

편집자 관점에서의 통계적 리뷰

2022. 8. 20.

연세대학교 원주의과대학

정밀의학과 · 의학통계학과

강대용

Assessment on the adequacy of the statistical methods

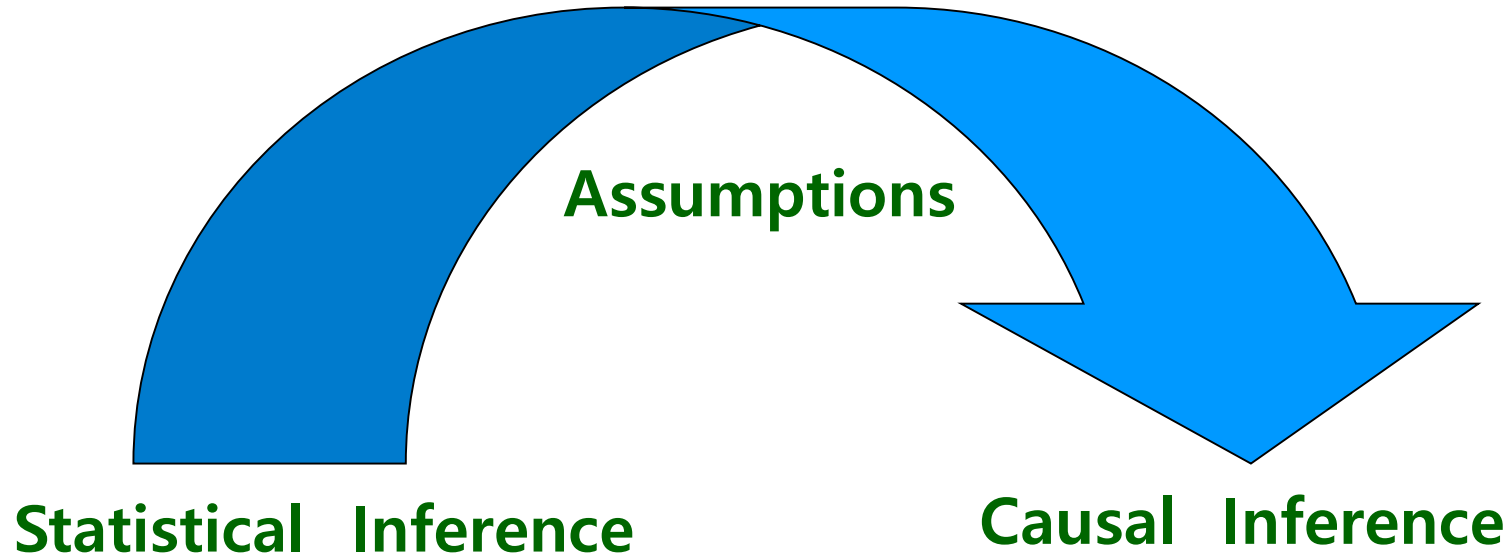
평가항목	항 목 설 명	대한의학학술지 편집인협의회 (KoreaMed, 2004. 10)
1	적절한 통계방법을 사용하였음*	
2	통계가 필요하나 사용하지 않았음	
3	p-value를 제시하고 있으나 통계방법 언급 없음	
4	통계방법을 나열하고 있으나 어느 분석에 적용한 것인지 알 수 없음	
5	틀린 방법 적용 1) 자료의 성격과 전혀 맞지 않는 방법 사용 2) ANOVA를 t-test로 분석 3) 비모수를 모수로 분석 4) 빈도 비교에서 기대치가 5 이하인 칸이 있음에도 Fisher's exact test를 쓰지 않았거나, 범주를 줄여 그 칸을 없애려는 시도를 하지 않음 5) 상관관계를 회귀분석으로 분석 6) 통계용어를 잘못 사용 7) 기타 경우 †	
6	적용통계는 맞게 언급했으나 결과를 제대로 제시하지 못함 1) 어느 군과 어느 군의 비교 시에 나온 p-value인지 불명확 함 2) p-value 또는 confidence interval을 제시하지 않고 유의하다고 주장 3) 양측 / 단측 언급 없음, 유의수준 언급 없음 4) 통계량이나 자유도를 제시하지 않음 5) 계산이 틀렸을 가능성이 농후한 경우 (유의하지 않을 것 같은데 유의 하다고 주장) 6) 결론을 잘못 유도 (통계결과를 확장하여 결론을 내림) 7) 기타 경우 ‡	

* 아래 2~6 항목에 해당사항 없을 때

† 교란변수의 존재로 logistic analysis가 필요한 경우에 χ^2 -pearson 으로만 분석

‡ 4 X 2 table에 Fisher's exact test를 사용했다고 쓰고 p-value 제시 없이 결론지음

What do you want to know?



통계학이란 ? “데이터의 지식화 (learning from data)”

역사적으로 인간은 수량적으로 표현된 경험적사실들인 데이터로부터 일반적인 추론을 해왔다.

데이터의 지식화는 다음 3가지 요소로 구성된다.

1. 타당하고 신뢰성 있는 자료를 만드는 일
2. 자료를 분석해내는 능력
3. 분석결과의 전략적 활용

(통계적 사고 , 허명회 2006)

What makes it difficult for Medical Research?

✓ **the research target is
'Human'**

- ethical problems
- limit of study design
- problems caused from the limit of study design

✓ **distortion of research results
occurs when we have no
enough time**

- need for comparing several analytic results
- lack of reflections in the discussion part

✓ **data noise
data incomplete**

- outliers
- missing value
- there is no data without 'noise'

✓ **when we use inappropriate statistical methods in data analysis**



EXCEL



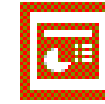
MSACCESS



SAS



IBM SPSS



POWERPNT



Hwp

1. Descriptive Statistics :



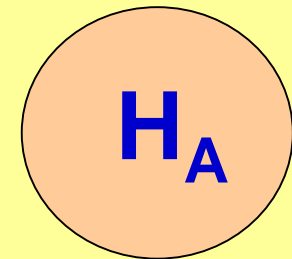
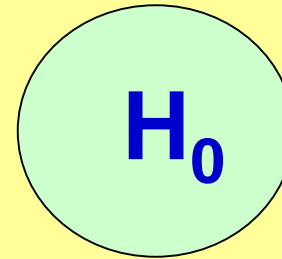
we work on 'data cleaning' while calculating DS of each observed variable
e.g., n, missing value, checking outlier, category regrouping, ...

