



브레인 PhD 논문특강 논문 통계 특강

2021년 12월 18일

목차

- 통계의 정의
- 통계 용어와 자료 정리
- 가설의 설정
- 기본 통계분석법의 종류와 선택 (모수적 기법)
- 기본 통계분석법의 종류와 선택 (비모수적 기법)
- 실전 통계분석

통계의 정의

통계 (Statistic)

- 분석의 대상이 되는 집단에 대하여 실시한 조사 및 실험의 결과로 얻어진 관측값(수치) 또는 그 관측값들의 요약된 형태를 의미

통계학 (Statistics)

- 특별한 의도나 목적을 가지고 조사·연구할 때 자료의 수집방법뿐만 아니라 수집한 자료를 과학적이고 논리적인 이론에 의하여 정리·분석하는 학문
- 관심의 대상에 대한 관련된 자료를 수집하고, 그 자료를 요약·정리하여 해석하며, 이로부터 불확실한 사실에 대한 결론이나 일반성 등을 이끌어내는데 필요한 이론과 방법을 제시해 주는 학문

☞ 우리에게 통계란?



통계

익숙한 통계

- 표/그래프를 해석하면 어떤가요?
- 동전을 5번 던질 때 앞면이 2번 나올 확률은 얼마?
- 주머니 속에 빨간 구슬 10개 파란 구슬 7개...
- 표준편차 공식은? 정규분포함수 식은? 신뢰구간은?



통계학의 분류

기술통계학 (descriptive statistics)

- 자료를 편리하게 정보를 나타내는 방식으로 정리, 요약, 설명하는 방법론
- 수치 기법, 그래프 기법

추론통계학 (inferential statistics)

- 모집단으로부터 얻은 표본자료를 사용하여 모집단 전체에 대한 특성을 추측하고, 모집단에 대한 일련의 의사결정방법을 연구하는 방법론

통계 용어 및 자료의 정리

모집단 (population)

- 조사 혹은 관심 대상이 되는 원소 하나 하나의 전체 집합이나 모임
- 일반적으로 모집단은 상당히 크고, 무한할 수도 있음.

표본 (sample)

- 모집단의 특성을 파악하기 위하여 추출된 모집단의 일부 (subset)
- 모집단의 일부분인 표본으로부터 조사된 자료를 표본자료(sample data)라고 함.

통계 용어 및 자료의 정리

변수 (variable)

• 독립변수

- 실험에서 자극을 주는 변수
- 어떤 것의 원인이 되는 변수
- 영향을 주는 변수

• 종속변수

- 자극에 대한 반응이나 결과를 나타내는 변수
- 독립변수의 변화에 따라 변함

Ex) A약물 주입 후 혈관 직경이 변화할 것이다.

통계 용어 및 자료의 정리

척도에 따른 데이터 분류

- 데이터 - 정성적 (Qualitative) / 정량적 (Quantitative) 데이터 (“사칙연산”)
 - 범주형 (Categorical) / 연속형 (Continuous) 데이터
 - ex) 혈액형, 성별, 질병반응 (분류형) / 신장, 나이, 연봉 (수치형)
- 명목척도: 응답안들을 상호배타적으로 분류하기 위하여 임의의 문자나 숫자로 구분 (순서상관 없음) ex) 성별, 지역, 직업 등
- 서열척도: 조사대상들의 특성을 서열로 나타냄 ex) 선호도, 직급 등
- 간격척도: 범주, 서열, 거리의 정보를 가짐 (숫자간의 차이는 절대적인 의미)
 - ex) 리커트 척도 (1-5; very bad – very good)
- 비율척도: 범주, 서열, 거리, 비율의 정보를 가짐 (각각의 값이 절대적인 의미, 평균 및 산술계산 가능)

변수의 척도/인과관계에 따른 통계 기법 선택

독립변수	종속변수	목적	모수적 검정법	비모수적 검정법
범주형	범주형	두 변수간의 관계	카이제곱 검정	Fisher's Exact Test McNemar Test
범주형 (범주 n=2)	연속형	두 그룹간의 평균 비교	독립표본 T 검정	Mann-Whitney Test Kolmogorov-Smirnov Test Wald-Wolfowitz runs Test
			대응표본 T 검정	Wilcoxon signed rank Test Sign Test
범주형 (범주 >2)	연속형	여러 그룹간의 평균 비교	분산분석 (1-way ANOVA)	Kruskal-Wallis Test Median Test
			분산분석 (2-way ANOVA)	Friedman's 2-way ANOVA
연속형	연속형	두 변수간의 관계	상관분석	Spearman's Correlation Kendall's tau test Stuart's tau test
			회귀분석	

가설의 설정

가설

- "그럴것이다 " 라고 추측하는 부분
 - 귀무가설 (null hypothesis) = 영가설(H_0)
 - 연구자가 주장하는 대립가설과 반대
 - 연구자 입장에서는 원치 않는 (기각되기를 바라는) 가설
 - 대립가설 (alternative hypothesis) = 연구가설(H_1)
 - 연구자가 원하는 주장 혹은 연구자의 신념
- 연구자의 최종목적은 귀무가설을 기각하고 대립가설을 채택하는 것!

가설의 설정

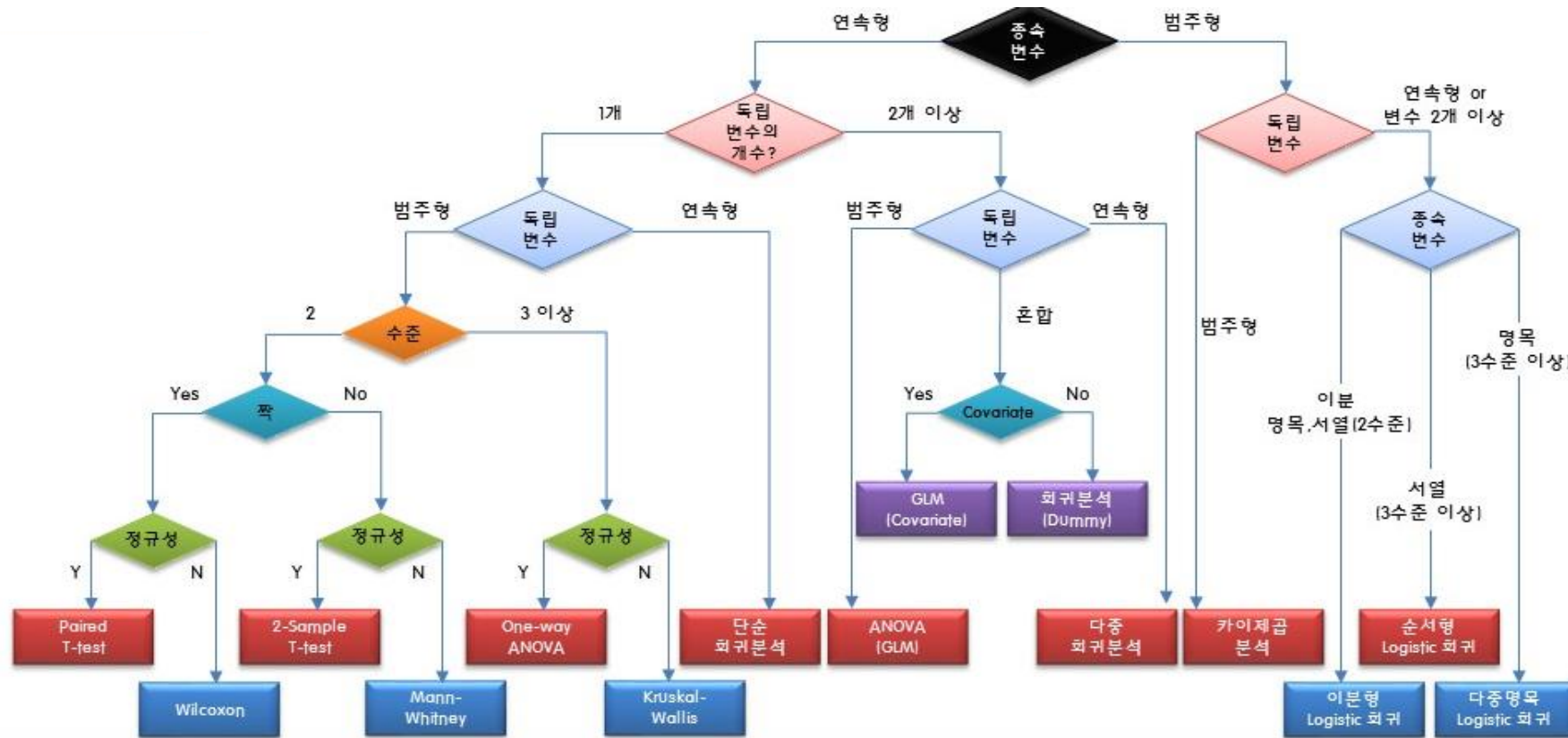
성별에 따른 혈류 속도의 차이가 있는가?

