$$\mathcal{Z}_0 = \left[z_0 - \frac{d}{2}, z_0 + \frac{d}{2}\right], d = 0.1 \quad (15.35)$$

- 지역 오발견 비율은 (15.36)과 같이 정의하여

$$\text{fdp}(z_0) = \frac{N_0(\mathcal{Z}_0)}{N(\mathcal{Z}_0)} \quad (15.36)$$

- 지역 오발견율의 추정치를 (15.31)-(15.33)과 같이 유도

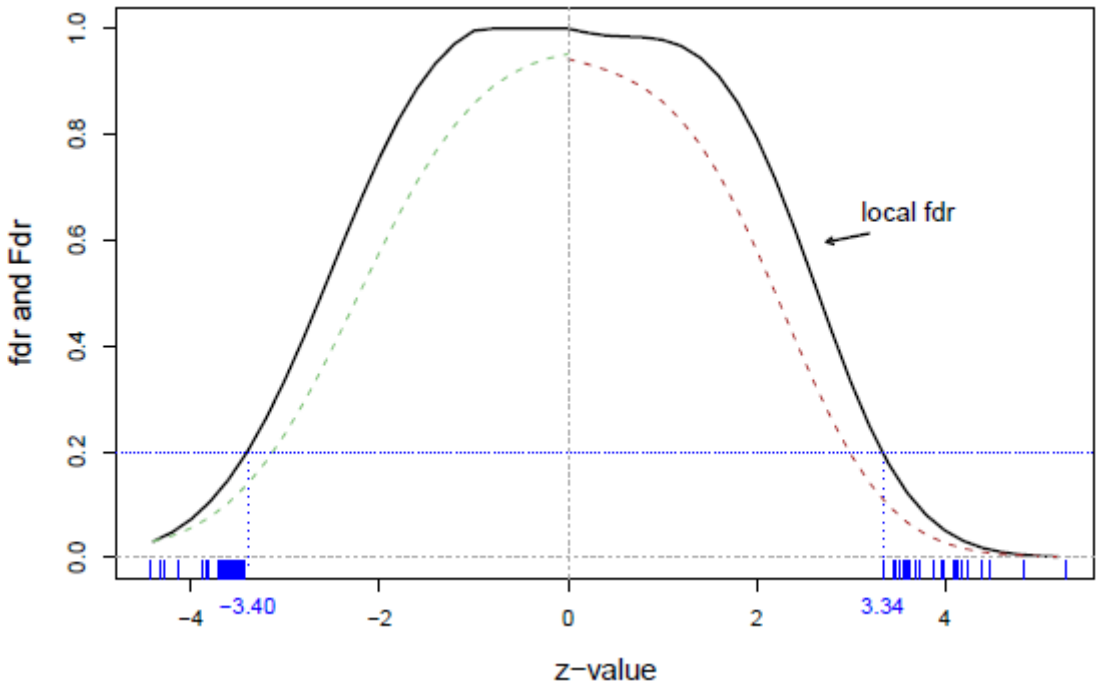
$$\widehat{\text{fdr}}(z_0) = \frac{N\pi_0 f_0(z_0)d}{N(\mathcal{Z}_0)} \quad (15.37)$$

