Statistical Inference II - lecture 4

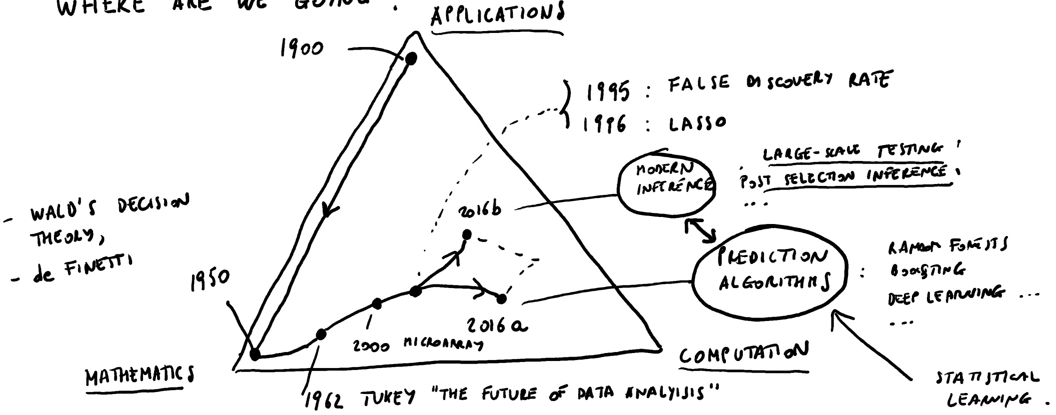
12 April 2021 09:13

REFERENCES:

EFRON & HASTIE (2016) EPILOGUE p. 448
CHAPTER 15, SECTIONS 15.1, 15.2 p 271 - 278

GOEMAN & SOLARI (2014) STATISTICS IN MEDICINE (TUTORIAL)

WHERE ARE WE GOING?



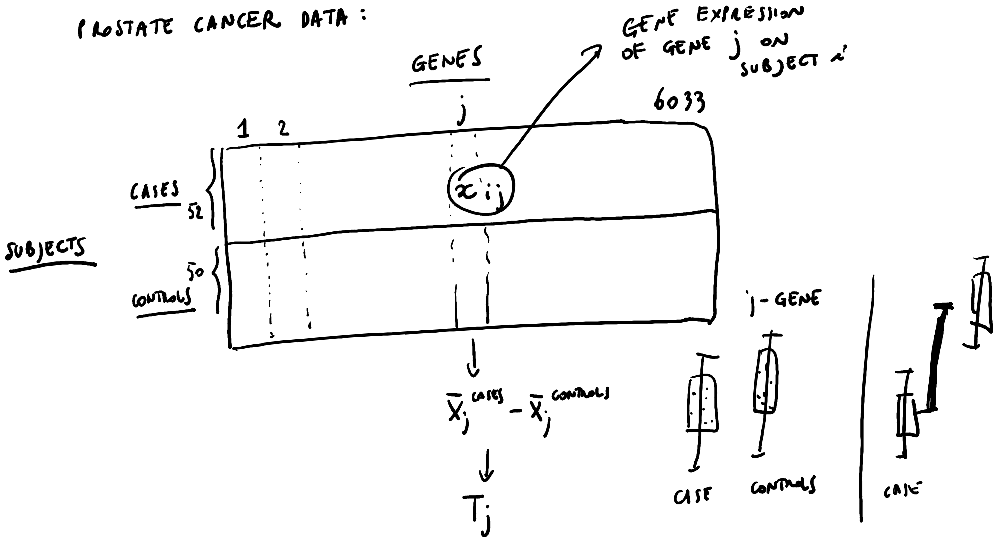
WHEATHER OR NOT WE CAN PREDICT THE FUTURE OF STATISTICS, WE CAN AT LEAST EXAMINE THE PAST TO SEE HOW WE'VE GOTTEN WHERE WE ARE.

CONFORMAL INFERENCE.

BREIMAN (2001)

MICROARRAY: BIOMEDICAL DEVICE THAT ENABLED THE MEASUREMENT OF "ACTIVITY" OF THOUSANDS OF GENES.

PROSTATE CANCER DATA:



$p_1 \; p_2 \; \ldots\ldots \; p_j \; \ldots\ldots \; p_m$

$H_j : \mu_j = 0$

LARGE-SCALE TESTING: FIND ONLY A FEW "INTERESTING" GENES AMONG MANY

PROBLEM: MANY TESTS.

