Statistical Methods in Medical Research

Joshua Naranjo

Department of Statistics
Western Michigan University

Outline

1. Survey of statistical methodology
2. Survey of contradicted studies
3. Threats to correct outcomes

1979 Methods Survey

"Use of Statistical analysis in *The New England Journal of Medicine*", by Emerson and Colditz (1983)

Methods: Reviewed 760 articles from January 1978 through December 1979). At least two statisticians reviewed each article for the presence of statistical methods in each of 21 categories.

Results:

- "A reader who is conversant with descriptive statistics (percentages, means, standard deviations) has statistical access to 58 percent of the articles."
- "Understanding t-tests increases this access to 67 percent."
- "Addition of contingency tables increases access to 73 percent."

1979 Methods Survey

Table 2. Statistical Content and Accessibility of *Journal* Articles.

PROCEDURE	ARTICLES CONTAINING METHODS	ACCUMULATED BY ARTICLE	ACCESSIBILITY BY ARTICLE-METHOD
	no. (%)	*no.* (%)	(%)
No statistical methods or descriptive statistics only	443 (58)	443 (58)	(40)
t-test	179 (24)	509 (67)	(56)
Contingency tables	112 (15)	551 (73)	(66)
Non-parametric tests	45 (6)	571 (75)	(70)
Epidemiologic statistics	39 (5)	585 (77)	(73)
Pearson correlation	55 (7)	598 (79)	(78)
Simple linear regression	37 (5)	621 (82)	(81)
Analysis of variance	33 (4)	636 (84)	(84)
Transformation	26 (3)	650 (86)	(87)

1979 Methods Survey

Non-parametric correlation	15 (2)	662 (87)	(88)
Life table	24 (3)	674 (89)	(90)
Multiple regression	19 (3)	686 (90)	(92)
Multiple comparisons	13 (2)	698 (92)	(93)
Other methods	17 (2)	708 (93)	(94)
Adjustment and standardization	13 (2)	718 (95)	(96)
Multiway tables	12 (2)	728 (96)	(97)
Power	13 (2)	737 (97)	(98)
Other survival analysis	11 (1)	747 (98)	(99)
Regression for survival	6 (1)	753 (99)	(99)
Cost–benefit analysis	6 (1)	758 (100)	(100)
Sensitivity analysis	2 (0)	760 (100)	(100)
Totals:			
Article-methods used	1120		
Articles		760	

1979 Methods Survey

Table 3. Statistical Content of 332 Original Articles, According to Study Design.

PROCEDURE	ORIGINAL ARTICLES (n = 332)	PROSPECTIVE (n = 182)	RETROSPECTIVE (n = 18)	CROSS-SECTIONAL (n = 132)
	NO. (%)	DEVIATION FROM OVERALL % *		
No statistical methods or descriptive statistics only	91 (27)	−11	−16	+18
t-tests	147 (44)	+4	−11	−4
Contingency tables	91 (27)	+9	+23	−16
Non-parametric tests	38 (11)	+3	+6	−5
Epidemiologic statistics	33 (9)	0	+40	−5
Pearson correlation	40 (12)	0	−1	0
Simple linear regression	28 (8)	+1	+3	−1
Analysis of variance	25 (8)	+1	−8	−1
Transformation	23 (7)	0	−1	0

1979 Methods Survey

Non-parametric correlation	13 (4)	0	+4	+1
Life table †	36 (11)	+7	+4	−11
Multiple regression	15 (5)	+1	+1	−1
Multiple comparisons	11 (3)	+2	−3	−1
Adjustment and standardization	9 (3)	+2	−3	−3
Multiway tables	12 (4)	+1	+8	−3
Power	10 (3)	+1	+3	−2
Cost–benefit analysis	3 (1)	0	−1	0
Other methods	12 (3)	0	−3	+1

*Deviations from overall percentage were calculated by subtracting the overall percentage of original articles (e.g., 27 in row one) from the percentage of articles in each study design using the procedure (e.g., descriptive statistics only). For prospective studies, 16 per cent used descriptive statistics only, giving 16 − 27 = −11. Likewise for retrospective studies, the use in this group (11 per cent) minus the overall percentage (27 per cent) gives a residual of −16. Cross-sectional studies used only descriptive statistics in 45 per cent of studies and so yield a residual of +18 (45 − 27 = 18).