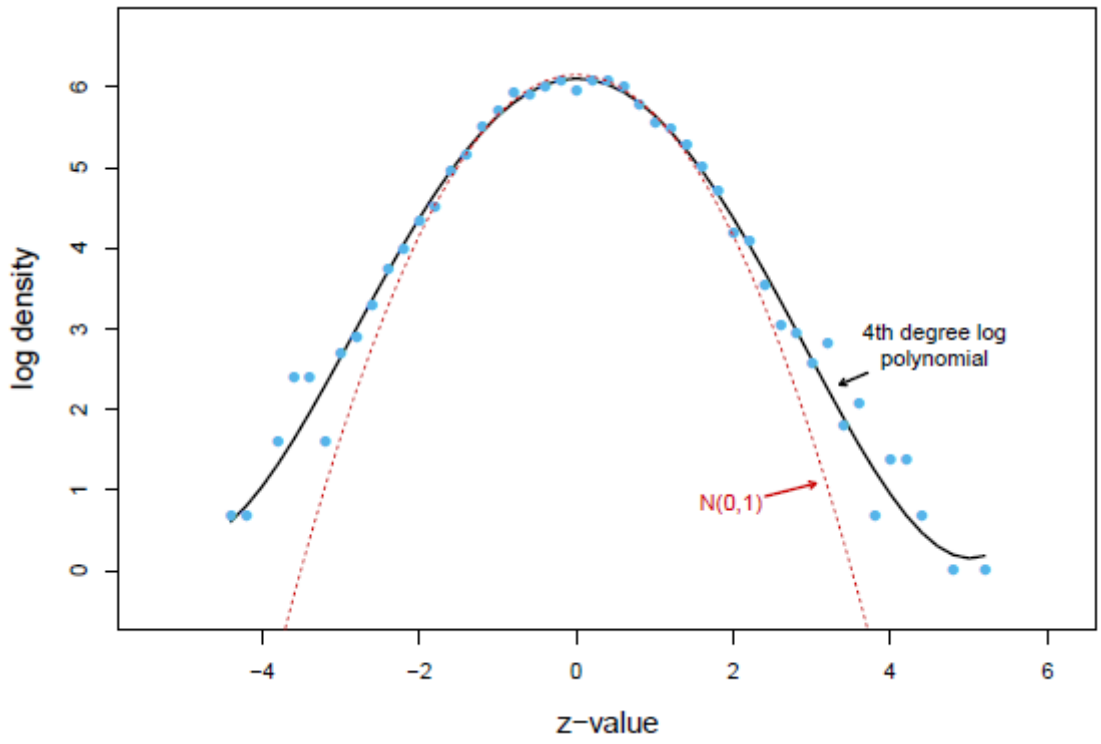
- 기본 베이즈 정리를 통해 (15.38),(15.39)로 재정의 가능

$$\text{fdr}(z) = \frac{\pi_0 f_0(z)}{f(z)} \quad (15.38)$$

$$\widehat{\text{fdr}}(z_0) = \frac{\pi_0 f_0(z_0)}{\hat{f}(z_0)} \quad (15.39)$$

- 전립선암 데이터의 분포를 나타낸 그림 2개를 통해 대부분(93%)의 유전자는 영향이 없음을 확인





- 왜 0.2?

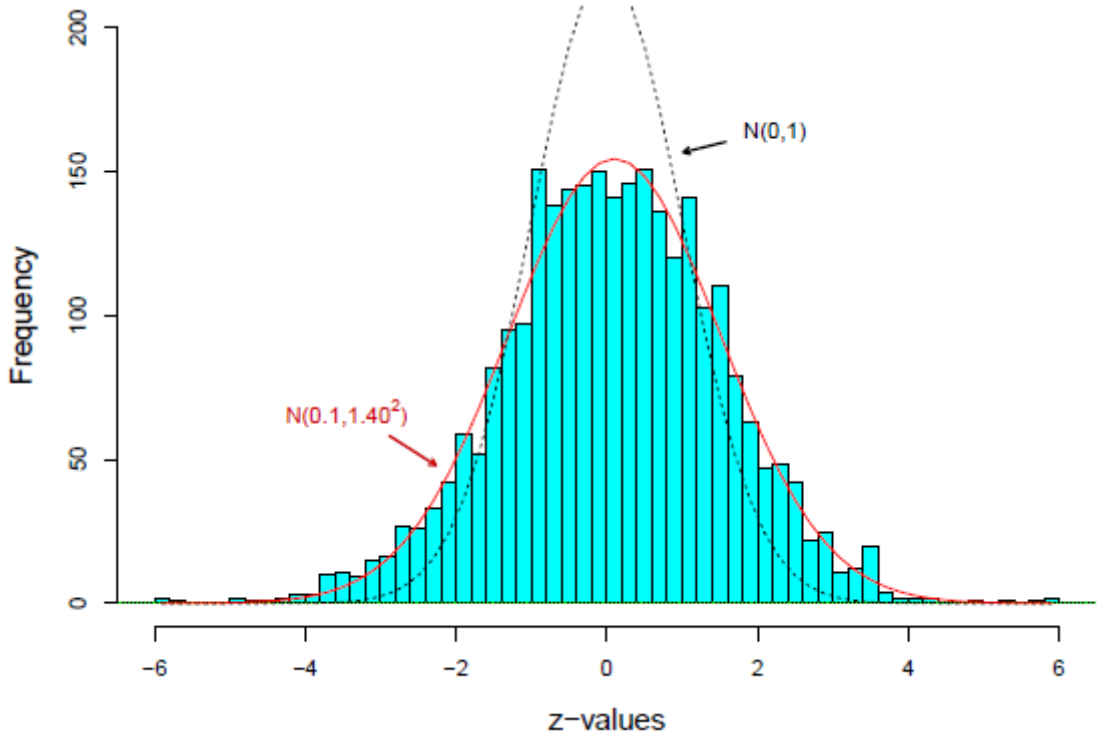
$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z) \quad (15.22)$$