<u>SINGLE TEST</u>: IS IMPORTANT TO <u>CONTROL</u> THE TYPE I ERROR

$$\mathbb{P}(\text{TYPE I ERROR}) \leq \alpha$$

<u>MANY TESTS</u>

CONTROL OF $\begin{cases} \text{FAMILYWISE ERROR RATE (FWER)} \\ \text{FALSE DISCOVERY RATE (FDR)} \\ \text{FALSE DISCOVERY PROPORTION (FDP)} \end{cases}$

<u>SETTING</u>

m NULL HYPOTHESIS: $H_1, H_2, \dots, H_m$

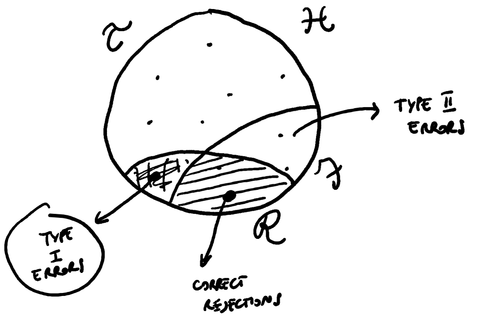
m_1 FALSE

$m_0 = m - m_1$ TRUE

P-VALUE $p_1 \quad p_2 \quad p_m$

<u>GOAL</u>

$\mathcal{R} \simeq \mathcal{F}$



REJECTIONS

$$\mathcal{R} = \{ H_i : p_i \leq c \} \quad \longleftarrow \text{"DISCOVERIES"}$$

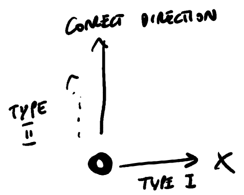
c ↑ CRITICAL VALUE

TYPE I ERRORS $\quad \mathcal{T} \cap \mathcal{R} \quad \leftarrow$ FALSE DISCOVERIES

II ERRORS $\quad \mathcal{F} \setminus \mathcal{R}$

$\mathcal{F} \cap \mathcal{R} \quad \longleftarrow$ TRUE DISCOVERIES

USUALLY TYPE I ERRORS ARE MORE PROBLEMATIC — $\begin{cases} \underline{\text{TYPE I}} \rightarrow \text{MISLEADING SCIENTIFIC RESULT} \\ \underline{\text{TYPE II}} \rightarrow \text{MISSING OUT A CORRECT SCIENTIFIC RESULT.} \end{cases}$



		HYPOTHESES		
		TRUE	FALSE	
DECISION	REJECT	$\underline{V}$ = NUMBER OF TYPE I ERRORS	U	$R = \# \mathcal{R}$
	NOT REJECT	$m_0 - V$	$m_1 - U$	$m - R$
		m_0	m_1	m
		DON'T OBSERVE		↑ OBSERVE

FALSE DISCOVERY PROPORTION <u>(FDP)</u>

DON'T OBSERVE OBSERVE

FALSE DISCOVERY PROPORTION (FDP)

$$Q = \frac{V}{\max(R,1)} = \begin{cases} V/R & \text{IF } R>0 \\ 0 & \text{OTHERWISE} \end{cases}$$

PROPORTION OF FALSE REJECTIONS AMONG REJECTIONS

FAMILYWISE ERROR RATE

$$FWER = \mathbb{P}(Q>0) = \mathbb{P}(V>0)$$

PROBABILITY OF HAVING AT LEAST ONE FALSE REJECTION

FALSE DISCOVERY RATE

$$FDR = \mathbb{E}(Q)$$

EXPECTED PROPORTION OF FALSE REJECTIONS

$$\text{CONTROL AT LEVEL } \alpha \quad \begin{cases} FWER \leq \alpha \\ FDR \leq \alpha \end{cases}$$

FWER AND FDR

Q	$\mathbb{1}$
0	0
x	1

- $0 \leq Q \leq 1$, THEN $Q \leq \mathbb{1}\{Q>0\}$

$$\mathbb{E}(Q) \leq \mathbb{E}(\mathbb{1}\{Q>0\})$$
$$\mathbb{P}(Q>0)$$
$$\underline{FDR} \leq \underline{FWER} \quad \underline{\leq \alpha}$$

- IF $\boxed{m_0 = m}$, THEN $V = R$ AND Q IS BERNOULLI

$$\mathbb{E}(Q) = \mathbb{P}(Q>0)$$
$$FDR = FWER.$$

- IF $\boxed{m=1}$ THEN $\mathbb{P}(\text{TYPE I ERROR}) = FDR = FWER$.