- 검정 방법: t-test (두 그룹의 평균 비교)
 - 귀무가설: 성별에 따른 혈류 속도의 차이가 없다.
 - 대립가설: 성별에 따라 (남성과 여성은) 혈류 속도에 차이가 있다.

피부질환 크기에 따라 반응정도의 차이가 있는가? (피부질환크기 (>1 , $1\sim0.5$, <0.5) / 반응정도: neg, pos)

- 검정 방법: chi-square test
 - 귀무가설: 피부질환 크기에 따라 반응 정도에 차이가 없다.
 - 대립가설: 피부질환 크기에 따라 반응 정도에 차이가 있다.

통계분석법의 선택



통계분석 소프트웨어

엑셀 (excel)

- 사용이 용이하며, 간단한 통계분석 가능

미니탭 (minitab)

- 교육용으로 제작, 사용이 용이

SPSS(Statistical Package for Social Sciences)

- 사회과학을 위한 통계패키지로 현재 많은 분야에서 사용
- 메뉴방식으로 사용이 편리

SAS(Statistical Analysis System)

- 통계전문가를 위한 통계패키지
- 다양한 통계분석뿐만 아니라 대용량의 자료처리 가능

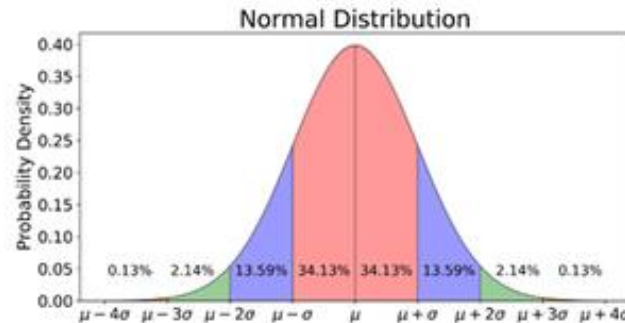


정규성 검정



정규분포 (Normal Distribution)

- 특정 값들의 출현 비율이 중심(평균)을 기준으로 좌우 대칭 형태



$z\sigma$	Percentage within CI	Percentage outside CI
0.674 σ	50%	50%
1 σ	68.268 9492%	31.731 0508%
1.645 σ	90%	10%
1.96 σ	95%	5%
2 σ	95.449 9736%	4.550 0264%
2.576 σ	99%	1%
3 σ	99.730 0204%	0.269 9796%
3.2906 σ	99.9%	0.1%
4 σ	99.993 666%	0.006 334%
5 σ	99.999 942 6697%	0.000 057 3303%
6 σ	99.999 999 8027%	0.000 000 1973%
7 σ	99.999 999 999 7440%	0.000 000 000 2560%

- 정규성 검정은 표본의 모집단이 정규분포를 이루는지에 대한 분석
 - 귀무가설: 표본의 모집단이 정규분포를 이루고 있음
 - 대립가설: 표본의 모집단이 정규분포를 이루고 있지 않음
 - 유의수준 95%신뢰구간에서 p값이 0.05보다 크게 나타나야 귀무가설을 채택할 수 있음.



정규성 검정



정규성 검정 방법

- Kolmogorov-Smirnov (K-S test)
 - 50개 이상의 대표본에 적합 (절대적 기준은 아님)
- Shapiro-Wilk (S-W test)
 - 50개 미만의 소표본에 적합 (절대적 기준은 아님)
- 두 가지 검정방법에서 어느 한 개라도 $p > 0.05$ 일 경우, 해당 표본은 정규성을 만족한다고 가정
- 정규성 검정 – 정규분포 – 모수검정/비모수검정

기본 통계분석법의 종류와 선택 (모수적 기법)

단일표본 t 검정 (1 sample t-test)

- 하나로 구성된 모집단의 평균값을 기준값과 비교하고자 할 때 사용
- 독립변수에 따른 종속변수는 정규분포를 만족해야 함
(일반적으로 범주 별 30개 이상의 표본을 확보하는 경우 모수검정으로 고려)

Ex) 제주대학교 대학원 학생들의 평균키가 175cm 일 때, 의학계열 대학원 학생 20명의 평균키를 측정하여 제주대학교 대학원 학생들의 평균키와 비교하고자 할 때 사용되는 분석방법

기본 통계분석법의 종류와 선택 (모수적 기법)

독립표본 t 검정 (2 sample t-test)

- 종속변수가 연속형으로 측정, 독립변수는 2개 수준의 범주형이 상태에서 짝을 이루지 않고 정규성을 만족할 때 사용
- 분석하고자 하는 변수가 나이, 키 등과 같은 연속변수로 평균과 표준편차로 나타내어 질 수 있는 형태의 자료를 분석 (두 집단 간 평균의 차이를 비교)
- 가정사항
 - 독립성: 독립변수의 그룹 군은 서로 독립
 - 정규성: 독립변수에 따른 종속변수는 정규분포를 만족
 - 등분산성: 독립변수에 따른 종속 변수 분포의 분산(변량이 평균으로부터 떨어져 있는 정도를 나타내는 값)은 각 군마다 동일해야 함
("Levene의 등분산 검정" 확인)
- 등분산 만족(= 분산같음, $p > 0.05$)하는지 여부에 따라 해당 통계량을 확인해야 함

✍ 기본 통계분석법의 종류와 선택 (모수적 기법)

■ 독립표본 t 검정 (2 sample t-test)

➔ T 검정

그룹 통계				
그룹구분	N	평균	표준 편차	표준 오차 평균
유산농도 결핵성	30	4.130	1.4118	.2578
바이러스성	30	2.050	.8038	.1467

독립표본 검정

		Levene의 등분산 검정		평균 등식에 대한 T 검정						
		F	유의수준	t	df	유의수준 (양쪽)	평균 차이	표준 오류 편차	차이의 95% 신뢰구간	
									하한	상한
유산농도	등분산을 가정할	4.964	.030	7.013	58	.000	2.0800	.2966	1.4863	2.6737
	등분산을 가정하지 않을			7.013	46.012	.000	2.0800	.2966	1.4830	2.6770

기본 통계분석법의 종류와 선택 (모수적 기법)

대응표본 t 검정 (paired t-test)

- 종속변수가 연속형으로 측정, 독립변수 1개, 종속변수 1개 (독립변수는 1개이지만 수준이 2개로서 서로 짝을 이루고 있음)
- 실험 이전의 집단과 실험 이후의 집단이 동일한 집단인 경우 사용, 한 개인이 서로 다른 두 조건에서 짝을 지어 한 쌍이 연구대상이 되는 경우
- 가정사항
 - 정규성: 독립변수에 따른 종속변수는 정규분포를 만족

Ex) A약물 주입 전후에 콜레스테롤 값의 차이를 확인

기본 통계분석법의 종류와 선택 (모수적 기법)

일원배치 분산분석 (one-way analysis of variance (ANOVA))

- 1개 독립변수의 수준이 3개 이상인 경우 적용
- 3개 이상의 범주를 갖는 독립변수에 대한 종속변수의 평균 차이를 검정
- 가정사항
 - 독립성: 독립변수의 그룹 군은 서로 독립적
 - 정규성: 독립변수에 따른 종속변수는 정규분포를 만족해야 함
 - 등분산성: 독립변수에 따른 종속변수 분포의 분산은 각 군마다 동일해야 함
- 대립가설: ~의 변화에 따라 "적어도 하나의 그룹"에서 ~차이가 있다 (모든 그룹 아님)

기본 통계분석법의 종류와 선택 (모수적 기법)

일원배치 분산분석 (one-way analysis of variance (ANOVA))

- 사후분석 (post-hoc analysis) 수행: 어느 그룹간 차이가 발생했는지 확인함



- Tukey: 집단간 차이를 가장 정밀하게 감지, 각 그룹의 샘플수가 같을 때 사용
- Bonferroni: 각 그룹의 샘플수가 다를 때 사용
- Duncan: 사회과학 분야에서 주로 사용

기본 통계분석법의 종류와 선택 (비모수적 기법)

Mann-Whitney U test (Wilcoxon rank sum)

- 독립표본 t 검정 (2 sample t-test)의 수행할 수 있는 3가지의 기본 가정 사항(정규성, 독립성, 등분산성)중 만약 정규성을 만족하지 못하는 경우 수행 (비모수통계)
- 두 집단의 분포가 동일한지를 조사하는 기법, 두 집단의 관측치가 통합이 되고 크기 순으로 순위가 부여
- 가정사항: 없음
- 보건의료계통에서 적은샘플수로 통계 검증을 하고자 할 때, 주로 사용

기본 통계분석법의 종류와 선택 (비모수적 기법)