†Life table here includes the original categories of life table, regression for survival, and other survival.

2005 Methods Survey

"Statistical Methods in *The Journal*", by Horton and Switzer (2005)

Methods: We have updated previous surveys of Original Articles in 1979 and 1989 with data from 311 articles published from January 2004 through June 2005.

Results: Our findings show a continued trend toward increased use of newer and more complex methods not typically included in introductory or second-level statistics courses. The increasing sophistication of statistical methods has potential implications for medical and statistical educators.

2005 Methods Survey

Table 1. Statistical Content of Original Articles in the *New England Journal of Medicine* over Time.*

Statistical Procedure	Original Articles Containing Methods			Accumulation by Article†
	1978–1979	1989	2004–2005	2004–2005
			number (percent)	
No statistical methods or descriptive statistics only	91 (27)	14 (12)	39 (13)	39 (13)
t-Tests	147 (44)	45 (39)	80 (26)	42 (14)
Contingency tables	91 (27)	41 (36)	166 (53)	47 (15)
Nonparametric tests	38 (11)	24 (21)	85 (27)	53 (17)
Epidemiologic statistics	33 (10)	25 (22)	110 (35)	55 (18)
Pearson's correlation	40 (12)	22 (19)	10 (3)	56 (18)
Simple linear regression	28 (8)	10 (9)	19 (6)	56 (18)
Analysis of variance	25 (8)	23 (20)	50 (16)	61 (20)
Transformation	23 (7)	8 (7)	31 (10)	62 (20)

2005 Methods Survey

Nonparametric correlation	13 (4)	1 (1)	14 (5)	65 (21)
Survival methods	36 (11)	37 (32)	190 (61)	74 (24)
Multiple regression	15 (5)	16 (14)	160 (51)	122 (39)
Multiple comparisons	11 (3)	10 (9)	70 (23)	127 (41)
Adjustment and standardization	9 (3)	10 (9)	3 (1)	128 (41)
Multiway tables	12 (4)	11 (10)	39 (13)	136 (44)
Power analyses	10 (3)	4 (3)	121 (39)	211 (68)
Cost–benefit analysis	3 (1)	0	1 (<1)	212 (68)
Sensitivity analysis‡	0	0	18 (6)	223 (72)
Repeated-measures analysis	—	—	37 (12)	249 (80)
Missing-data methods	—	—	26 (8)	272 (87)
Noninferiority trials	—	—	11 (4)	283 (91)
Receiver-operating characteristic	—	—	7 (2)	288 (93)
Resampling	—	—	5 (2)	293 (94)

2005 Methods Survey

Resampling	—	—	5 (2)	293 (94)
Principal component analysis and cluster analysis	—	—	5 (2)	298 (96)
Other methods§	12 (4)	10 (9)	13 (4)	311 (100)
Total				
Articles	332	115	311	311
Article-method uses	637	311	1310	
Average uses of methods per article	1.9	2.7	4.2	

* The data in columns 2 and 3 are from Emerson and Colditz.[2] Dashes denote no data.
† The accumulation by article indicates the number and percentage of articles including only that procedure and those listed above it (excluding methods listed further down in the table). For example, a total of 53 papers used only descriptive statistics, t-tests, contingency tables, and nonparametric tests (17 percent of the 311 papers reviewed).
‡ Sensitivity analysis was not reported for the 1978–1979 or the 1989 survey of Original Articles, so a value of zero was assumed.
§ The 13 articles listed as "other" for the 2004–2005 survey included five statistical procedures for genetics and two meta-analyses.

