연세대학교 박상언 교수님의 비모수통계학 수업을 듣고 간단하게 정리하였다. 수업의 전반적인 내용은 **비모수적인 방법론의 접근방법, 검정과 추정, 비모수회귀** 이다.

<목차>

1. Independent sample의 경우
2. Dependent sample의 경우
3. CDF 이용 검정 (Kolmogorov-Smirnov type)
4. Nonparametric Density Estimation and Regression

1. Independent sample

(그룹끼리) independent한 표본의 경우 추정,검정에 대한 내용이다.

1.1.1 Binary response : one group

$H_0 : p = p_0$, Data : 크기 n의 binomial sample

- test statistic
  - T (성공개수, 주로 1의 개수)
- decision rule
  - exact rule : binomial dist로 $Pr(T \geq t/p = p_0)$을 계산
  - approximate rule : 우리가 일반적으로 알고 있는 large sample의 경우 normal dist로 근사

$$Z = \frac{T - np_0}{\sqrt{np_0(1-p_0)}}$$

1.1.2 Binary response : two group

.	class1	class2	total
population1	O_{11}	O_{12}	n_1
population2	O_{21}	O_{22}	n_2
total	C_1	C_2	N

1. Homogeneity (동질성)

- $H_0 : p_{11}/p_{1.} = p_{21}/p_{2.}$ (i.e. P(class1 / pop1) = P(class2 / pop2))
- test statistic : 원래는 이항분호를 통해 exact 하게 해야하지만 정규분포로 근사한다.

$$\frac{O_{11}}{n_1} \sim N(p_1, \frac{1}{n_1}\frac{C_1}{N}\frac{C_2}{N}), \frac{O_{21}}{n_2} \sim N(p_2, \frac{1}{n_2}\frac{C_1}{N}\frac{C_2}{N})$$

2. Independent (독립성)

- $H_0 : p_{ij} = p_{i.}p_{.j}$
- test statistic :

$$T = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi_1^2$$

동질성검정의 검정통계량을 제곱하면 독립성검정의 검정통계량과 같은 값을 가진다.

1.1.3 Polytomous response : several group

.	class1	class2	...	class c	total
population 1	O_{11}	O_{12}	...	O_{1c}	n_1
population 2	O_{21}	O_{22}	...	O_{2c}	n_2
...	...	...	...	...	...
population r	O_{r1}	O_{r2}	...	O_{rc}	n_r
total	C_1	C_2	...	C_c	N

1.1.2와 똑같이 두 가지 방법 가능

1.2.1 Numeric data : two group

각 표본을 rank를 매긴후에 검정하는 방법

- data : $\{X_1, X_2, ..., X_n\} \sim F, \{Y_1, Y_2, ..., Y_n\} \sim G$
- $H_0 : P(X < Y) = 0.5 \ (same \ as \ F = G)$
- $R(X_i)$ 는 해당 표본의 rank를 의미하며 가장 작은 수가 1, 가장 큰 수가 N의 값을 가진다.
- test statistic
  - Mann-Whitney test : $M = \sum_{i=1}^{m} \sum_{j=1}^{n} I(Y_i < X_j)$
  - Wilcoxon rank sum test : $W = \sum_{i=1}^{n} R(X_i)$
  - 관계 : $M = W - \frac{n(n+1)}{2}$
- decision rule
  - approximate p-value by 정규화 : $W = \frac{W - n\frac{N+1}{2}}{\frac{nm}{N-1}\frac{1}{N}\sum_{i=1}^{N}(R_i - \frac{N+1}{2})}$

1.2.2 Numeric data : Several group

이처럼 rank를 매긴다. (따라서 ANOVA처럼 정규성이 필요없다)

treatment / replication	.	sum
1	$R(X_{11})\ R(X_{12})\ ...\ R(X_{1n_1})$	R_1
2	$R(X_{21})\ R(X_{22})\ ...\ R(X_{2n_2})$	R_2
...	...	...
k	$R(X_{k1})\ R(X_{k2})\ ...\ R(X_{kn_k})$	R_k

- H_0 : All of k population distribution function are identical
- TSS : $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (R(X_{ij}) - \frac{N+1}{2})^2$
- SST : $\sum_{i=1}^{k} n_i(\bar{R}_i - \frac{N+1}{2})^2$
- SSE :$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (R(X_{ij}) - \bar{R}_i)^2$
- decision rule
  - Kruskal-Wallis Test : $T = \frac{SST}{TSS/(N-1)} \sim \chi^2_{k-1}$
  - F-test : 원래 정규분포인 경우만 해당하지만 근사적으로 사용가능. 하지만 Kruskal 더 선호.

2. Dependent sample

data가 paired data인 경우이다. 예를 들어, 질병의 여부에 대해 조사를 하는데 (엄마, 딸) 이런 식의 dependent한 결과를 내는 데이터들을 분석하는데 있어서 이전의 independent한 경우와는 다른 접근방법이 필요하다.