Wilcoxon signed-rank test

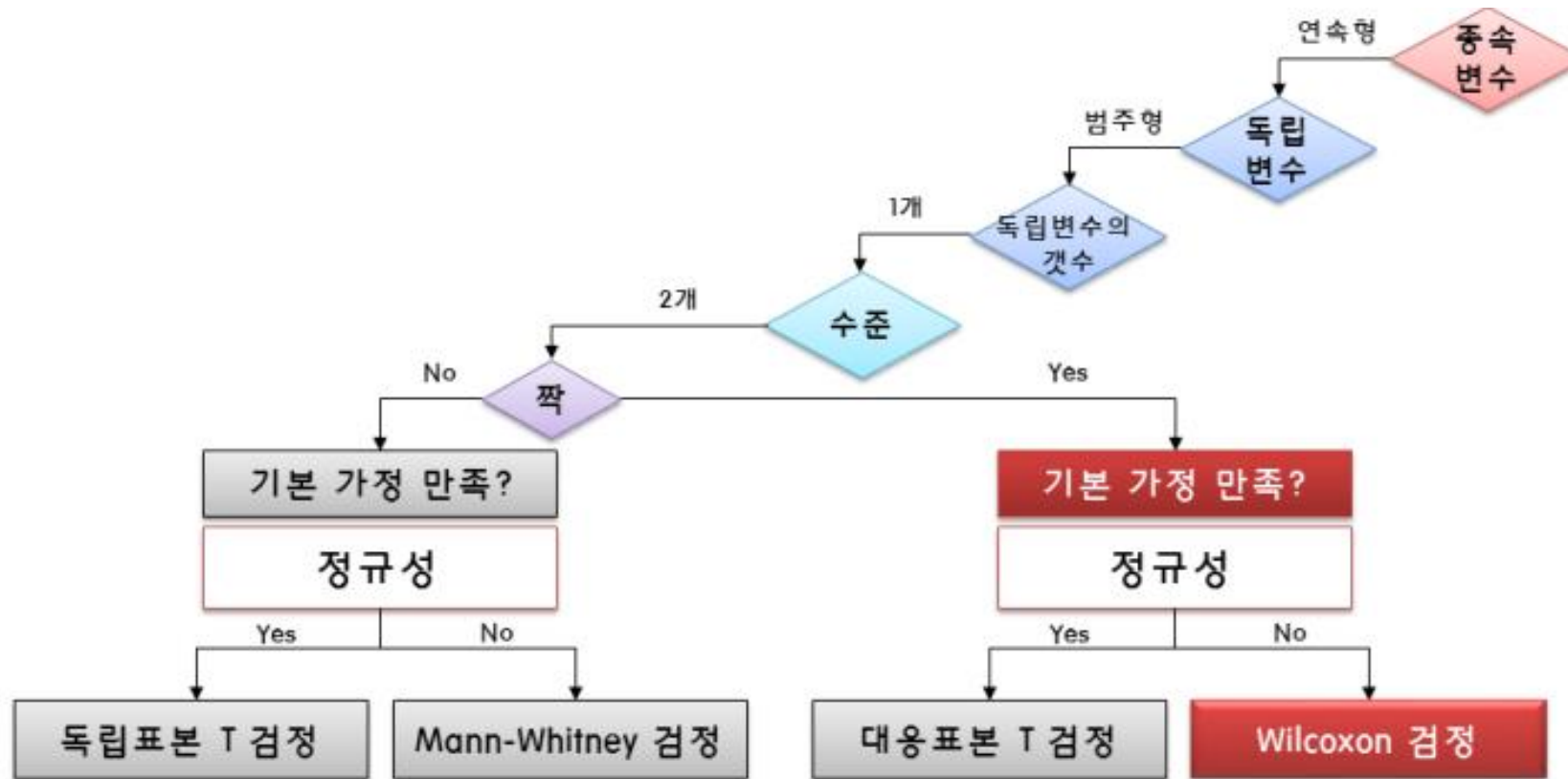
- Paired t-test와 맞 대응되는 (정규성 만족하지 못하는 경우) 비모수적 통계기법
- 두 집단이 짝을 이루고 있지만 정규성을 만족하지 못하는 경우에 사용
- 가정사항: 없음
- 보건의료계통에서 적은샘플수로 통계 검증을 하고자 할 때, 주로 사용

기본 통계분석법의 종류와 선택 (비모수적 기법)

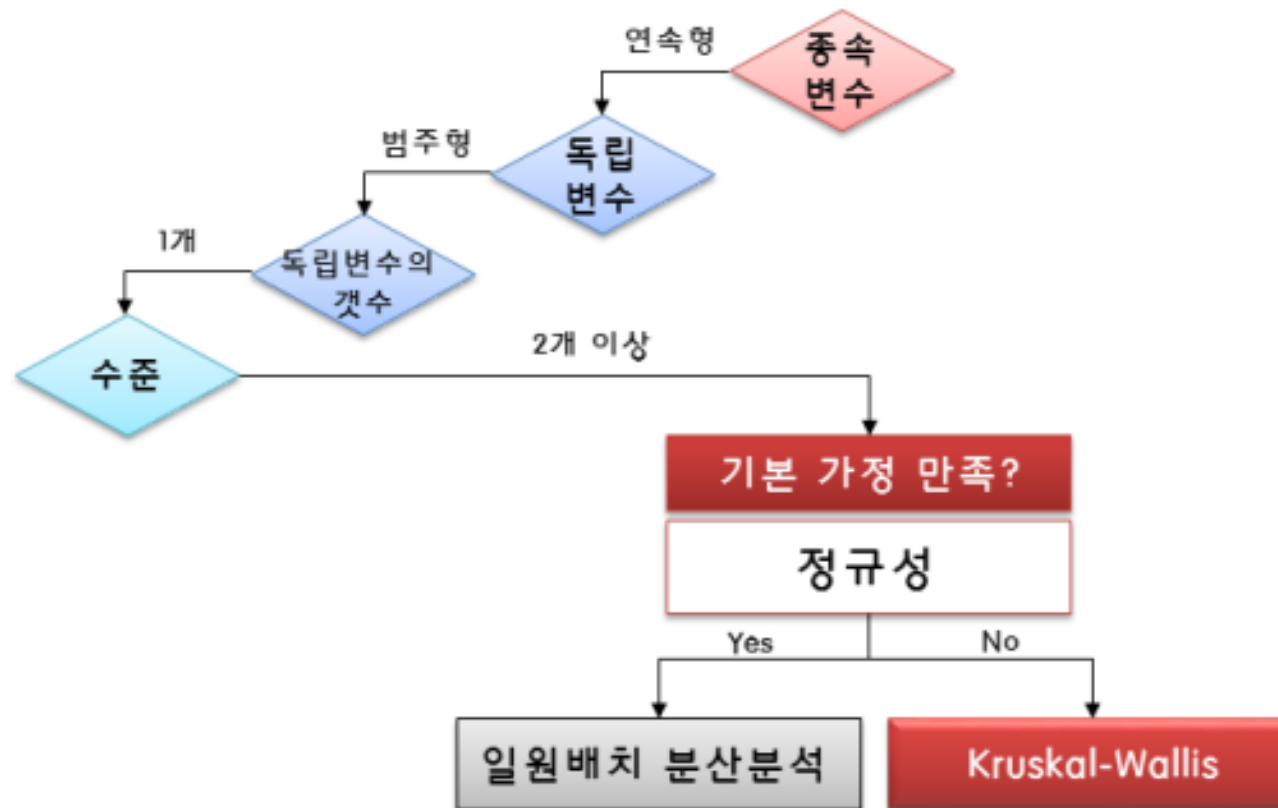
Kruskal-Wallis H test

- One-way ANOVA의 비모수검정, 세 집단 이상의 집단 분포를 비교
- 가정사항: 없음
- 사후검정 (post-hoc analysis)
 - Bonferroni Correction method 이용
 - Kruskal-Wallis H test에서 집단간 유의성이 있을 때, 각 집단을 짝으로 묶어서 t-test의 비모수적인 방법인 Mann-Whitney U검정 실시
(단, 유의수준은 Mann-Whitney U검정 수로 나누어 줌)
Ex) 3개의 집단의 경우 사후검정은 세번 실시해야 하므로 (a-b, b-c, a-c)
이때, 유의수준은 $0.05/3=0.0167$ 을 사용)