EXAMPLE $m=100$, $m_0 = 80$, REJECT IF $p \leq 5\%$

$\mu = 0$ TRUE HYP

$\mu = 2$ FALSE HYP

	I	II	III				
R	20	17	23	5	...		
V	4	5	6	0	...	$\longrightarrow$ $\underline{0.983}$	FWER
V/R	0.2	0.29	0.31	0	...	$\longrightarrow$ $\underline{0.232}$	FDR

NULL P-VALUES

$q_1, \dots, q_{m_0}$ P-VALUES FROM TRUE NULL HYPOTHESIS.

WE SAY THAT WE HAVE VALID NULL P-VALUES IF

$$\mathbb{P}(q_i \leq u) \leq u$$

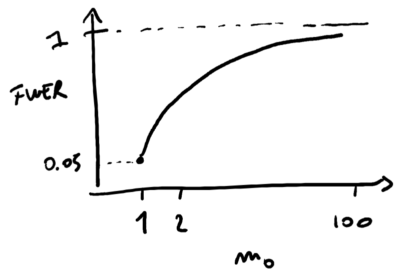
$$q_i \sim U(0,1) \qquad \mathbb{P}(U(0,1) \leq u) = u$$

$$\text{FWER} = \mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V = 0)$$

$$= 1 - (1 - c)^{m_0}$$

$q_1, \dots, q_{m_0}$ IID $U(0,1)$

$c = 0.05$



$$\underbrace{\mathbb{P}(V \geq 1)}_{} \underset{\text{MARKOV INEQ.}}{\leq} \frac{\mathbb{E}(V)}{1} = \underbrace{E(V)}_{} \leq \alpha$$

FDR $\leq$ FWER $\leq$ PFER   PER FAMILYWISE ERROR RATE

$$\mathbb{E}(V) = \mathbb{E}\left(\sum_{i=1}^{m_0} \mathbb{1}\{q_i \leq c\}\right)$$

$$= \sum_{i=1}^{m_0} \mathbb{E}\left(\mathbb{1}\{q_i \leq c\}\right)$$

$$\underset{q_i \sim U(0,1)}{=} \sum_{i=1}^{m_0} \mathbb{P}(q_i \leq c)$$

$$= m_0 c$$

$m = 1000$

$m_0 = 100$

$$R = \{H_i : p_i \leq \underline{0.05}\}$$

$$\uparrow \frac{0.05}{1000}$$

$\underline{E(V) \leq 5\%}$

1000

BONFERRONI METHOD

REJECT THE HYPOTHESIS WITH P-VALUE LESS THAN $\frac{\alpha}{m}$

$$\mathcal{R} = \left\{ H_i : p_i \le \frac{\alpha}{m} \right\}$$

IT CONTROLS FWER $\le \alpha$

PROOF

$$FWER = \mathbb{P}\left(\bigcup_{i=1}^{m_0} \left\{ q_i \le \frac{\alpha}{m} \right\} \right) \underset{\text{BOOLE}}{\le} \sum_{i=1}^{m_0} \mathbb{P}\left(q_i \le \frac{\alpha}{m} \right) \underset{\text{VALIDITY}}{\le} \underbrace{\frac{m_0 \alpha}{m}}_{\pi_0} \le \alpha$$

ADJUSTED P-VALUES $\tilde{p}_j = \min(m \cdot p_j, 1)$

$$\mathcal{R} = \left\{ H_i : \tilde{p}_j \le \alpha \right\}$$

HOLM METHOD : UNIFORMLY MORE POWERFUL THAN BONFERRONI

SEQUENTIAL BONFERRONI :

P-VALUES	$\underline{0.001}$	$\underline{0.01}$	0.02	0.6	$\frac{0.05}{4} = 0.0125$
	R	R	NR	NR	
	/	/	$\underline{0.02}$	0.6	$\frac{0.05}{2} = 0.025$
			R	NR	
				0.6	$\frac{0.05}{1}$
				NR	

BENJAMINI-HOCHBERG (1995) JRSS-B 74K CITATIONS

EM ALGORITHM 64K

KAPLAN-MEIER 59K

COX' PH MODEL 54K

LASSO 39K

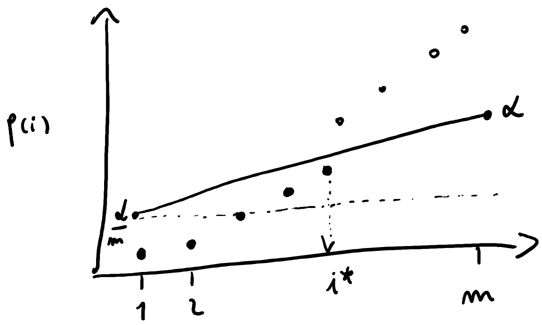
BENJAMINI-HOCHBERG METHOD

SORT P-VALUES $p_{(1)} \le p_{(2)} \le \dots \le p_{(m)}$

IF $p_{(i)} > \frac{i\alpha}{m} \; \forall i \quad \mathcal{R} = \emptyset$ STOP.