2. Statistical Testing (検定) :



Statistical inference
(1:1, 1:k, sub-group)

under significant level $\alpha=5\%$



3. Statistical Interpretation :



(highly) significant, limit significant, borderline significant, not significant
(difference, association, correlation, influence with adjustment, ...)

decision making with 'p-value'

4. Interpretation through 'Medicine' and 'Public Health'

Study Designs

Observational study

Unit of study

Descriptive study

Analytical study

Hypothesis ?

Ecological study (Correlation study)

Population

Cross-sectional study (Prevalence study)

Individuals

Case-control study (Case-reference study)

Individuals

Cohort study (Follow-up study, Prospective study)

Individuals

Hybrid designs :

Nested case-control design, Case-cohort design

Case-crossover design, Case-time-control design

Experimental study, Quasi-experimental study

Experiment

Randomized controlled trials (Clinical study)

Patients

Field trials

Healthy people

Community trials (Community intervention study)

Community

Randomization ?
Intervention ?

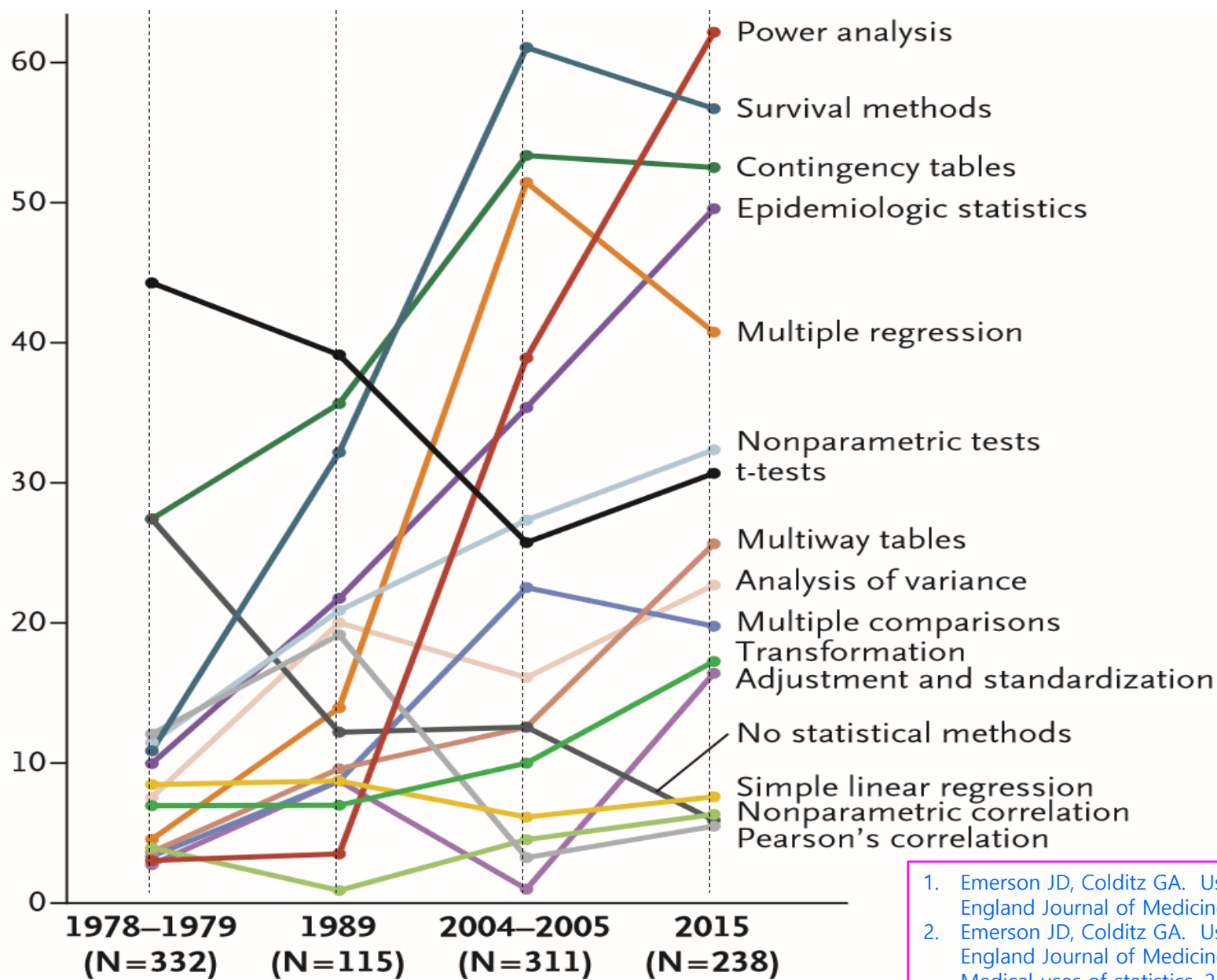
Categories of statistical procedures used to assess the statistical content in the articles

자료 성격	권고 통계분석 방법
사례보고, 임상연구, 치료결과분석 등	No statistical method or Descriptive study
진단능력평가, 참고치 정하기	Sensitivity, Specificity, ROC curve
짝을 이룬 두 그룹간 평균비교	Paired t-test, Wilcoxon signed rank test *
독립적인 두 그룹간 평균비교	t-test, Wilcoxon rank sum test *, Mann-Whitney U test *
독립적인 세 그룹 이상 평균비교 (또는 군간비교)	ANOVA (with multiple comparison), Kruskal-Wallis test *
동일인에 대한 3회 이상 반복측정자료의 평균비교	Repeated measures of ANOVA, Friedman test *
두 그룹 또는 세 그룹 이상 빈도 비교	Chi-squared test *, Fisher's exact test *
동일인에 대한 반복측정 빈도 비교	McNemar's test *
두 연속변수간 상관관계 분석	Pearson's correlation, Spearman's rho *
두 개 이상 독립변수와 종속변수와의 관계 분석	Simple linear regression, Multiple (logistic) regression
생존율 추정, 생존율 비교 생존형 자료의 회귀분석	Life table, Kaplan-Meier method Log-rank test, Cox's proportional hazard model (HR)
역학적 통계량 분석	Incidence, Prevalence, Risk ratio (RR), Odds ratio (OR)

* 비모수적 방법

Source: Emerson JD, Colditz GA. Use of Statistical Analysis in The New England Journal of Medicine. *N Engl J Med* 1983; 309: 709-713.