2005 Methods Survey: Accessibility

Accessibility jumps at:

- Multiple regression (24 to 39 percent)
- Power analysis (44 to 68 percent)
- Repeated measures (72 to 80 percent)
- Missing data (80 to 87 percent)

2005 Methods Survey: Trends

Statistical Procedure	1979 n=332 (%)	1989 n=115 (%)	2005 n=311 (%)
t-Tests	44	39	26
Contingency tables	27	36	53
Nonparametric tests	11	21	27
Epidemiologic statistics	10	22	35
Pearson correlation	12	19	3
Survival methods	11	32	61
Multiple regression	5	14	51
Multiple comparisons	3	9	23
Power analysis	3	3	39

2005 Methods Survey: Trends

Categories added in 2005 survey:

Statistical Procedure	1979 n=332 (%)	1989 n=115 (%)	2005 n=311 (%)
Repeated measures analysis	.	.	12
Missing-data methods	.	.	8
Noninferiority trials	.	.	4
Receiver-operating characteristic	.	.	2
Resampling	.	.	2
PCA and cluster analysis	.	.	2
Other methods	4	9	4

2015 Design and Methods Survey

"*Statistical Use in Clinical Studies: Is There Evidence of a Methodological Shift?*", by Yi et al. (2015)

Methods: Reviewed 838 eligible clinical study articles that were published in 1990, 2000, and 2010 in four journals New England Journal of Medicine, Lancet, Journal of the American Medical Association and Nature Medicine.

Exclusion criteria: Comments, case reports, systematic reviews, meta-analyses, genome-wide analyses and articles did not involve primary or secondary data analysis were excluded.

Results: Over time, there has been increased attention on the details of selecting a sample and controlling bias, and a higher frequency of utilizing complex statistical methods.

2015 Design and Methods Survey

Table 1: Study types

Categories	Total n=838	1990 n=223 (%)	2000 n=314 (%)	2010 n=301 (%)
Drug trial	62	70	61	56
Medical device	4	3	6	3
Operation method	6	3	8	6
Others*	28	24	25	35

*Training, diet, physical therapy, etc.

2015 Design and Methods Survey

Table 2: Study designs

Categories	Total n=838	1990 n=223 (%)	2000 n=314 (%)	2010 n=301 (%)
RCT	81	74	83	84
Case-control	5	11	5	1
Cohort	6	2	4	10
Descriptive*	8	12	8	5

*Cross-sectional study, sample survey and prevalence study

2015 Design and Methods Survey

Table 3: Study comparison type

Categories	Total n=838	1990 n=223 (%)	2000 n=314 (%)	2010 n=301 (%)
Difference	55	75	49	47
Superiority	25	15	27	31
Non-inferiority	17	8	21	18
Equality	3	2	3	4

2015 Design and Methods Survey

Table 4: Statistical software

Categories	Total n=838	1990 n=223 (%)	2000 n=314 (%)	2010 n=301 (%)
SAS	36	14	42	47
SPSS	30	25	32	34
STATA	9	3	11	11
R	4	1	4	5
Other	9	17	8	3
No mention	12	40	3	1

2007 Statistical Quality in *NEJM* and *Nat Med*

"The Use of Statistics in Medical Research: A Comparison of *The New England Journal of Medicine* and *Nature Medicine*", by Strasak et al.(2007)

Methods: We contrast the top journal for basic science *Nature Medicine* with the top journal for clinical science *NEJM* on current practices regarding the use of statistics in medicine. A sample of 53 papers (*NEJM* = 31, *Nat Med* = 22) was selected for a detailed quality assessment of statistical methods using a standardized 46-item checklist.

Inclusion Criteria: (1) the use of inferential methods beyond descriptive statistics and (2) the use of one or more elementary statistical significance tests to evaluate the primary outcome measure.