2.1.1 Binary response : two dependent group

.	Y = 0	Y = 1
X = 0	a	b
X = 1	c	d

- $H_0 : P(X=0) = P(Y=0)$ (marginal homogeneity)
- method1 : McNemar test
  - under assumption, $b \sim N(\frac{1}{2}(b+c), \frac{b+c}{4})$
  - test statistic : $T = \frac{(b-c)^2}{b+c} \sim \chi_1^2$
- method2 : Covariance term
  - $V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) - 2COV(\bar{X}, \bar{Y})$
  - $COV(\bar{X}, \bar{Y}) = \frac{1}{n}COV(X, Y) = \frac{1}{n}(E(XY) - E(X)E(Y)) = \frac{1}{n}(p_{11} - p_{1+}p_{+1}) = \frac{1}{n}(p_{00}p_{11} - p_{10}p_{01})$
- method3 : Cochran's Q-test
  - $H_0 : p_{i1} = p_{i2}\ for\ all\ i$

since $V(X_i) = p_{i1}(1-p_{i1})$, 귀무가설 하에서

- $V(X_i) = V(Y_i) = \frac{R_i}{2}(1 - \frac{R_i}{2})$
- $cov(X_i, Y_i) = cov(X_i, R_i - X_i) = -V(X_i)$
- 따라서 correlation이 -1
- test statistic : $\frac{\sum_{i=1}^{2}(C_i - \frac{C_1+C_2}{2})^2}{\sum_{i=1}^{r}\frac{R_i}{2}(1-\frac{R_i}{2})*2} \sim \chi_1^2$
- Lemma
  - $(W_1, ..., W_c)$ has a multivariate normal distribution with common mean and common variance and common correlation, then $\sum_{i=1}^{c}(W_i - \bar{W})^2/\sigma^2(1-\rho)$ has a chi-squared distribution with df(c-1)
- method4 : Paired Difference
  - 우리가 흔히 알고 있는 방법(통입때 배움)
- 4가지 방법 모두 같은 검정통계량이 나온다.

2.1.2 Binary response : more than dependent group

- Cochran's Q-test

2.2.1 Numeric response : two dependent group

- data : $(X_1, Y_1), ..., (X_n, Y_n)$
- $H_0 : P(+) = P(-), where\ P(+) = Pr(X > Y)$
- Sign Test
  - test statistic (tie는 버린다)
    - $T = total\ number\ of\ +$
  - decision rule
    - exact rule : binomial 이용 $p(T \geq t / B(n, 0.5))$
    - approximate rule : $\frac{T - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$
- Wilcoxon Signed Rank Test
  - sign test와 다른 점은 차이에 대해 가중치를 주는 것
  - Laplace dist 이외에는 sign test보다 most powerful
- test statistic
  - exact : $T^+ = \sum I(sign(D_i) = +)R_i$
  - approximate : $T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$

2.2.2 Numeric response : more than two dependent group

block / treatment	1	2	...	k
1	X_{11}	X_{12}	...	X_{1k}
2	X_{21}	X_{22}	...	X_{2k}
...	...	...	...	...
b	X_{b1}	X_{b2}	...	X_{bk}

- H_0 : each ranking of the random variables within a block is equally likely
- within block에서 rank를 매기므로 block effect가 사라진다.
- Friedman Test
    - $TSS^A : \sum_{i=1}^{b}\sum_{j=1}^{k}(R(X_{ij}) - \frac{k+1}{2})^2$
    - $SST : b\sum_{j=1}^{k}(\bar{R}_j - \frac{k+1}{2})^2$
    - $SSE : \sum_{i=1}^{b}\sum_{j=1}^{k}(R(X_{ij}) - \bar{R}_j)^2 = TSS^A - SST$
    - test statistic

$$T = \frac{SST}{TSS^A/b(k-1)} \sim \chi^2_{k-1}$$

- Quade Test
    - consider block range : $S_{ij} = Q_i[R(X_{ij}) - \frac{k+1}{2}]$

3. CDF 이용 검정 (Kolmogorov-Smirnov type)

3.1 Comparison of CDF's : one group

- data : independent random sample of size n from $F_n(x)$
- $H_0 : F(x) = F_0$
- Kolmogorov Goodness of Fit Test
    - $H_1 : F(x) \neq F_0, T = sup_x|F_0(x) - F_n(x)|$
    - decision rule
        - exact
        - asymptotic : $P(\sqrt{n}T^+ \leq t) \approx 1 - exp(-2t^2)$
- Cramer-von Mises Test
    - average of the squared difference :

$$C_n^2 = \int (F_n(x) - F_0(x))^2 dF_0(x)$$

    - $nC_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}(F_0(x_{i:n}) - \frac{2i-1}{2n})^2$
- Anderson-Darling Test
    - 표준화해서 비교한다.
    - power가 가장 좋다. (덜 보수적)
    - 이전 두가지 test보다 양 끝점의 이야기를 들어주는 편이다.
    - average of the standardized squared difference :