Percent of Research Articles Using a Particular Analysis



- Emerson JD, Colditz GA. Use of statistical analysis in The New England Journal of Medicine. *NEJM* **1983**; 309: 709-13.
- Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. In: Bailar JC III, Mosteller F, eds. *Medical uses of statistics*. 2nd ed. Waltham, MA: NEJM Books, **1992**: 45-57.
- Horton NJ, Switzer SS. Statistical methods in the Journal. *NEJM* **2005**; 353: 1977-9.
- Sato Y, Gosho M. Statistical Methods in the Journal - An Update. *NEJM* **2017**; 376: 1086-7.

측정 수준에 따른 확률변수의 구분 :

① 범주형 변수 (categorical variable)

- **명목척도(nominal scale) :** 단지 범주로만 의미가 있다.
ex) 성별, 혈액형, 치료방법 등
- **순위척도(ordinal scale) :** 명목 + 대소 관계를 나타냄.
가감승제와 같은 수학적 계산은 무의미하다.
ex) 증상의 정도, 학력수준, 사회경제적 수준 등

② 연속형 변수 (continuous variable)

- **구간척도(interval scale) :** 측정치 간의 '간격'에 의미가 있는 경우
ex) IQ, 온도의 경우 $10^{\circ}\text{C} \sim 15^{\circ}\text{C}$, $20^{\circ}\text{C} \sim 25^{\circ}\text{C}$ 의 5°C 는 본질적으로 같다.
가감은 가능하나, 승제는 불가능함 (→ 比(ratio)의 개념은 갖지 못함)
ex) $100^{\circ}\text{C} / 50^{\circ}\text{C} \neq 212^{\circ}\text{F} / 122^{\circ}\text{F}$ (∵ 절대 0°C , 0°F 이 아니라 인위적으로 정한 것임)
- **비율척도(ratio scale) :** 'age'
절대 영점을 가지게 되므로 수학적으로 가장 완벽한 형태의 변수 (가감승제 모두 가능함)
ex) 80세는 20세에 비해 60살 더 많고(구간), 4배(비율) 더 살았다.

Data Entry 유의사항

■ 범주형 자료

- 적절한 형태의 숫자 코드 할당 (조사지 위에 함께 적어 놓는 것을 추천!!)
- 이진수 자료의 경우 **0/1**을 사용 추천 (주로, '예'=1 / '아니오'=0)

■ 연속형 자료

- 해당 자료를 측정한 **그대로** 기록 (크기를 줄여서 입력 지양)
- 측정 단위는 일관성을 유지

■ 한 사람 당 여러 개의 형식을 사용하는 경우

- **id** 부여 (자료의 결합을 위해) - 중요!!

■ 날짜와 시간의 문제 : 조사 / 입력 형식의 통일

■ 결측치(**missing value**)의 입력 : 가능한 한 default value (. 또는 공백) 사용

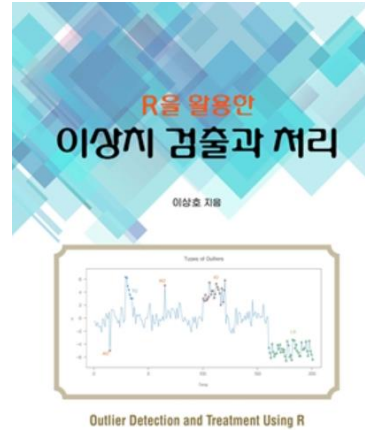
오류검토

- 원본 대조 / **double entry**
 - 둘 다 완벽한 것은 아님. 그러나 오류를 최소화 할 수 있다.
- 오류 검토
 - 범주형 자료
 - 검토가 비교적 쉬움
 - 나타나서는 안 되는 범주의 값이 입력되어 있다면 분명히 오타
 - 빈도표(frequency table)의 활용이 효과적
 - 연속형 자료
 - 오타발생 가능성 높음. 찾아내기는 어려움 (예: 소수점의 문제 등)
 - 범위(range) 검토 (Max – Min check)가 효과적
 - 날짜 자료
 - 쉬운 문제는 아님
 - 이상값의 확인 (예: 20070230)
 - 논리적인 검토 (예: 생년과 나이의 일치, 연구시작일 검토 등)
- 자료의 수정
 - 해당 자료가 잘 못 입력된 자료라는 **명백한 증거**가 있는 경우에 한함
 - 단지 이상한 값이 있다는 이유만으로 자료 수정은 곤란, 위험 함