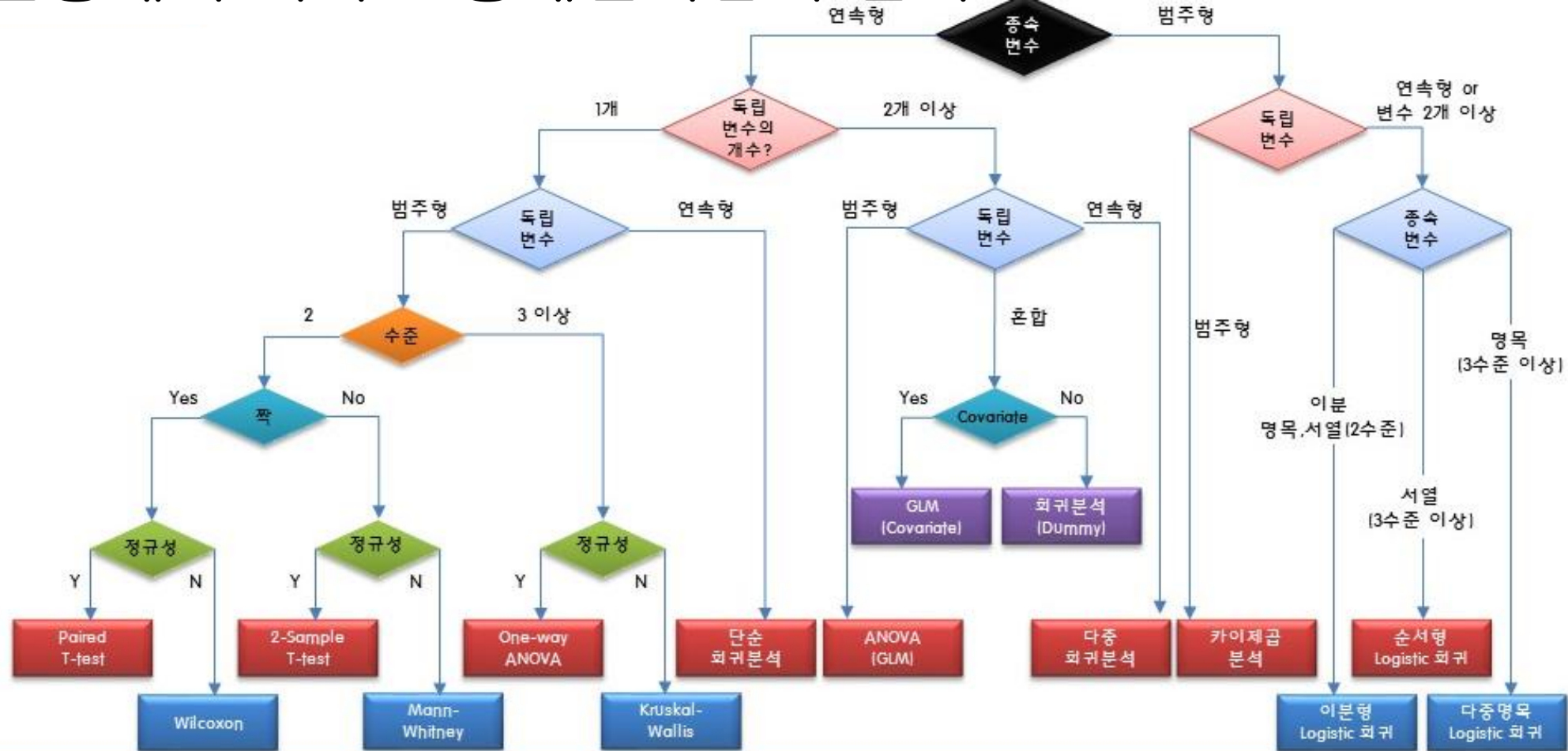
☞ 평균비교-분석 (2그룹)



☞ 평균비교-분석 (3그룹 이상)



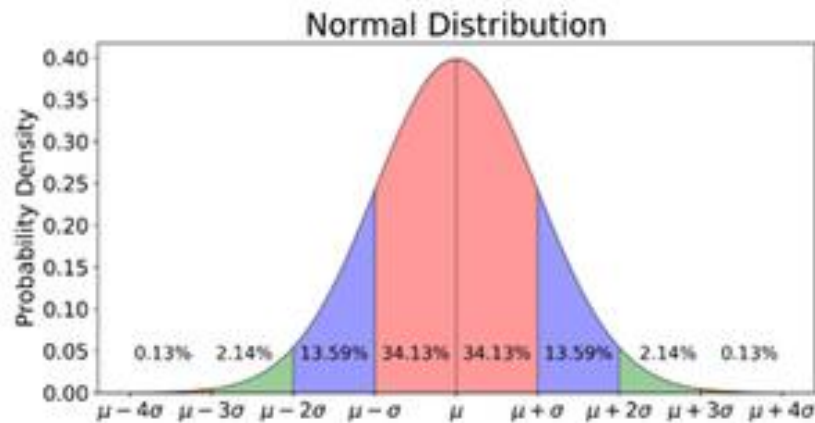
실전통계의 시작 - 통계분석법의 선택



1) 정규성 검정

■ 정규분포 (Normal Distribution)

- 특정 값들의 출현 비율이 중심(평균)을 기준으로 좌우 대칭 형태



$z\sigma$	Percentage within CI	Percentage outside CI
0.674σ	50%	50%
1σ	68.268 9492%	31.731 0508%
1.645σ	90%	10%
1.96σ	95%	5%
2σ	95.449 9736%	4.550 0264%
2.576σ	99%	1%
3σ	99.730 0204%	0.269 9796%
3.2906σ	99.9%	0.1%
4σ	99.993 666%	0.006 334%
5σ	99.999 942 6697%	0.000 057 3303%
6σ	99.999 999 8027%	0.000 000 1973%
7σ	99.999 999 999 7440%	0.000 000 000 2560%

■ 정규성 검정은 표본의 모집단이 정규분포를 이루는지에 대한 분석

- 유의수준 95%신뢰구간에서 p값이 0.05보다 크게 나타남 (정규분포 만족)

■ 정규성 검정 방법

1) Kolmogorov-Smirnov (K-S test)

- 50개 이상의 대표본에 적합 (절대적 기준은 아님)

2) Shapiro-Wilk (S-W test)

- 50개 미만의 소표본에 적합 (절대적 기준은 아님)

- 두 가지 검정방법에서 어느 한 개라도 $p > 0.05$ 일 경우,
해당 표본은 정규성을 만족한다고 가정

- 정규성 검정 - 정규분포 - 모수검정/비모수검정

2) 기술통계, 신뢰구간

■ 기술통계량

(평균, 표준편차 등) 확인

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
나이 * 성별	50	100.0%	0	.0%	50	100.0%
신장 * 성별	50	100.0%	0	.0%	50	100.0%
체중 * 성별	50	100.0%	0	.0%	50	100.0%

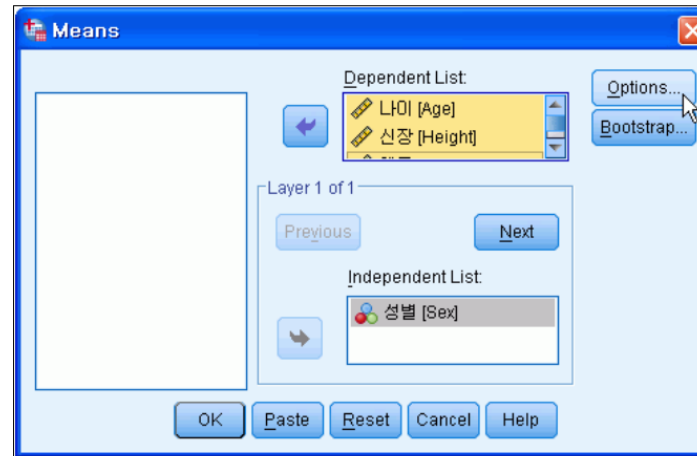
Report

성별		나이	신장	체중
Male	Mean	52.85	169.19	64.50
	N	26	26	26
	Std. Deviation	17.334	10.088	15.700
	Std. Error of Mean	3.399	1.978	3.079
Female	Mean	58.37	171.38	70.21
	N	24	24	24
	Std. Deviation	16.862	8.293	11.692
	Std. Error of Mean	3.442	1.693	2.387
Total	Mean	55.50	170.24	67.24
	N	50	50	50
	Std. Deviation	17.162	9.242	14.079
	Std. Error of Mean	2.427	1.307	1.991

	Sex	Age	Height	Weight
1	0	51	186	67
2	0	79	170	57
3	1	73	161	86
4	1	44	157	80
5	1	30	182	57
6	0	28	157	66
7	1	51	168	49

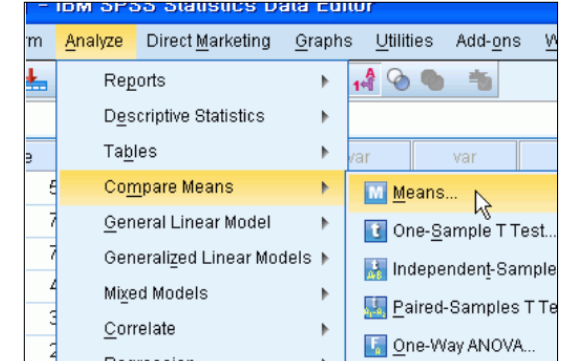
1. Data 입력

Sex, Age, Height, Weight



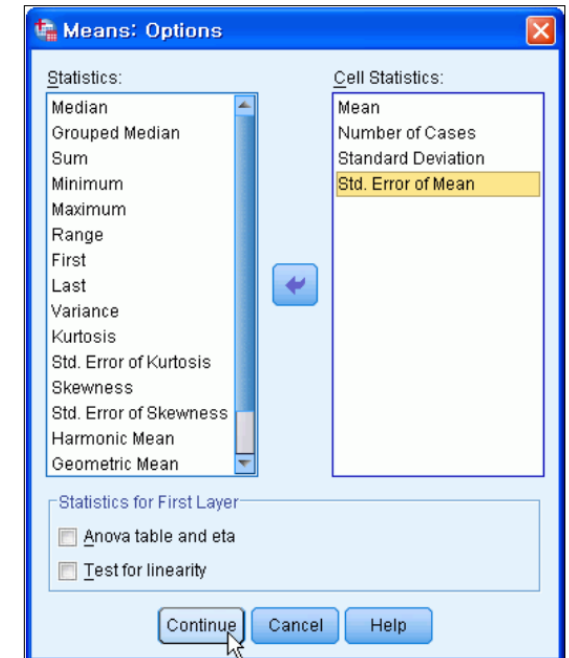
3. 설정

[Dependent List]에 나이, 신장, 체중을,
[Independent List]에 성별을 넣고
[Options]를 누른다.