$$A_n^2 = n\int \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x)$$

    - $A_n^2 = \sum_i^n \frac{2i-1}{n}(logF_0(x_{i:n}) + log(1 - F_0(x_{n+1-i:n})))$

3.2 Comparison of CDF's : two group

- data : $\{X_1, X_2, ..., X_n\} \sim F, \{Y_1, Y_2, ..., Y_m\} \sim G$

- $H_0 : F(x) = G(x)$
- Kolmogorov Goodness of Fit Test
  - test statistic : $T = sup_x|F_n(x) - G_m(x)|$
- Cramer-von Mises Test
  - $H_{mn}(x) = nF_n(x)/(n+m) + mG_m(x)/(n+m)$

$$C_n = \frac{nm}{n+m}\{\sum_{i=1}^{n}(F_n(x_i) - G_m(x_I))^2 + \sum_{i=1}^{m}(F_n(y_i) - G_m(y_i))^2\}$$

3.3 Numerical Summary : AUC

- auc가 커질수록 두 집단은 다른 것 (잘 나누어졌다 in classification)

3.4 Comparison of quantile : one & two group

- qqplot

4. Nonparametric Density Estimation and Regression

4.1 Nonparametric Density Estimation

- Nonparametric maximum likelihood
  - Empirical (Nonparametric) likelihood : $p_i = Pr(X = x_i)$

$$L(F) = \prod_{i=1}^{n} p_i$$

- 이에 대한 해는 $\hat{p}_i = \frac{1}{n}$

우리가 지금까지 사용해왔던 EDF는 미분이 불가능하다. 따라서 density를 구하기 위해서는 다른 방법들을 생각해야 한다.

- Histogram
  - 우리가 흔히들 알고 있는 히스토그램으로 f 를 estimate
- Naive density estimate
  - $f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h}$ 을 이용하여

$$\hat{f}(x) = \frac{1}{2h}(F_n(x+h) - F_n(x-h))$$

- Kernel estimate

$$\hat{f}(x) = \frac{1}{nh}\sum_{j=1}^{n} K(\frac{x - X_j}{h})$$

- kernel function K를 이용한다. 주로 K는 unimodal and symmetric.
- k와 h를 잘 정해야 한다.

4.2 Nonparametric Regression

- regression function : $Y_i = m(X_i) + \epsilon_i$
- Nonparametric 방법은 위의 m(x)를 일반적인 회귀처럼 dist of the error and functional form of m(x) 를 가정하지 않는다.
- Smoothing, Local Averaging
  - local weighted average : $\hat{m}(x) = \sum_{i=1}^{n} w_i(x)Y_i$
  - 특정 point의 m을 estimate하기 위해서 해당 point 근처에 있는 y values를 평균을 낸다. 근처의 s개의 observation으로 평균을 낸다고 하면 s를 span이라고 부른다.

- span을 고르는 방법은 Leave-out cross validation 방법사용

- Nadaraya-Watson estimator

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h})Y_i}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})}$$

- Limitation of NW estimator and other methods
  - NW estimator는 local constant fit이기 때문에 이를 조금 더 improve하면

$$min\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1(X_i - x))^2 K(\frac{x-X_i}{h})$$

- spline : $\sum_{i=1}^{n}(Y_i - m(X_i))^2 + \lambda\int(m''(u))^2du$

번외 : Model evaluation

Confusion matrix

Actual / predict	Positive	Negative
Positive	a	b
Negative	c	d

- evaluation based on Actual

1. Sensitivity(TPR, recall) : $\frac{a}{a+b}$
2. Specificity(TNR) : $\frac{d}{c+d}$

- evaluation based on Predict

1. Precision(PPV) : $\frac{a}{a+c}$
2. NPV : $\frac{d}{b+d}$

Evaluation Index

- Accuracy = $(\frac{a+b}{a+b+c+d})TPR + (\frac{c+d}{a+b+c+d})TNR = \frac{a+b}{a+b+c+d}$
- F1 score = $\frac{2}{\frac{1}{Recall}+\frac{1}{Precision}}$: recall 과 precision의 조화평균

The effect of the proportion in the learning sample

- 모델을 training 할 때(2 class classification), 모집단의 비율에 맞춰서 가장 좋지만 그렇지 못한 경우가 많다. 이런 경우는 cutoff에 대해 잘 고민해야 한다. 일반적으로 cutoff를 0.5로 한다(unbalanced가 아닌 경우가 많으니). 하지만 training sample과 validation sample의 balanced 정도가 다르다면 cutoff를 조정해야한다.
- unbalanced in classification : training sample는 주로 overfitting을 막기 위해 1:1로 한다. 하지만 validation sample이 unbalanced 이고 이를 통해 model evaluation을 하면 TPR, PNR을 영향을 받지 않지만 PPV, NPV 가 영향을 받는다. 주의해야한다.