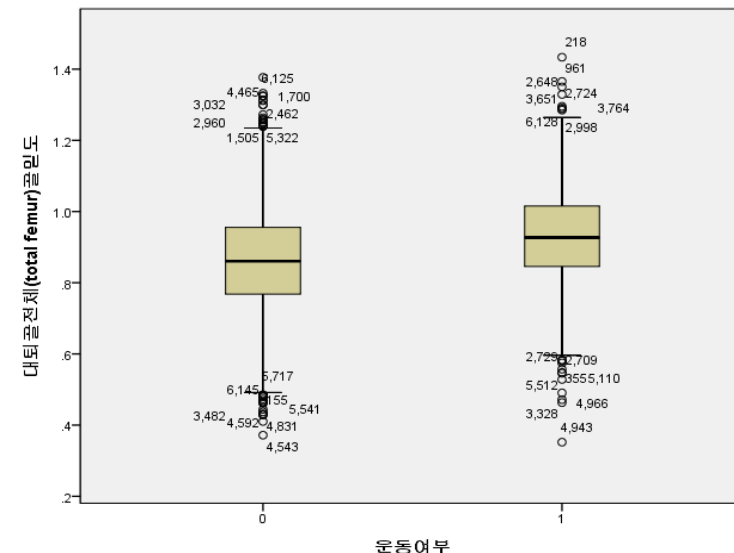
(extreme) outliers

(극)이상치

- 자료의 전반적인 값들과 구별되는 값 / 다른 값들과 병립될 수 없는 값
- 실제 관찰 값일 수도 있고 / 잘못 입력된 값일 수도 있음
 - 예, 키가 210인 여성
 - 만일 연구 대상이 일반인/초등학생이면? 만일 체중과 함께 검토한다면?
- 통계분석방법의 종류에 따라 결과에 심각한 영향을 미칠 수도
 - 예, Student's t-test vs. Wilcoxon's test
 - 따라서 자료 내에 이상치가 있는지 검토하는 것은 매우 중요
- **이상치에 대한 검토 !!!**
 - Range check이 효과적 / Graphical method (histogram, scatter plot 등) 사용도 효과적
 - 특정 통계모형 내에서도 가능 (예: 회귀분석 등)
- 이상치의 처리
 - 무분별한 삭제는 곤란
 - 최선의 방법 : **with / without analysis**
 - 분석결과가 서로 비슷하면 ok
 - 결과가 서로 상이하면 이상치에 영향을 많이 받지 않는 분석법 (예 : 자료의 변환, 비모수적 방법 등)



자유아카데미



정규성 검정 (normality test) :

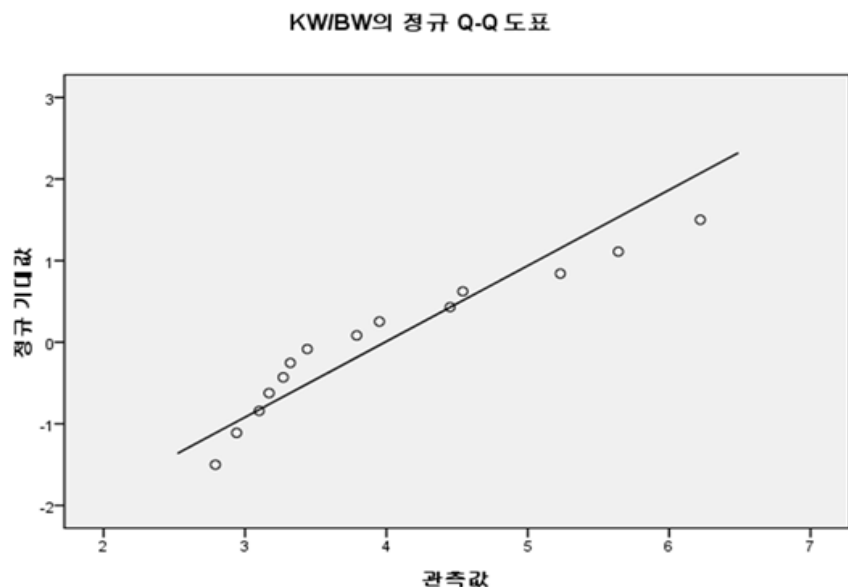
주어진 자료가 정규 모집단에서 랜덤하게 뽑힌 것인지 확인하고자 할 때 사용한다. 정규 모집단이라는 것이 확인되면 **모수적 검정(ex. Independent t-test)**를 하고, 아니면 **비모수적 검정(ex. Mann-Whitney U-test)**를 수행한다. 히스토그램을 통해 자료의 분포가 정규분포와 유사한지 확인해 볼 수 있고, 통계적 검정방법으로는 콜모고로프-스미르노프 검정과 샤피로-윌크 검정을 주로 사용한다.

- **콜모고로프-스미르노프 검정 (Kolmogorov-Smirnov test) :** 자료의 가장 작은 값부터 가장 큰 값까지의 누적상대빈도가 이론적 정규분포에서의 누적상대빈도와 얼마나 다른가를 측정하여 검정하는 방법이다.
- **샤피로-윌크 검정 (Shapiro-Wilk test) :** 자료 값들과 표준정규점수와의 선형상관 관계를 측정하여 검정하는 방법이다.

두 검정법의 가설은 "주어진 자료가 정규분포를 따른다"이므로 유의확률이 0.05 보다 클 때 정규성 가정을 만족한다고 할 수 있다. 표본크기(sample size)가 충분히 클 때, 콜모고로프-스미르노프 검정법을 사용하고, 작을 때는 샤피로-윌크 검정법을 사용한다. 이 외에 **크라머-본 미세스(Cramér-von Mises)**, **앤더스-달링 (Anderson-Darling) 검정** 등이 있다.

생체신장이식 과정 중 이식신장의 실제무게를 전향적으로 측정하여 공수여자간의 체격불일치 척도로 공여신장무게와 신장수여자체중 간의 비율(KW/BW) kidney weight(g) / recipient body weight(kg)을 측정하였다. 14명의 비율값이 정규분포를 따른다고 할 수 있는가?

KW/BW	3.79	2.79	2.94	3.95	4.45	4.54	5.64	6.22	5.23	3.10	3.17	3.44	3.32	3.27
-------	------	------	------	------	------	------	------	------	------	------	------	------	------	------



데이터 탐색

종속변수(D): KW/BW [kwbw]

요인(F):

통계량(S)...
도표(D)...
 옵션(O)...

표시
 모두(B) 통계량 도표

확인 불여넣기(B)

데이터 탐색: 도표

상자도표
 요인수준들과 함께(F)
 종속변수들과 함께(D)
 지정없음(N)

기술통계
 줄기와 잎그림(S)
 히스토그램(H)

검정과 함께 정규성도표(O)

Levene 검정이 있는 평균-산포
 없음
 제공값 추정(P)
 변환(T) 제공값: 자연로그
 변환하지 않음(U)

계속 취소 도움말

정규성 검정

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	통계량	자유도	유의확률	통계량	자유도	유의확률
KW/BW	.195	14	.155	.893	14	.090

비모수적 통계분석 (nonparametric analysis)