2. 선택

[MENU]-[Analyze]-[Compare Means]-[Means]



4. [Options]

[Statistics]에서 원하는 항목을 선택하여 [Cell Statistics]에 넣고,
[Continue]-[OK]를 누른다.

2) 기술통계, 신뢰구간

■ 신뢰구간 (confidence interval, CI)

- 이상적 분석: 알고자 하는 집단 전체(모집단)를 조사하는 것.

불가능한 경우가 많고, 비효율적

- 통계적 분석: 집단 전체(모집단)에서 일부의 표본을 추출하여 그 표본의 정보를 분석하여 모집단의 정보를 추측하는 방법.

경제적이지만, 어느 정도의 오차가 있을 수 있다.

→ 표본의 평균을 이용하여 모집단의 평균을 추정하는 경우 어느 정도의 오차가 있을 수 있으므로 특정한 하나의 값으로 나타낼 수 없다. 그러므로 모집단의 평균은 일정한 구간으로 추정할 수 있는데 이를 신뢰구간이라 한다.

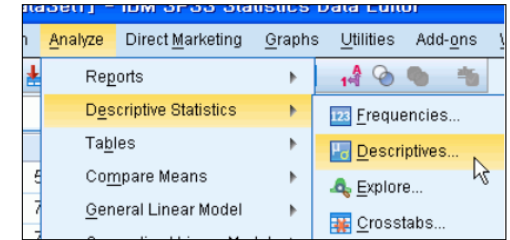
2) 기술통계, 신뢰구간

■ 신뢰구간 (confidence interval, CI)

	Sex	Age	Height	Weight
1	0	51	186	67
2	0	79	170	57
3	1	73	161	86
4	1	44	157	80
5	1	30	182	57
6	0	28	157	66

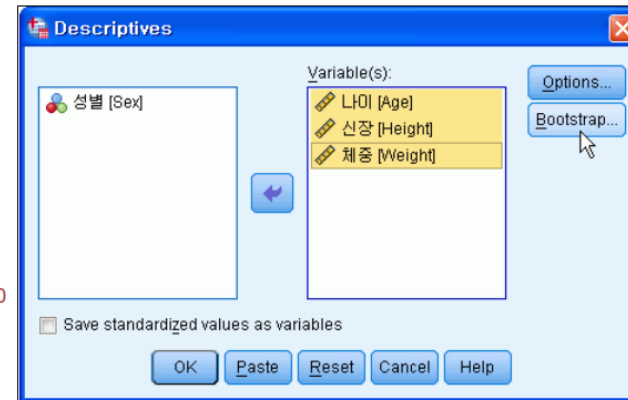
1. Data 입력

Sex, Age, Height, Weight



2. 선택

[MENU]-[Analyze]-[Descriptive Statistics]-[Descriptives]

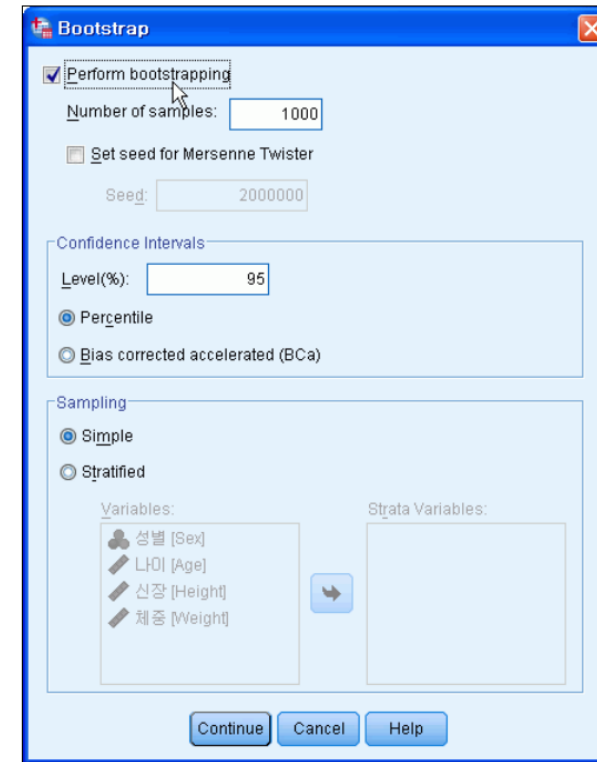


3. 설정

[Variables]에 나이, 신장, 체중을 넣고

[Bootstrap]을 누른다.

Sex는 연속변수가 아니므로 해당사항이 없음.



4. [Bootstrap]

[Perform bootstrapping]을 선택하고,

[Continue]-[OK]를 누른다.

Descriptive Statistics

	Statistic	Bias	Std. Error	Bootstrap ^a	
				95% Confidence Interval	
				Lower	Upper
나이	N	50	0	50	50
	Minimum	27			
	Maximum	82			
	Mean	55.50	.07	2.33	51.26 60.36
	Std. Deviation	17.162	-.248	1.132	14.641 19.033
신장	N	50	0	50	50
	Minimum	155			
	Maximum	186			
	Mean	170.24	.02	1.32	167.76 172.90
	Std. Deviation	9.242	-.118	.593	7.891 10.228
체중	N	50	0	50	50
	Minimum	46			
	Maximum	92			
	Mean	67.24	.03	1.98	63.52 71.36
	Std. Deviation	14.079	-.143	.922	12.128 15.708
Valid N (listwise)	N	50	0	50	50

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

나이의 평균은 55.50, 95% CI 은 51.26 - 60.36

신장의 평균은 170.24, 95% CI 은 167.76-172.90

체중의 평균은 67.24, 95% CI 은 63.52-71.36

3) 신뢰도분석 (intra-class correlation coefficient, ICC)

■ 관찰자간, 관찰자내의 신뢰도 검정 (ex., tumor size)

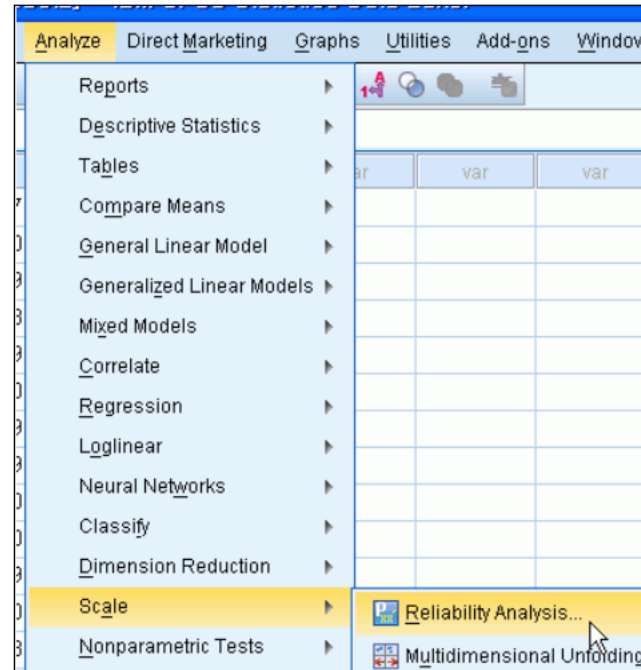
	reader1	reader2	reader3
1	7.8	7.9	7.7
2	10.2	10.7	9.8
3	9.8	9.7	8.9
4	8.8	8.9	8.0
5	10.0	9.9	9.0
6	10.0	10.1	9.1

1. Data 입력

reader1 (연구자 No.1, cm)

reader2 (연구자 No.2, cm)

reader3 (연구자 No.3, cm)

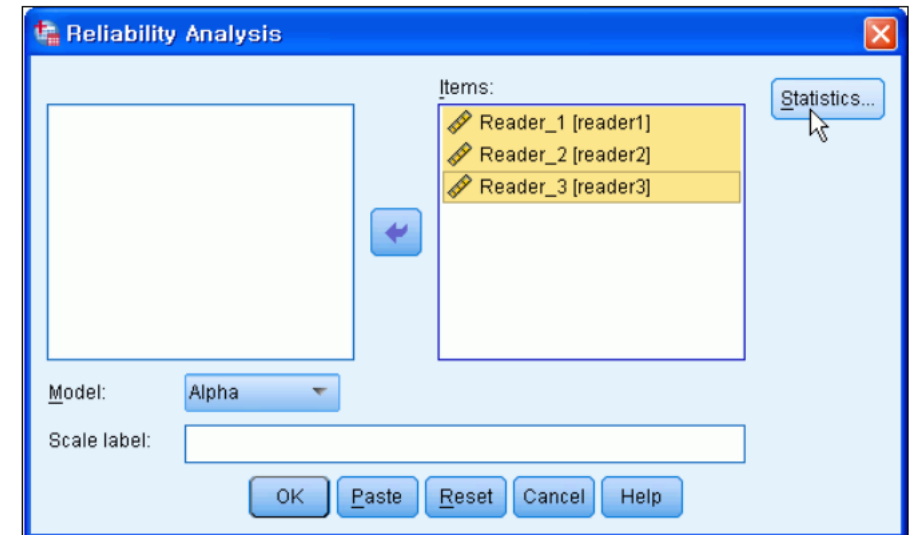


2. 선택

[MENU]
-[Analyze]
-[Scale]
-[Reliability Analysis]

3. 설정

[Items]에 비교하고자 하는 변수를 넣고, [Statistics]를 누른다.