2007 Statistical Quality in *NEJM* and *Nat Med*

Table 1: Deficiencies related to study design

	NEJM n=31		*Nat Med* n=22	
Category	*n*	%	*n*	%
No sample size calculation	13	41.9	22	100.0
Failure to use and report randomization	1	3.2	13	59.1
Method of randomization not clear	6	19.4	4	18.2

2007 Statistical Quality in *NEJM* and *Nat Med*

Table 2: Deficiencies related to analysis

	NEJM (n=31)		*Nat Med* (n=22)	
Category	*n*	%	*n*	%
Incompatibility of test with data	4	12.9	1	4.5
No multiple-comparison correction	11	35.5	6	27.3
Special errors with Students t test				
No mention of check assumptions	16	51.6	13	59.1
No multiple-comparison correction	2	6.5	3	13.6
Special errors with chi-square tests				
No Yates correction if small *n*	4	12.9	0	0.0
Expected cell counts < 5	9	29.0	0	0.0

2007 Statistical Quality in *NEJM* and *Nat Med*

Table 3: Deficiencies related to reporting

	NEJM (n=31)		*Nat Med* (n=22)	
Category	*n*	%	*n*	%
Failure to state one or two-tailed	8	25.8	20	90.9
Failure to state paired or unpaired	18	58.1	17	77.3
Using mean but no measure of variability	2	6.5	0	0.0
Giving SE instead of SD for description	8	25.8	16	72.7
Failure to define ±; unlabeled error bars	2	6.5	4	18.2
Mean/SD for non-normal or ordinaldata	4	12.9	0	0.0
No confidence intervals for effect size	14	45.2	21	95.5
'p = NS', 'p<.05' instead of exact p value	6	19.4	19	86.4
'Nonsignificant' interpreted as 'no effect'	1	3.2	0	0.0
Disregard Type II error when nonsignificant	5	16.1	5	22.7

2005 Contradicted research outcomes

Questions:

1. How often do medical studies result in wrong findings?
2. What are the primary causes of wrong findings?
   - Statistical
   - Otherwise

Paper:
"*Contradicted and Initially Stronger Effects in Highly Cited Clinical Research*", by Ioannidis (2005)

Methods: We looked at all original clinical research studies published in 3 major general clinical journals (*NEJM*, *JAMA*, *Lancet*) or high-impact-factor specialty journals in 1990-2003 and cited more than 1000 times in the literature.

2005 Contradicted research outcomes

Results: Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these,

- 7 were contradicted by subsequent studies
- 7 found effects stronger than those of subsequent studies
- 20 found effect confirmed by subsequent studies
- 11 remained largely unchallenged

Conclusion: Contradiction and initially stronger effects are not unusual in highly cited research of clinical interventions and their outcomes.

2005 Contradicted research outcomes

Table 1: Contradicted research and current state of knowledge

Highly Cited Study	Other Research	Current state of knowledge
1. Nurses Health 2. PEPI	RCT (n=16,608)	Etrogen/progestin do not protect but increase CAD risk in postmenopausal women.
3. Health Pros	RCT (n=6996)	Vitamin E supplement does not reduce CAD in men.
4. Nurses Health	RCT (n=2545)	Vitamin E supplement does not reduce CAD in women.
5. CHAOS	RCT (n=9541)	Vitamin E supplement does not prevent coronary events

2005 Contradicted research outcomes

Table 1: Contradicted research and current state of knowledge

Highly Cited Study	Contradicting Research	Current state of knowledge
6. HA-1A Sepsis	RCT (n=2199)	HA-1A did not improve survival in gram-negative sepsis.
7. Rossaint et al (nitric oxide)	MA RCT (n=535)	Nitric oxide does not improve survival in respiratory distress.