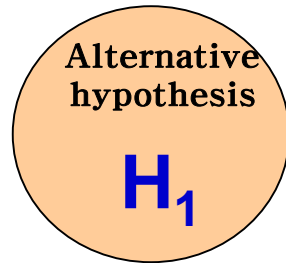
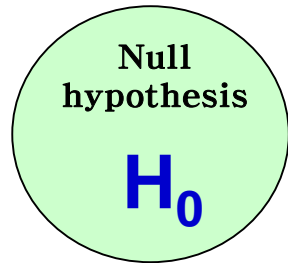
모수적(parametric) 통계분석은 기본적으로 모집단의 정규성, 등분산성, 측정치의 연속성을 가정하거나 조건으로 하는 방법이다. 하지만 실제 임상연구 자료에서는 정규분포의 가정을 만족하지 않아도 큰 문제가 되지 않는 경우가 많고, 표본수가 작아서 정규분포의 가정을 만족하지 않는 경우의 연구도 많다. 따라서 **비모수적 방법은 표본수가 크지 않고, 모집단의 분포가 정규분포를 따르지 않는 경우에 효율적인 통계분석 방법이 될 수 있다.**

→ 비모수적 통계분석의 기본은

- 모집단은 연속성이어야 하며, 그 분포에 대한 가정은 필요 없다.
- 측정치 보다는 그들의 상대 순위(rank) 혹은 서열(order)에 기반한다.
- 평균(mean)을 비교하는 것이 아니라 중위수(median)을 비교한다.
- 변이 정도는 표준편차(sd)가 아니라 범위(range)나 사분위수(IQR)로 표기한다.

비모수적 방법	모수적 방법	비모수적 방법
↑Δ τ (Π) / □ □ Δ Π	Paired t-test	Wilcoxon signed rank test
⊕ □ ξ Π) / □ □ Δ Π	Independent two-sample t-test	Wilcoxon rank sum test Mann-Whitney U test
⊕ □ ξ 8) / τ □ Δ Π	One-way ANOVA Two-way ANOVA	Kruskal-Wallis test Friedman test
Π □ □ □ □ □ □ □ □ □ □	Pearson's correlation	Spearman's rank correlation Kendall's tau





Errors in Hypothesis Testing



H_0 : No difference between effects of two drugs

H_1 : Not H_0

$H_0: \mu=120$ mmHg vs. $H_1: \mu \neq 120$ mmHg

Decision \ True		True	
		H_0 is True	H_0 is False
Decision	Fail to Reject H_0	 Ex. = Control	 β Type II error Ex. = Control
	Reject H_0	 α Type I error Ex. \neq Control	 Ex. \neq Control “Power (1-β)”

Type I error = $P(\text{positive} \mid H_0 \text{ true}) = \text{“False Positive”}$

Type II error = $P(\text{negative} \mid H_0 \text{ false}) = \text{“False Negative”}$

▪ **유의수준, level of significance $\alpha=5%$** , 이는 **우연**에 의해 연구결과가 참이 되어버리는 것(false positive)을 5%까지 허용하겠다는 연구자의 의지이다.

▪ **유의확률, p-value**, 내 데이터를 이용한 가설검정의 결과를 p-value로 요약됨. H_0 이 참인 연구결과가 **우연**에 의해 얻어지는 확률을 의미한다.

→ H_0 가 참인 확률(p-value)이 연구자가 미리 정해놓은 유의수준 $\alpha=5%$ 보다 작으면 H_0 이 거짓이다. H_0 를 기각한다 ('차이가 없다는 것이 아니다'). **반증적**으로 연구자가 원하는 H_1 을 증명한다. (all or none)

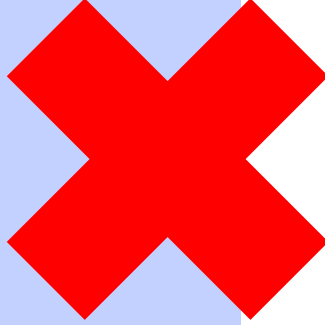
(Frequentist)귀무가설이 참/거짓 판단 vs. (Bayesian)연구가설이 참이 되는 확률 계산

→ **Bayesian** 입장에서는 *all uncertainty is measured by probability*이며 *continuous learning process* 여서 임상연구를 중간에 조기종료 / 수정이 가능할 수 있다. 이러한 Bayesian concept이 임상시험에 많이 도입되면서 adaptive design이 가능하게 되었다.

→ Baye's Theorem :

$$P(\theta | data) = \frac{P(data | \theta)}{P(data)} \times P(\theta)$$

p-value, 의사결정

$0.01 \leq p < 0.05$	→ significant	}	
$0.001 \leq p < 0.01$	→ highly significant		
$p < 0.001$	→ very highly significant		
$p > 0.05$	→ not statistically significant		
$0.05 \leq p < 0.10$	→ a trend toward statistical significance is sometimes noted		

- ❖ Analyses were performed with SAS software, version 9.1 (SAS Institute, Cary, NC). A 2-sided probability value 0.05 was considered statistically significant.
- ❖ ... whereas the difference in subjects with IFG/IGT was of borderline statistical significance (11.0 vs. 5.3%; $P = 0.06$).
- ❖ A probability of 0.05 was the limit of statistical significance.
- ❖ There was a borderline significant improvement in survival for the experimental arm with a median time to death of 63.2 months compared with 52.2 months in the standard cisplatin/paclitaxel arm (log-rank $P = .05$, one-tail)
- ❖ The multiplicative interaction term between vitamin D deficiency and hypertension had a borderline statistical significance ($P=0.08$ for both 2- and 3-category 25-OH D models).

The value of p -value in Biomedical Research

DB Panagiotakos, *The Open Cardiovascular Medicine Journal*, 2008, 2, 97-99.

- p -value 0.05가 '유의함(significance)'을 판단하는 절대적인 기준인가?
- p -value에 의존한 통계적 유의성 해석의 한계 有 (동일한 효과에서도 표본수에 따라 p -value가 달라진다)
 - 의사결정에서 p -value는 중요한 기준은 되지만 이것의 해석에는 한계가 있다.
- p -value는 관찰된 효과의 중요성 정도는 설명하지 못한다. 즉, p -value가 작다고 association이 강하다는 의미는 아니다.
- p -value의 수준은 sample size와 밀접한 관련이 있다.
 - p -value에 의한 해석에 의존한다면, 상대적으로 N수가 적은 '희귀질환'의 경우, 해당 의약품의 효과를 입증하기 위한 sample size에 도달하지 못하여 임상적 유용성 개선의 입증이 어렵게 된다.