3) 신뢰도분석 (intra-class correlation coefficient, ICC)

■ 관찰자간, 관찰자내의 신뢰도 검정 (ex., tumor size)

4. [Statistics]

Descriptives for에서 [Item],
[Intraclass correlation coefficient],
Model에서 [Two-Way Random],
Type에서 [Consistency]를 선택한다.
[Continue]-[OK]를 누른다.

Case Processing Summary			
		N	%
Cases	Valid	50	100.0
	Excluded ^a	0	.0
	Total	50	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics	
Cronbach's Alpha	N of Items
.959	3

Item Statistics			
	Mean	Std. Deviation	N
Reader_1	9.466	.8884	50
Reader_2	9.900	.8564	50
Reader_3	9.000	.8507	50

3명의 연구자가 측정한 평균값과 표준편차를 보여준다.

ICC(Intraclass Correlation Coefficient)	신뢰도
80-100	매우 신뢰
60-79	중등도 신뢰
59 이하	신뢰할 수 없음

Intraclass Correlation Coefficient							
	Intraclass Correlation ^a	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig.
Single Measures	.887 ^b	.828	.930	24.626	49	98	.000
Average Measures	.959	.935	.976	24.626	49	98	.000

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
b. The estimator is the same, whether the interaction effect is present or not.

[Intraclass Correlation Coefficient]표에서

Average Measures의 ICC(Intraclass Correlation Coefficient) = 0.959,
즉, 3명의 연구자간에 시행한 측정이 95.9% 일치한다는 의미이고, 이는 통계적으로 유의하다 (p=0.000).

4) 평균비교

1개군과 알려진 특정값의 비교

A군 vs. 특정값

A검사 결과 나타난 평균치가 기존에 정상이라고 알려진 값과 유의한 차이가 있는가?

독립된 2개군의 비교

A군 vs. B군

A군과 B군사이에 AST의 차이가 있는가?

... 상응하는 비모수적 방법

짝을 지은 2개군의 비교

치료전 vs. 치료후

치료전과 후의 안압에 차이가 있는가?

... 상응하는 비모수적 방법

독립된 3개 이상 군의 비교

A군 vs. B군 vs. C군 ...

A, B, C 군의 혈압에 유의한 차이가 있는가?

... 상응하는 비모수적 방법

A군			B군			C군		
0hr	1hr	3hr	0hr	1hr	3hr	0hr	1hr	3hr
..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..

반복되는 (짝을 이루는) 측정값을 가진 군간의 유의성을 알아보고자 하는 경우 (paired t-test의 확장개념)

3번씩 반복측정을 한 A, B, C 군이 있다. 군간에는 유의한 차이가 있는가?

One-sample t-test

t-test (student t-test, independent-sample t-test)

Mann-Whitney U-test

Paired t-test

Wilcoxon signed rank test

One-way ANOVA

Kruskal Wallis test

Repeated measures ANOVA

일원배치 분산분석 (one-way analysis of variance (ANOVA))

■ 사후분석 (post-hoc analysis) 수행: 어느 그룹간 차이가 발생했는지 확인함



일원 분산 분석: 사후분석 - 다중 비교

등분산을 가정함

- ☐ LSD
- ☐ Bonferroni
- ☐ Sidak
- ☐ Scheffe
- ☐ R-E-G-W의 F
- ☐ R-E-G-W의 Q
- ☐ S-N-K
- ☒ Tukey 방법
- ☐ Tukey의 b
- ☐ Duncan
- ☐ Hochberg의 GT2
- ☐ Gabriel
- ☐ Waller-Duncan
- 제1종/제2종 오류 비율: 100
- ☐ Dunnett
- 제어 범주(Y): 마지막
- 검정
 - ☒ 양쪽검정(2)
 - ☐ < 제어(O)
 - ☐ > 제어(N)

등분산을 가정하지 않음

- ☐ Tamhane의 T2
- ☐ Dunnett의 T3
- ☐ Games-Howell
- ☐ Dunnett의 C

유의 수준(E): 0.05

계속 취소 도움말

- Tukey: 집단간 차이를 가장 정밀하게 감지, 각 그룹의 샘플수가 같을 때 사용
- Bonferroni: 각 그룹의 샘플수가 다를 때 사용
- Duncan: 사회과학 분야에서 주로 사용

Kruskal-Wallis H test

- One-way ANOVA의 비모수검정, 세 집단 이상의 집단 분포를 비교
- 가정사항
없음
- 사후검정 (post-hoc analysis)
 - Bonferroni Correction method 이용
 - : Kruskal-Wallis H test에서 집단간 유의성이 있을 때,
각 집단을 짝으로 묶어서 t-test의 비모수적인 방법인 Mann-Whitney U검정 실시
(단, 유의수준은 Mann-Whitney U검정 수로 나누어 줌)
Ex) 3개의 집단의 경우 사후검정은 세번 실시해야 하므로 (a-b, b-c, a-c) 이때,
유의수준은 $0.05/3=0.0167$ 을 사용)

5) 상관분석

연속변수로 구성된 두 변수간의 선형적인 관련성이 있는가에 대한 분석
독립변수와 종속변수의 개념이 없다 (=원인과 결과에 대한 개념이 없는 data분석)

ex) 키와 발의 크기에 대한 상관성분석

if) 결과값 = 키가 크면 발이 큰 것인지, 발이 크면 키가 큰 것인지 알 수 없음
: 원인과 결과에 대한 설명이 확실한 경우 = 회귀분석 등을 이용

2개군의 선형 관련성 분석

A군과 B군간에는 어느 정도의 상관성을 가지고 있는가? (A군이 증가하면 B군도 증가하는지? A군이 증가하면 B군은 감소하는지? 등)

Pearson correlation

... 상응하는 비모수적 방법

Spearman correlation

5) 상관분석

Correlations							
		Systole BP	Diastole BP	Body Temperature	Pulse Rate	Respiration Rate	O2 Saturation
Systole BP	Pearson Correlation	1	.973**	.216	.215	.130	.194
	Sig. (2-tailed)		.000	.113	.116	.346	.156
	N	55	55	55	55	55	55
Diastole BP	Pearson Correlation	.973**	1	.240	.174	.113	.167
	Sig. (2-tailed)	.000		.077	.203	.412	.222
	N	55	55	55	55	55	55
Body Temperature	Pearson Correlation	.216	.240	1	.549**	-.115	-.005
	Sig. (2-tailed)	.113	.077		.000	.402	.971
	N	55	55	55	55	55	55
Pulse Rate	Pearson Correlation	.215	.174	.549**	1	-.114	-.011
	Sig. (2-tailed)	.116	.203	.000		.407	.935
	N	55	55	55	55	55	55
Respiration Rate	Pearson Correlation	.130	.113	-.115	-.114	1	.826**
	Sig. (2-tailed)	.346	.412	.402	.407		.000

Systole BP vs. Diastole BP	r = 0.973	p = 0.000	2군간에 강한 양(+)적 선형 상관관계가 있으며, 통계적으로 유의하다.
Body Temperature vs. Pulse Rate	r = 0.549	p = 0.000	2군간에 중등 양(+)적 선형 상관관계가 있으며, 통계적으로 유의하다.
Respiration Rate vs. O ₂ Saturation	r = 0.826	p = 0.000	2군간에 강한 양(+)적 선형 상관관계가 있으며, 통계적으로 유의하다.

5) 상관분석 - 편상관분석 (partial correlation)

Ex) 병상수와 병원내 사망자 간의 Pearson correlation coefficient (r)=0.202, p =0.047 이다.

→ 큰 병원일수록 위험하다.

BUT) 병상수가 많은 병원은 사망자수는 물론 완치 환자수 또한 많다.