FIND $i^* = \max\left\{ i : p_{(i)} \le \frac{i\alpha}{m} \right\}$

$$R = \left\{ H_i : p_i \le \frac{i^* \alpha}{m} \right\}$$



THEOREM

ASSUME $p_1, \dots, p_m$ INDEPENDENT, AND NULL p-VALUES ARE $U(0,1)$.

THEN BH METHOD

$$FDR = \pi_0 \alpha \le \alpha$$

PROOF:

$m_0 = 0$ NOTHING TO PROVE

$\boxed{m_0 > 1}$ $\quad V_i = \mathbb{1}\{ H_i \text{ REJECTED} \}$ for $i \in T = \{ i : H_i \in \mathcal{H}_0 \}$

$$Q = \sum_{i \in T} \frac{V_i}{R \vee 1}$$

CLAIM: $\boxed{\mathbb{E}\left(\frac{V_i}{R \vee 1} \right) = \frac{\alpha}{m} \quad \forall i \in T}$

$$FDR = \sum_{i \in T} \mathbb{E}\left(\frac{V_i}{R \vee 1} \right) = m_0 \frac{\alpha}{m} = \pi_0 \alpha$$

IF $R = k$, THEN H_i IS REJECTED IIF $p_i \le \frac{k\alpha}{m}$

$$V_i = \mathbb{1}\left\{ p_i \le \frac{k\alpha}{m} \right\} \text{ for } i \in T$$

ALSO $\boxed{V_i \, \mathbb{1}\{R = k\} = V_i \, \mathbb{1}\{ R(p_i \downarrow 0) = k \} \quad \forall i}$

IF p_i IS REJECTED, SET $p_i = 0$ AND $R(p_i \downarrow 0) = R$

IS NOT REJECTED, THEN $V_i = 0$

LET $\mathcal{F}_i = \{ p_1, p_2, \dots, p_{i-1}, p_{i+1}, \dots, p_m \}$

$$\left(\frac{V_i}{\;} \middle| \mathcal{F}_i \right) = \sum^{m} \mathbb{E}\left(\mathbb{1}\left\{ p_i \le \frac{k\alpha}{m} \right\} \middle| \mathcal{F}_i \right) \overset{\text{INDEP}}{\cdot} \mathbb{E}\left(\mathbb{1}\{ R(p_i \downarrow 0) = k \} \middle| \mathcal{F}_i \right)$$

$$\mathbb{E}\left(\frac{V_i}{R \vee 1} \,\Big|\, \mathcal{F}_i\right) = \sum_{k=1}^{m} \frac{\mathbb{E}\left(\mathbb{1}\left\{p_i \le \frac{k\alpha}{m}\right\} \Big| \mathcal{F}_i\right) \cdot \mathbb{E}\left(\mathbb{1}\left\{R(p_i \downarrow 0) = k \,|\, \mathcal{F}_i\right\}\right)}{k}$$

$$= \sum_{k=1}^{m} \frac{\left(\frac{\cancel{k}\alpha}{m}\right) \cdot \mathbb{1}\left\{R(p_i \downarrow 0) = k\right\}}{\cancel{k}}$$

$$= \frac{\alpha}{m} \underbrace{\sum_{k=1}^{m} \mathbb{1}\left\{R(p_i \downarrow 0) = k\right\}}_{=1}$$

$$= \frac{\alpha}{m}$$

TOWER PROPERTY

$$\mathbb{E}\left(\frac{V_i}{R \vee 1}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{V_i}{R \vee 1} \,\Big|\, \mathcal{F}_i\right)\right) = \frac{\alpha}{m} \quad \square$$

- BH CONTROLS FDR ALSO UNDER THE ASSUMPTION OF "POSITIVE DEPENDENCE"

- FOR THE CASE OF GENERAL DEPENDENCE

$$FDR \le \alpha H_m \quad \text{WITH } H_m = \sum_{i=1}^{m} \frac{1}{m} \quad \text{HARMONIC NUMBER}$$

BENJAMINI-YEKUTIELI PROCEDURE WHICH IS BH AT LEVEL α / H_m $\square$

$m = 6033$ GENES

REJECT AT $\alpha = 5\%$	478 REJECTIONS
BONFERRONI	3
BH	21