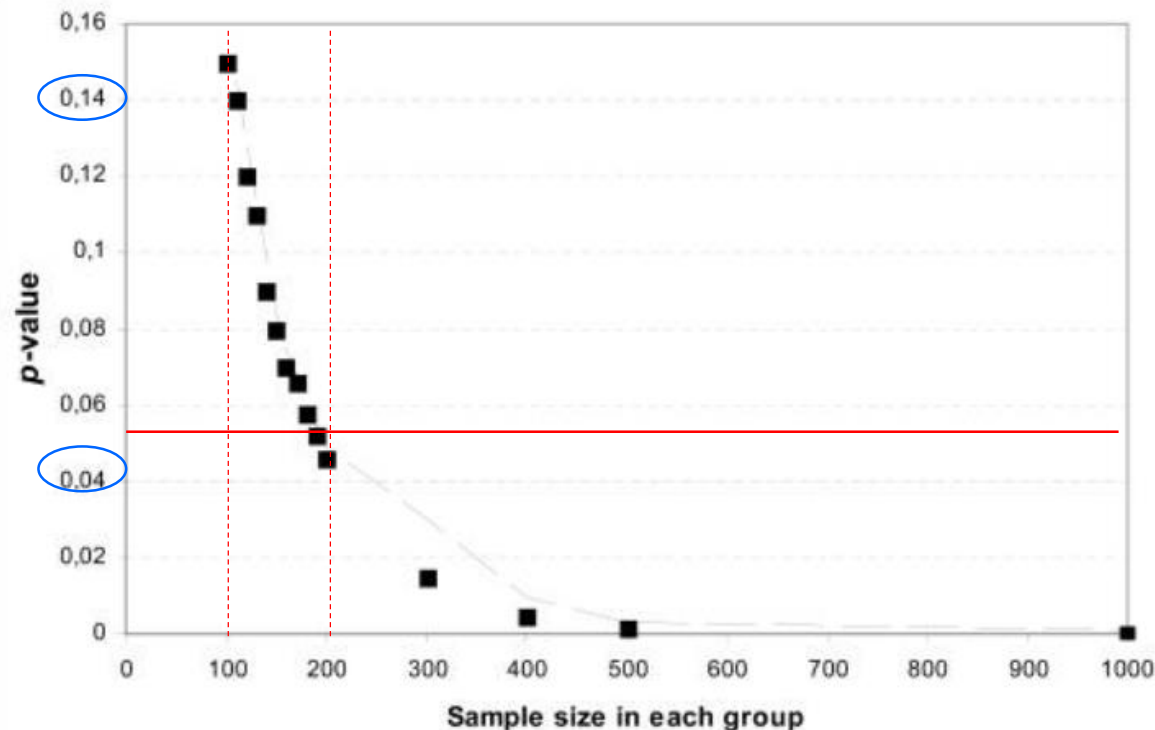


Table 1. Key Questions to Ask When the Primary Outcome Is Positive.

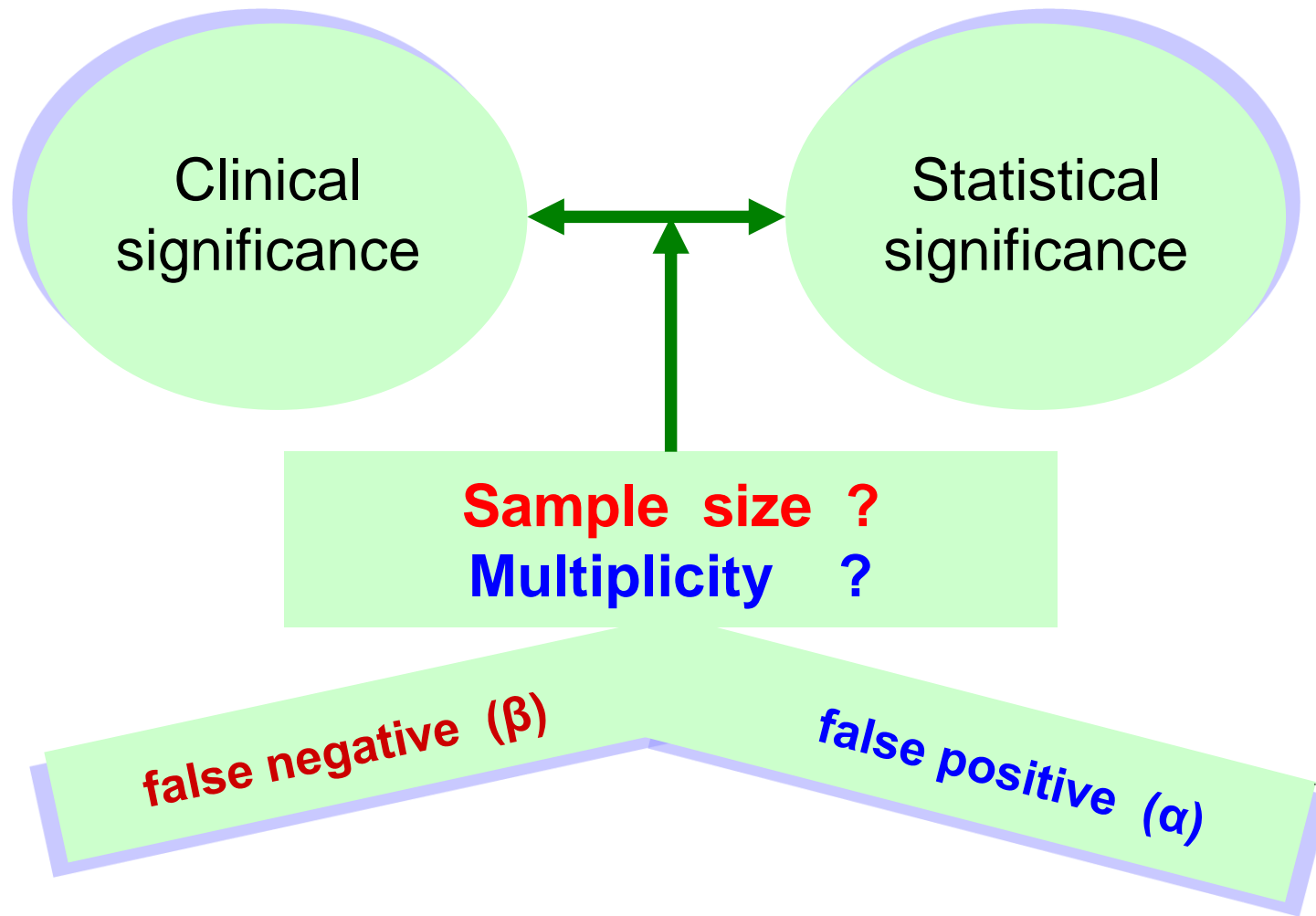
Stuart et al., N Engl J Med 2016;375:971-9.