그러므로 완치환자수에 대한 요인(제3의 교란요인)을 배제할 경우에도 과연 서로 상관관계가 있을까?

→ 병상수와 사망자수의 상관계수 r =0.103 으로 약한 양(+)적 선형관계를 가지고 있으나, p =0.319로 통계적으로 유의하지 못하다. 완치환자수를 고려할 경우 병상수와 사망자수의 상관성은 없다 (p =0.319).

Correlations

		병상수	사망자수
병상수	Pearson Correlation	1	.202*
	Sig. (2-tailed)		.047
	N	97	97
사망자수	Pearson Correlation	.202*	1
	Sig. (2-tailed)	.047	
	N	97	97

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

Control Variables			병상수	사망자수
완치환자수	병상수	Correlation	1.000	.103
		Significance (2-tailed)	.	.319
		df	0	94
	사망자수	Correlation	.103	1.000
		Significance (2-tailed)	.319	.
		df	94	0

6) 회귀분석 - 단순회귀분석 (simple regression analysis)

독립변수 1 개 (연속변수)

종속변수 1 개 (연속변수)

연속변수로 구성된 두 변수가 직선적 관련성을 가지고 있는지 그리고 그를 반영하는 수학 방정식($Y=B_0+B_1X$)을 구하고자 하는 경우

두 변수의 관계는 직선적 관련이 있어야 함.

Tip) 초기 단계에 두 변수에 대한 그래프를 작성하여 직선적 성향을 가졌는지를 관찰하고 만일 포물선 또는 log함수 등의 관계가 있을 경우 적절한 변형을 통하여 직선형으로 자료를 변형한 후에 이 통계기법을 적용하면 좋음

6) 회귀분석 - 단순회귀분석 (simple regression analysis)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.363 ^a	.132	.114	.8059

a. Predictors: (Constant), Alkaline phosphatase

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.843	1	4.843	7.457	.009 ^a
	Residual	31.822	49	.649		
	Total	36.665	50			

a. Predictors: (Constant), Alkaline phosphatase

b. Dependent Variable: Total bilirubin

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.565	.434		1.303	.199
	Alkaline phosphatase	.009	.003	.363	2.731	.009

a. Dependent Variable: Total bilirubin

R Square (결정계수, r^2) = 0.132

→ 13.2%의 설명력을 가지고 있다.

($p=0.009$ 이므로 통계적으로 의미가 있다.)

$Y=B_0+B_1X$ 에서, $B_0 = 0.565$, $B_1 = 0.009$
즉, $TB = 0.565 + 0.009 * ALP$

r^2 가 1 (=100%)에 가까울수록
설명력이 좋다.

'설명력이 좋다': 통계적 유의성을
의미하는 것이 아니고, '선형의
상관성이 좋다'라는 뜻

6) 회귀분석 - 다중회귀분석 (multiple regression analysis)

독립변수 n 개 (연속변수)

종속변수 1 개 (연속변수)

종속변수를 설명하는 독립변수가 2개 이상인 경우 이들의 관련성을 알고 그를 반영하는 수학적 방정식($Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 \dots + B_nX_n$)을 구하고자 하는 경우 (Simple regression analysis의 확장 개념).

6) 회귀분석 - 다중회귀분석 (multiple regression analysis)

Ex) Age[X_1], Stent Diameter(mm)[X_2], Stent Cross Sectional Area (mm²)[X_3], ALP(AI Phosphatase)[X_4], AST(sGOT)[X_5], ALT(sGPT)[X_6], Total Bilirubin[X_7]이 Stent Open Duration(days)[Y]에 영향을 주는가?

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	183.427	25.050		7.322	.000		
	Age	-.206	.063	-.229	-3.272	.002	.755	1.325
	Stent Diameter (mm)	-.571	.642	-.882	-.890	.377	.004	265.803
	Stent Cross Sectional Area (mm)	.004	.005	.798	.805	.424	.004	265.641
	ALP (AI Phosphatase)	-.249	.026	-.760	-9.761	.000	.610	1.640
	AST (sGOT)	-.009	.054	-.012	-.159	.874	.642	1.558
	ALT (sGPT)	-.322	.154	-.153	-2.093	.041	.691	1.447
	Total Bilirubin	-1.474	1.121	-.085	-1.315	.194	.893	1.119

a. Dependent Variable: Stent Open Duration (days)

A

다중공선성 확인 먼저!
VIF >10

B

C

6) 회귀분석 - 다중회귀분석

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	Total Bilirubin, AST (sGOT), Stent Diameter (mm), Age, ALT (sGPT), ALP (AI Phosphatase)	.	Enter

a. All requested variables entered.

b. Dependent Variable: Stent Open Duration (days)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.893 ^a	.798	.776	2.556

a. Predictors: (Constant), Total Bilirubin, AST (sGOT), Stent Diameter (mm), Age, ALT (sGPT), ALP (AI Phosphatase)

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1419.273	6	236.546	36.196	.000^a
	Residual	359.436	55	6.535		
	Total	1778.710	61			

a. Predictors: (Constant), Total Bilirubin, AST (sGOT), Stent Diameter (mm), Age, ALT (sGPT), ALP (AI Phosphatase)

b. Dependent Variable: Stent Open Duration (days)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	164.475	8.535		19.271	.000		
	Age	-.205	.063	-.228	-3.262	.002	.755	1.324
	Stent Diameter (mm)	-.056	.042	-.086	-1.339	.186	.892	1.121
	ALP (AI Phosphatase)	-.243	.024	-.741	-10.011	.000	.670	1.493
	AST (sGOT)	-.023	.051	-.032	-.454	.652	.722	1.386
	ALT (sGPT)	-.341	.152	-.162	-2.245	.029	.707	1.415
	Total Bilirubin	-1.491	1.117	-.086	-1.334	.188	.894	1.119

a. Dependent Variable: Stent Open Duration (days)

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	Age	Stent Diameter (mm)	ALP (AI Phosphatase)	AST (sGOT)	ALT (sGPT)	Total Bilirubin
1	1	6.935	1.000	.00	.00	.00	.00	.00	.00	.00
	2	.023	17.473	.00	.00	.01	.06	.27	.00	.33
	3	.016	20.643	.00	.18	.01	.00	.24	.00	.26
	4	.012	24.001	.00	.09	.57	.03	.00	.00	.16
	4	.009	27.737	.01	.05	.01	.84	.27	.00	.02
	6	.004	43.345	.04	.18	.31	.00	.21	.40	.23
	7	.001	82.621	.95	.49	.09	.07	.00	.59	.01

a. Dependent Variable: Stent Open Duration (days)

6개의 변수로 Stent Open Duration을 79.8% 설명할 수 있고, 통계적으로 유의하다 ($r^2=0.798$, $p=0.000$).

7) Pearson's Chi-square test

환자군과 대조군간의 유의한 차이가 있는가?

2 x 2 표	환자군	대조군
증상(+)	...	...
증상(-)	...	...

교차비(Odd ratio), 상대위험도(Relative risk)를 얻을 수 있다.

7) Pearson's Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7.924 ^a	1	.005	.006	.004
Continuity Correction ^b	6.869	1	.009		
Likelihood Ratio	8.317	1	.004		
Fisher's Exact Test					
Linear-by-Linear Association	7.862	1	.005		
N of Valid Cases	128				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.25.

b. Computed only for a 2x2 table

Disease group과 Control group 사이에는 유의한 차이가 있다. ($p=0.005$)

7) Pearson's Chi-square test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7.924 ^a	1	.005		
Continuity Correction ^b	6.869	1	.009		
Likelihood Ratio	8.317	1	.004		
Fisher's Exact Test				.006	.004
Linear-by-Linear Association	7.862	1	.005		
N of Valid Cases	128				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.25.

b. Computed only for a 2x2 table

Pearson Chi-Square 의 p값을 취하는 경우 vs. Fisher's Exact Test의 p값을 취하는 경우

→ 2 x 2 table의 하나의 셀이라도 5보다 작은 경우 (25% 기준!!!)

7) Pearson's Chi-square test

Odd ratio, Relative risk의 경우,

: 95% 신뢰구간에 "1"이 포함되지 않는 경우
통계적으로 유의하다고 판단

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Findings (Positive / Negative)	3.291	1.404	7.714
For cohort Group = Disease group	1.516	1.164	1.973
For cohort Group = Control group	.460	.249	.850
N of Valid Cases	128		

Odds ratio = 3.291 (95% 신뢰구간 = 1.404 – 7.714)
으로 통계적으로 유의하다.

Odds ratio란? Disease group이 Control group에
비해서 Finding Positive가 될 가능성이 3.291 배
높고, 이는 통계적으로 유의하다.

Relative risk = 1.516 (95% 신뢰구간 = 1.164 –
1.973) 으로 통계적으로 유의하다.

Relative risk란? Findings Positive이면 Negative인
경우에 비해서 Disease group이 될
가능성이 1.516 배 높고, 이는 통계적으로 유의하다.