$$\text{fdr}(z) = \frac{\pi_0 f_0(z)}{f(z)} \quad (15.38)$$

$$\text{Fdr}(z_0) = E\{\text{fdr}(z) | z \geq z_0\} \quad (15.44)$$

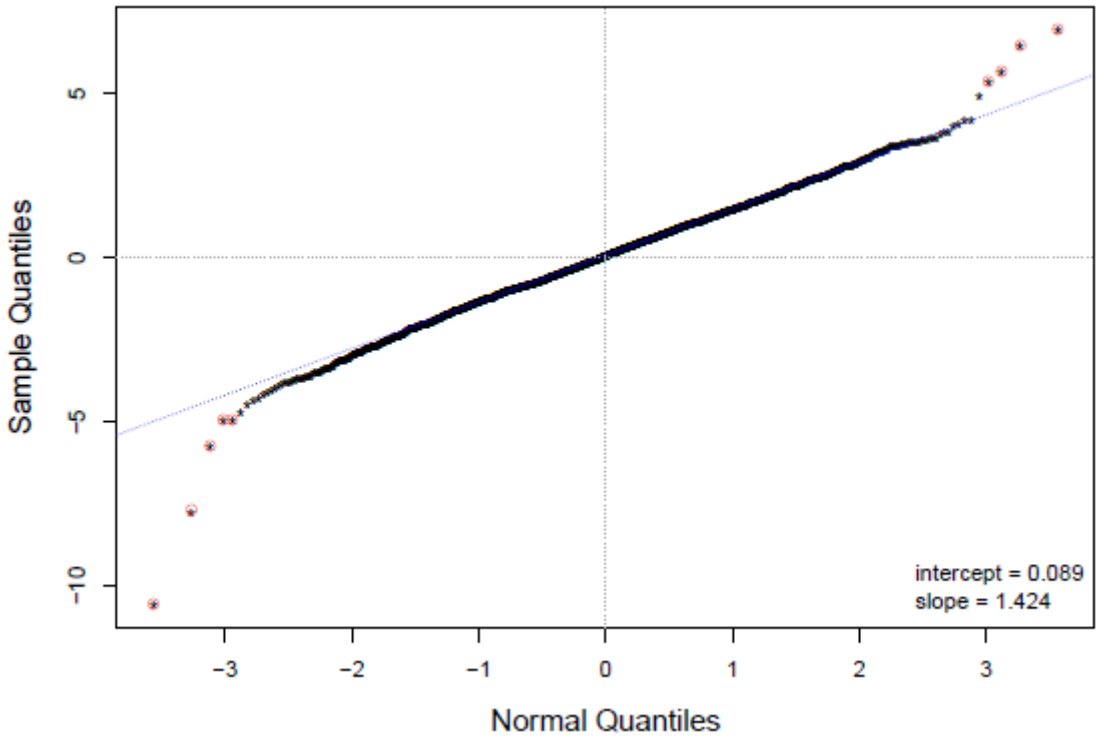
15.5

- 다중 검증의 경우 좀 더 현실적인 귀무 분포를 찾을 수도 있다



- 데이터: 경찰의 단속 시 인종차별 영향 유무
- 일반적인 정규 분포를 적용하기에는 중앙 부분이 너무 넓은 것을 알 수 있다
- MLE를 사용하여 분포 추정. QQ Plot을 통해서도 적합한지 검증