- Does a P value of <0.05 provide strong enough evidence?
- What is the magnitude of the treatment benefit?
- Is the primary outcome clinically important (and internally consistent)?
- Are secondary outcomes supportive?
- Are the principal findings consistent across important subgroups?
- Is the trial large enough to be convincing?
- Was the trial stopped early?
- Do concerns about safety counterbalance positive efficacy?
- Is the efficacy–safety balance patient-specific?
- Are there flaws in trial design and conduct?
- Do the findings apply to my patients?

Table 1. Questions to Ask When the Primary Outcome Fails.

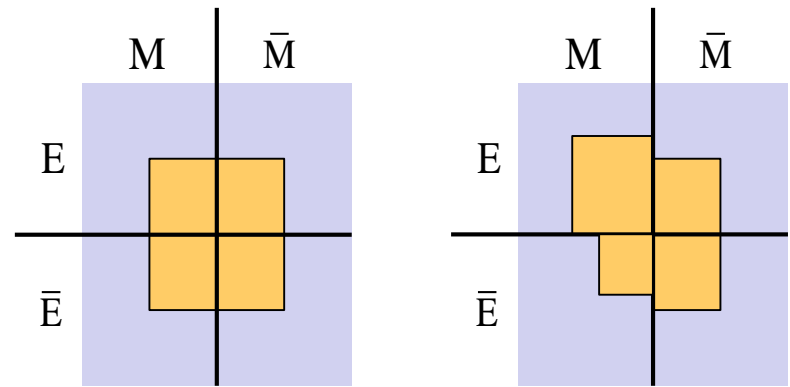
- Is there some indication of potential benefit?
- Was the trial underpowered?
- Was the primary outcome appropriate (or accurately defined)?
- Was the population appropriate?
- Was the treatment regimen appropriate?
- Were there deficiencies in trial conduct?
- Is a claim of noninferiority of value?
- Do subgroup findings elicit positive signals?
- Do secondary outcomes reveal positive findings?
- Can alternative analyses help?
- Does more positive external evidence exist?
- Is there a strong biologic rationale that favors the treatment?

Stuart et al., N Engl J Med 2016;375:861-70.



- ✓ Type III error : 연구가설 ???
- ✓ Bias 줄이기 위한 노력 + 시간

연구설계 단계 bias



Selection bias

- ✓ sampling frame bias : admission rate bias (Berksonian bias)
- ✓ non random sampling bias : detection bias
- ✓ non-converge bias : loss to follow-up bias, withdrawal bias

Non comparability bias

- ✓ lead time bias, length bias, historical control bias

Sample size bias

자료수집 과정에서의 information bias

instrument bias

data source bias

observer bias

- ✓ diagnostic suspicion bias
- ✓ exposure suspicion bias
- ✓ therapeutic bias (→ Blinding)

subject bias

- ✓ proxy respondent bias
- ✓ recall bias
- ✓ attention bias (“Hawthorne effect”)

분석 & 결과 해석 과정에서의 bias

confounding bias

analysis strategy bias

: missing data handling, outlier handling, unit of analysis

post-hoc analysis bias (← data dredging bias)

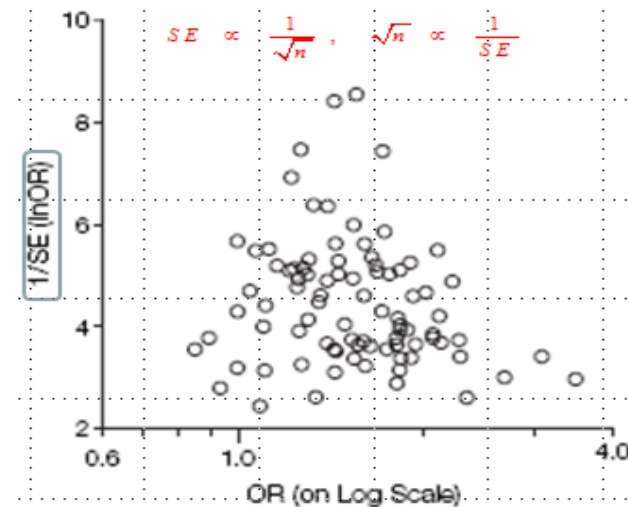
assumption bias

generalization bias (← lack of external validity)

significance bias

: statistical significance vs. biological significance

Publication bias (by Funnel plot, Egger의 회귀비대칭성 검정)