2005 Contradicted research outcomes

Table 2: Contradicted research designs

Highly Cited Study	Highly cited study design	Contradicting study design
1. Nurses Health	Cohort (n=48,470)	RCT (n=16,608)
2. PEPI	RCT (n=875)	RCT (n=16,608)
3. Health Pros	Cohort (n=39,910)	RCT (n=6,996)
4. Nurses Health	Cohort (n=87,245)	RCT (n=2,545)
5. CHAOS	RCT (n=2,002)	RCT (n=9,541)
6. HA-1A Sepsis	RCT (n=200)	RCT (n=2,199)
7. Rossaint et al	Case series (n=9)	MA RCT (n=535)

The Vitamin E study

Health Professionals Study (Rimm et al, 1993):

An observational study of 51,529 male health professionals ranging in ages 40-75 in 1986. Those who consume more than 100 IU of vitamin E a day for a period of 2 years saw a reduction in the risk of coronary heart disease.

Contradicted by:

Heart Outcomes Prevention Evaluation Study (Yusuf et al, 2000):

2,545 women and 6,996 men were randomly given either a placebo or 400 IU of vitamin E daily for a mean of 4.5 years. There was no significant difference between the number of deaths in either group.

The Vitamin E study

Statistical explanation for contradiction:

The Health Professionals study is an observational study and vulnerable to the 'healthy cohort' effect where subjects who choose to take Vitamin E is as a group healthier than those who don't.

Moral:
Large sample size does not protect a study from biased sampling. The bias is simply repeated on a larger scale.

The Estrogen Study

The Nurses Health Study (Stampfer et al, 1991):

A cohort study involving 48,470 participants. The study found that estrogen use decreases incidence of coronary artery disease and mortality from cardiovascular disease.

Contradicted by:

Risks and Benefits of Estrogen in Postmenopausal Women (Rossouw et al, 2002):

16,608 women were randomized to estrogen, estrogen plus progestin, or placebo. Trial was stopped after 5.2 years of follow-up for overall risk exceeding benefits. Estrogen plus progestin group had increased risk of CHD, stroke, and breast cancer.

The Estrogen Study

Statistical explanation for contradiction:

The Nurses Health study is an observational study and vulnerable to the 'healthy cohort' effect where subjects who choose to take estrogen is as a group healthier than those who don't.

Moral:
Large sample size does not protect a study from biased sampling. The bias is simply repeated on a larger scale.

The Estrogen Study

There were many studies supporting the findings of the Nurses Health Study.

> *"Of 16 prospective studies, 15 found decreased relative risks, in most instances, statistically significant."*
> *(Stampfer, 1990)*

Nurses Health Study:

- Heart disease: 44% decreased risk
- Stroke: No difference

Women's Health Initiative:

- Heart attack: 29% increased risk (37 vs 30 per 10,000)
- Stroke: 41% increased risk (29 vs 21 per 10,000)

The Nitric Oxide study

The Nitric Oxide Study (Rossaint et al. 1993):

Consisted of 9 patients with severe ARDS (Adult Respiratory Distress Syndrome). Concluded that inhalation of nitric oxide in those with severe ARDS reduces pulmonary-artery pressure and increases arterial oxygenation.

Contradicted by:

Nitric Oxide Inhalation for Acute Hypoxemic Respiratory Failure

(Sokol et al, 2009)
535 patients with acute hyperemic respiratory failure. Nitric oxide had no effect on mortality rates.

The Nitric Oxide study

Statistical explanation for contradiction:

- Sample Size (n=9)
- No Control Group
  - Placebo Effect
  - Regression Effect
    All patients started with extremely low oxygenation
- Heterogeneous Cohort
  4 patients had pneumonia, 4 had trauma and lung contusion.
  Some had kidney or liver failures.
- Surrogate endpoint (oxygenation, not mortality)

Threats to correct outcomes

Threats to study reliability: (*Are the results repeatable?*)

- Low sample size
- Heterogeneous cohort (population variance)
- Instrument reliability (e.g. pain scale, anxiety scale)

Threats to study validity: (*Does it do what is intended?*)

- Confounded effects (cohort vs RCT)
- Surrogate endpoint (Y) and surrogate markers (X) (e.g. urinary sodium $\rightarrow$ BP)
- Instrument validity (e.g. anxiety vs stress, memory loss)