!!! : 네이버 블로그
https://blog.naver.com/kjihoon0914/221214498142

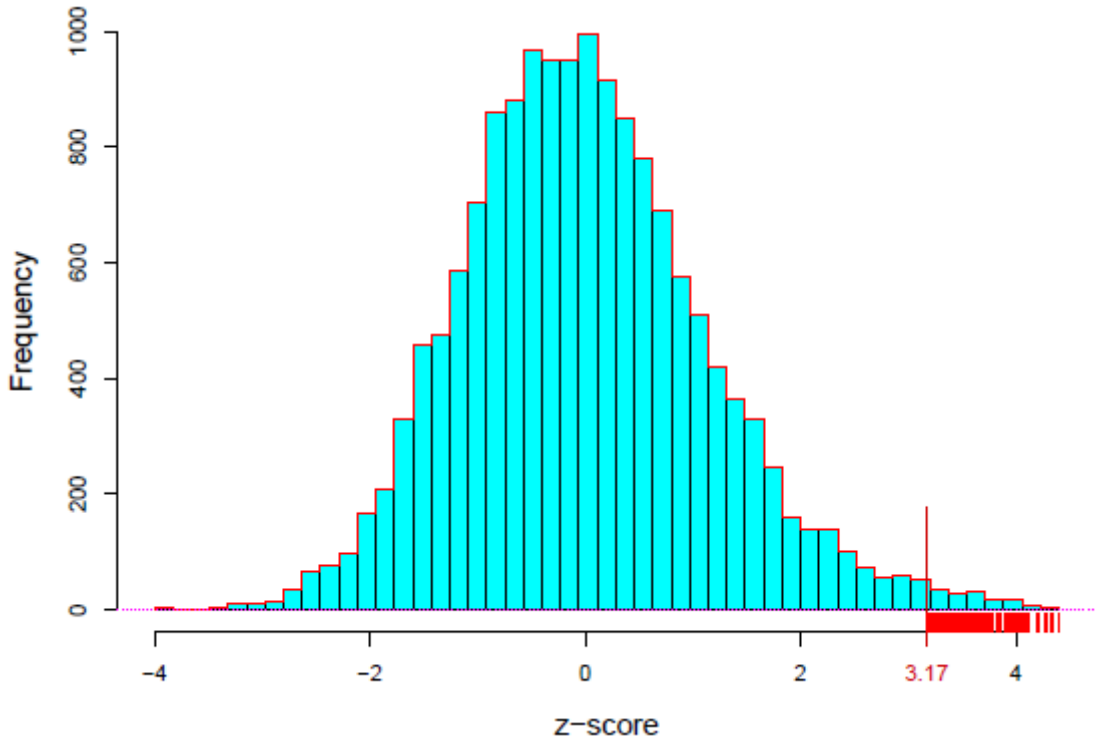


- MLE로 추정한 분포를 적용했을 경우 최우측 4명의 경찰관만 인종 차별적이라고 판정 반면에, 일반적인 정규 분포 적용 시 125명이라고 판정
- 일반적인 정규분포를 적용할 때의 문제점
  - 테일러 급수를 사용하여 예측하는 경우 꼬리 부분에서 발생할 수 있는 오차로 인한 결과의 왜곡
  - 일정 수준 이상의 상관관계가 존재할 경우 모집단분포를 왜곡
  - 경찰관 데이터에는 시간대 및 동네 등의 인자를 최대한 통제하려 했지만 미관측 공변량이 분명 존재할 수 있다
  - 위에서 설명한 것처럼 적합한 효과 크기를 모를 경우
- R에서는 locfdr 패키지를 사용하여 계산 가능
- 상황에 따라 데이터로부터 귀무 분포를 유추하는 것이 필요할 수 있다.
  - 다만, 추정값을 구하는 그 자체만으로도 변산도(variability)가 증가하기 때문에 일반적인 분포가 잘 맞는 경우는 그대로 적용