8) 생존분석 (survival analysis)

종속변수 즉 status를 나타내는 변수에는 두 가지 항목이 존재함 (하나는 원하는 event (death)이고, 다른 하나는 no event (censored)를 의미)

Ex) 연구기간이 5년이었는데 그 기간에 해당 질병으로 사망한 경우는 event가 발생한 것이고, 다른 원인에 의한 사망이나, 생존하는 경우, 또는 사망 여부를 알 수 없는 경우를 censored로 처리

→ 반드시 죽고 살고 로 나뉘는 것이 아니고 event vs. censored (해당질병에 의한 사망 vs. 기타 불확실한 모든 경우)로 이분된 값을 가짐.

예1) Stomach ca. 수술 후 5 year survival rate는 얼마인가?

예2) Stent의 수명은 어느 정도이고 12 month patency rate는 얼마일까?

8) 생존분석 – Kaplan–Meier method

1. 상태를 나타내는 이분된 비연속변수 (event vs. censored)
2. 기간을 나타내는 변수 (survival time)
3. 기타 (group, stage.. 등 비연속변수)

각군의 자료가 50개 이하인 경우

Event 발생 때 마다 자동으로 생존율이 계산되며,
그래프에 censored data가 나타남.

Lung ca.로 항암치료를 받은 환자들의
mean, median survival time은 얼마인가?

8) 생존분석 - Kaplan-Meier method

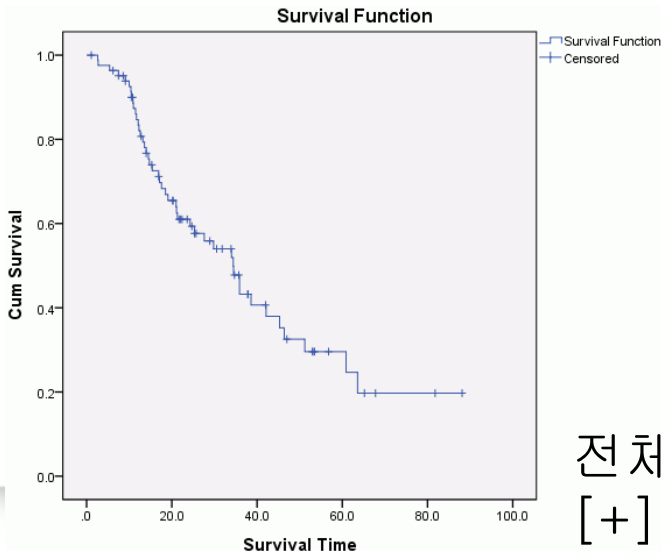
Case Processing Summary

Total N	N of Events	Censored	
		N	Percent
83	45	38	45.8%

Means and Medians for Survival Time

Mean ^a				Median			
Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound			Lower Bound	Upper Bound
40.478	3.901	32.833	48.124	34.400	4.103	26.358	42.442

a. Estimation is limited to the largest survival time if it is censored.



Survival Table

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	2.600	Event	.988	.012	1	81
2	2.700	Event	.976	.017	2	80
3	5.400	Event	.963	.021	3	79
4	6.200	Censored	.	.	3	78
5	7.500	Event	.951	.024	4	77
6	7.500	Censored	.	.	4	76
7	8.600	Censored	.	.	4	75
8	9.100	Event	.938	.027	5	74

Cummulative Proportion Surviving at the Time에 해당하는 숫자들이 해당 기간의 생존율이다.

2.6 개월에는 0.988 (98.8%),
2.7 개월에는 0.976 (97.6%),
5.4 개월에는 0.963 (96.3%),
7.5 개월에는 0.951 (95.1%) ...

전체 환자의 생존율을 그래프로 보여 주며,
[+] 표시는 censored data의 발생을 의미한다.

8) 생존분석 – Cox proportional hazard model

1. 상태를 나타내는 이분된 비연속변수 (event vs. censored)
2. 기간을 나타내는 변수 (survival time)
3. 기타 (group, stage, age.. 등 비연속변수 or 연속변수)

생존율에 영향을 주는 제3의 교란요인(Confounding factor)을 반영하고 이들의 영향력을 검정한다.

→ 제3의 교란요인들 각각이 연구기간 중 일정한 영향력을 가진다는 전제조건이 성립되어야 한다.

Lung ca.의 생존율에 영향을 주는 변수는 어떤 것이며, 각각의 odds ratio는 얼마인가? 즉 사망율이 몇 배나 증가하는가?

age (Age)

sex (Sex, Male = 1, Female = 2)

stage (Lung ca. Stage, Stage I=1, Stage II=2, Stage III=3, Stage IV=4)

trial (Clinical Trial Type, Control = 0, A Type Trial = 1, B Type Trial = 2)

smoking (Smoking Hx., Negative = 0, Positive = 1)

time (Survival Time, 환자가 살아있는 기간, month)

status (Survival Status, Event = 0, Censored = 1)

Lung ca.의 생존율에 영향을 주는 변수는
어떤 것이며, 각각의 odds ratio는 얼마인가?
즉 사망율이 몇배나 증가하는가?

8) 생존분석

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)
Step 1	trial			14.809	2	.001	
	trial(1)	-1.193	.579	4.249	1	.039	.303
	trial(2)	-2.344	.621	14.266	1	.000	.096
Step 2	sex	-1.454	.558	6.777	1	.009	.234
	trial			20.173	2	.000	
	trial(1)	-1.913	.637	9.013	1	.003	.148
Step 3	trial(2)	-2.937	.675	18.922	1	.000	.053
	sex	-1.156	.566	4.161	1	.041	.315
	trial			18.123	2	.000	
	trial(1)	-1.553	.643	5.835	1	.016	.212
	trial(2)	-2.861	.683	17.531	1	.000	.057
	smoking	1.166	.590	3.906	1	.048	3.210

sex	B = -1.156	(-)값을 가지므로 reference 즉 [Male]에 비해 사망율이 낮아진다는 의미
	p = 0.041	0.05보다 작으므로 사망율에 유의한 영향을 준다는 의미
	Exp(B) = 0.315	[Female]은 reference 즉 [Male]에 비해 사망률이 0.315 배로 감소한다는 의미
Trial	B =	Trial의 경우 3개이상의 분류로 되어 있으므로 다음줄에 있는 각각의 항목으로 표시된다.
	p = 0.000	0.05보다 작으므로 사망율에 유의한 영향을 준다는 의미 Trial에는 3개 분류가 있는데 이들 모두 통틀어서 유의한 영향을 준다는 의미
	Exp(B) =	Trial의 경우 3개이상의 분류로 되어 있으므로 다음줄에 있는 각각의 항목으로 표시된다.
Trial(1)	B = -1.553	(-)값을 가지므로 reference 즉 [Control]에 비해 사망율이 낮아진다는 의미
	p = 0.016	0.05보다 작으므로 사망율에 유의한 영향을 준다는 의미
	Exp(B) = 0.212	Trial(1) 즉 [A Type Trial]은 reference 즉 [Control]에 비해 사망률이 0.212배로 감소한다는 의미
Trial(2)	B = -2.861	(-)값을 가지므로 reference 즉 [Control]에 비해 사망율이 낮아진다는 의미
	p = 0.000	0.05보다 작으므로 사망율에 유의한 영향을 준다는 의미
	Exp(B) = 0.057	Trial(2) 즉 [A Type Trial]는 reference 즉 [Control]에 비해 사망률이 0.057배로 감소한다는 의미
smoking	B = 1.166	(+)값을 가지므로 reference 즉 [Negative]에 비해 사망율이 높아진다는 의미
	p = 0.048	0.05보다 작으므로 사망율에 유의한 영향을 준다는 의미
	Exp(B) = 3.210	[Positive]는 reference 즉 [Negative]에 비해 사망률이 3.210배로 증가한다는 의미

Exp(B) : Odds ratio [교차비, 위험도가 몇배 증가하는 것인가?]

Exp(B) < 1 : 위험도가 감소한다는 의미 (이때 B값은 (-)이다)

Exp(B) = 1 : 변화없다는 의미

Exp(B) > 1 : 위험도가 증가한다는 의미 (이때 B값은 (+)이다)

8) 생존분석 – Cox proportional hazard model

유의사항

Censored data (불확실한 자료)로 처리하여야 하는 경우

1. **loss to follow up**: 환자의 추적이 불가능한 경우 (이사)
 2. **drop out**: 환자의 치료 거부 또는 중단 (민간요법으로 전환)
 3. **termination of study**: 환자가 사망하기 전에 연구가 종료되는 경우
 4. **death from unrelated cause**: 관련이 없는 이유로 사망한 경우 (교통사고로 사망)
- 예를 들면 연구에 대상이 되는 질병으로 사망한 경우만을 event로 처리한다.
조사하고자 하는 것과 무관하게 Event가 일어난 경우는 모두 Censored data로 처리하여야 함.



참고문헌

<https://blog.naver.com/y4769>

<http://www.cbgstat.com/>



BRAIN PHD

감사합니다!

T H A N K Y O U

—
Q&A