✓ **Association (연관성 聯關性)**
동질성/독립성

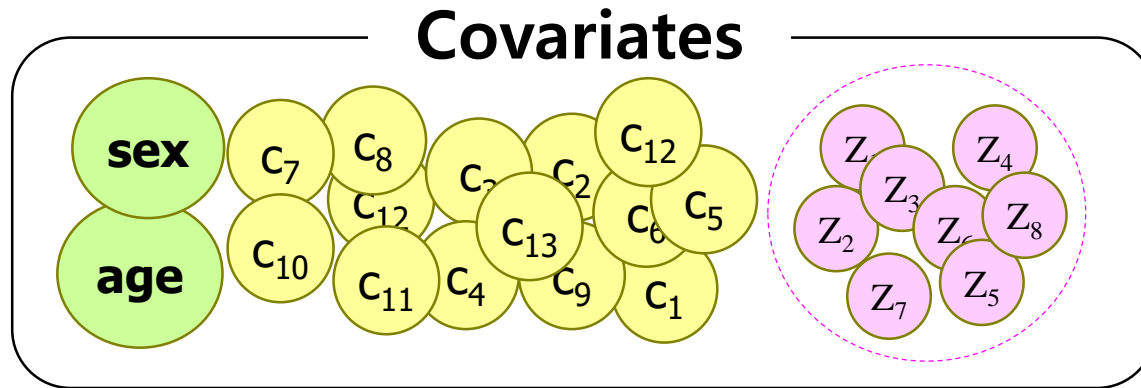
✓ **Correlation (상관성 相關性)**

✓ **Probability (개연성 蓋然性)**

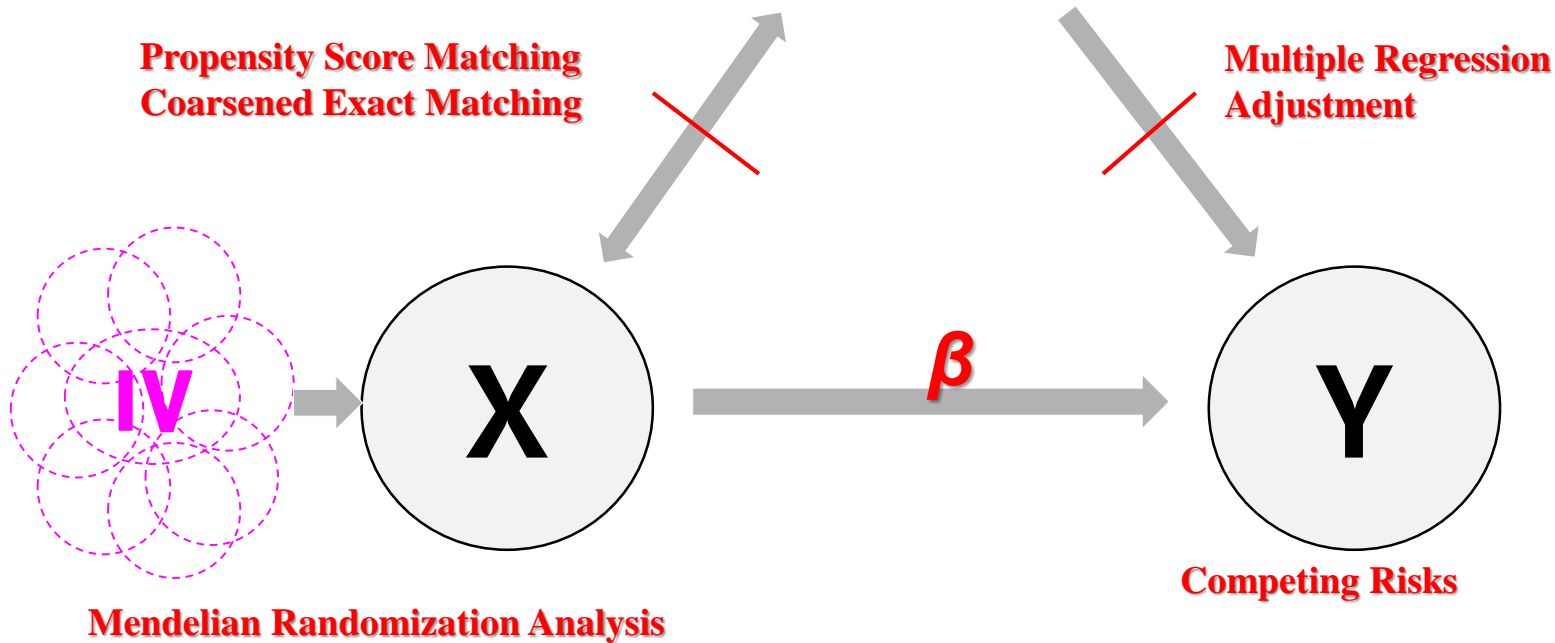
✓ **Causality (인과성 因果性)**

Statistical Methods for Causal Inference

- confounders variables
- unmeasured/unknown confounders
- stratification variables
- intermediate variables
- effect modifier / interaction effect



+ 'Time' 을 어떻게 보정할 것인가?



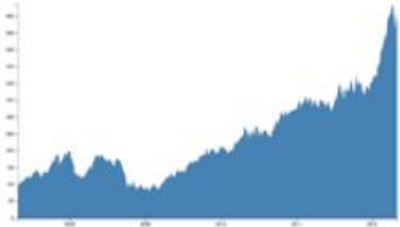
- Multiple Regression Analysis
- Logistic Regression Analysis
- Poisson Regression Analysis
- Cox's PHM
- Linear Mixed Model (LMM)
- Generalized Estimating Equation (GEE)

Data-Driven Documents

<http://d3js.org>

“Big Data”
Volume + Variety + Velocity
(Value + Veracity + Complexity)
Data Visualization

Area Chart



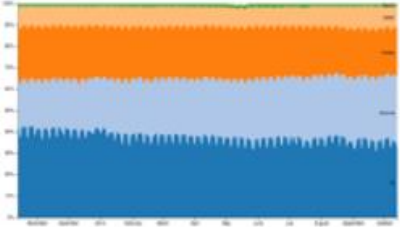
Line Chart



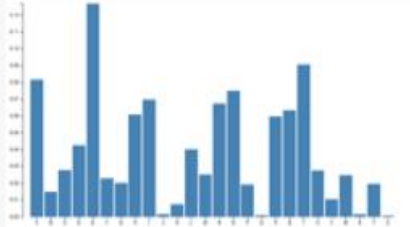
Bivariate Area Chart



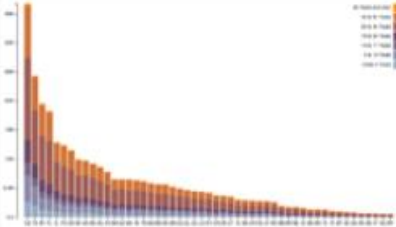
Stacked Area Chart



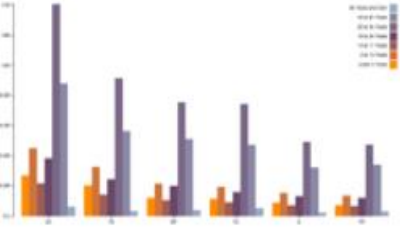
Bar Chart



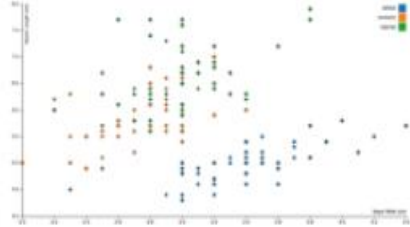
Stacked Bar Chart



Grouped Bar Chart



Scatterplot



Donut Chart