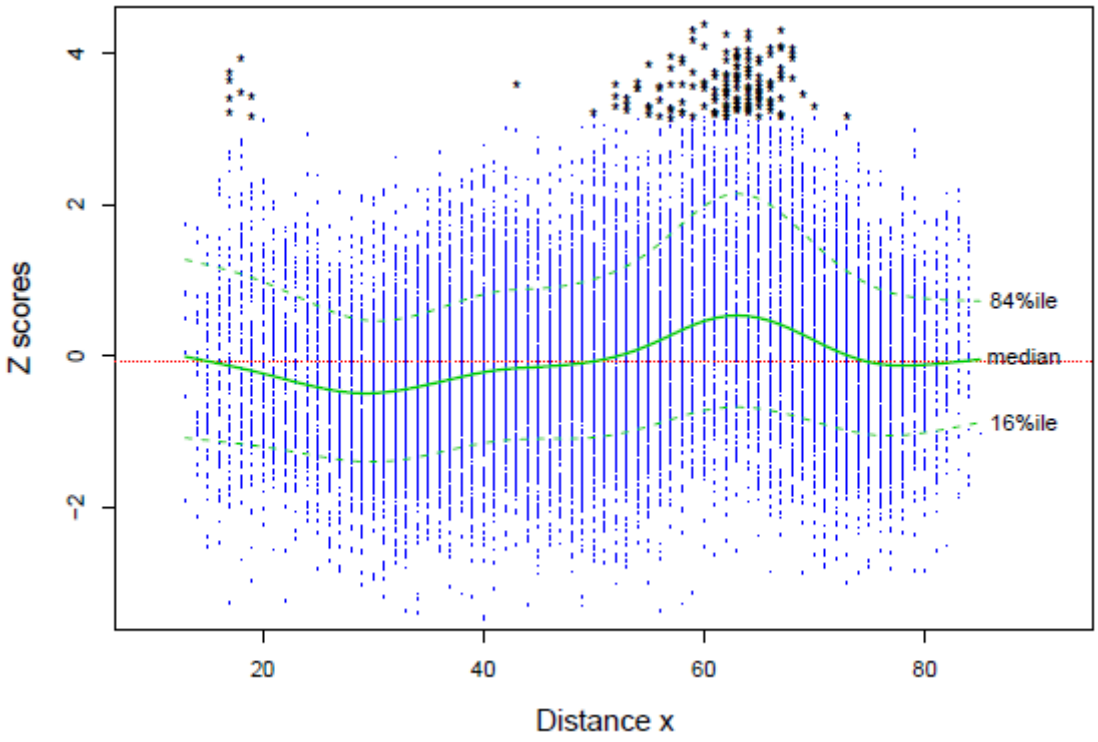
15.6

- FDR을 다루면서 6.4와 7.4에서 다뤘던 간접 증거에 대해 좀 더 살펴볼 필요가 있음

- 전립선암 데이터
  - 살펴보고 있는 유전자가 다른 유전자에 의해 영향을 받았을 수 있지만 정확히 어떤 유전자로부터 영향을 받았는지는 모름
- 확산 텐서 이미지 (뇌 연구)



  - locfdr을 통해 분포를 유추해도 일반적인 정규 분포와 큰 차이가 없는 것으로 판단



- 하지만, 뇌 뒤쪽에서부터의 거리를 인자로 놓고 분포를 추정할 경우 위 그림처럼 50~70 사이에 큰 편차가 있는 것을 확인
- 그렇다면 50~70은 별도의 그룹으로 생각해야 할까?

- 텐서 이미지 연구처럼 데이터 간의 연관성 및 적합성 또한 고려를 해야 하는데 이것은 과연 누구의 역할일까?
- 연관성을 수식으로 나타낼 수 있을까?