期刊論文常見的統計錯誤

台北醫學大學 生物統計暨研究諮詢中心

林彥光 副研究員

	NEMJ (文章數 =31)		*Nat Med* (文章數 =22)	
	n	%	n	%
研究設計				
無樣本數/檢定力計算	13	41.9	22	100.0
資料分析				
使用錯誤或較不適合的統計檢定	5	16.1	6	27.3
沒有校正多重比較 (multiple comparison)	11	35.5	6	27.3
方法說明				
沒有清楚且正確的定義或具體說明所有使用的統計檢定方法	20	64.5	20	90.9
結果呈現				
提供標準誤(SE)而不是標準差(SD)來描述資料	8	25.8	16	72.7
呈現 "p<0.05"、 "p>0.05"等，而非實際的 p 值	6	19.4	19	86.4
沒有提供主要效應值 (main effect size measures) 的信賴區間	14	45.2	21	95.5
結果解釋				
當結果為不顯著時，忽略討論型 II 錯誤	5	16.1	5	22.7
有多重顯著檢定時，沒有討論可能的問題	10	32.3	6	27.3

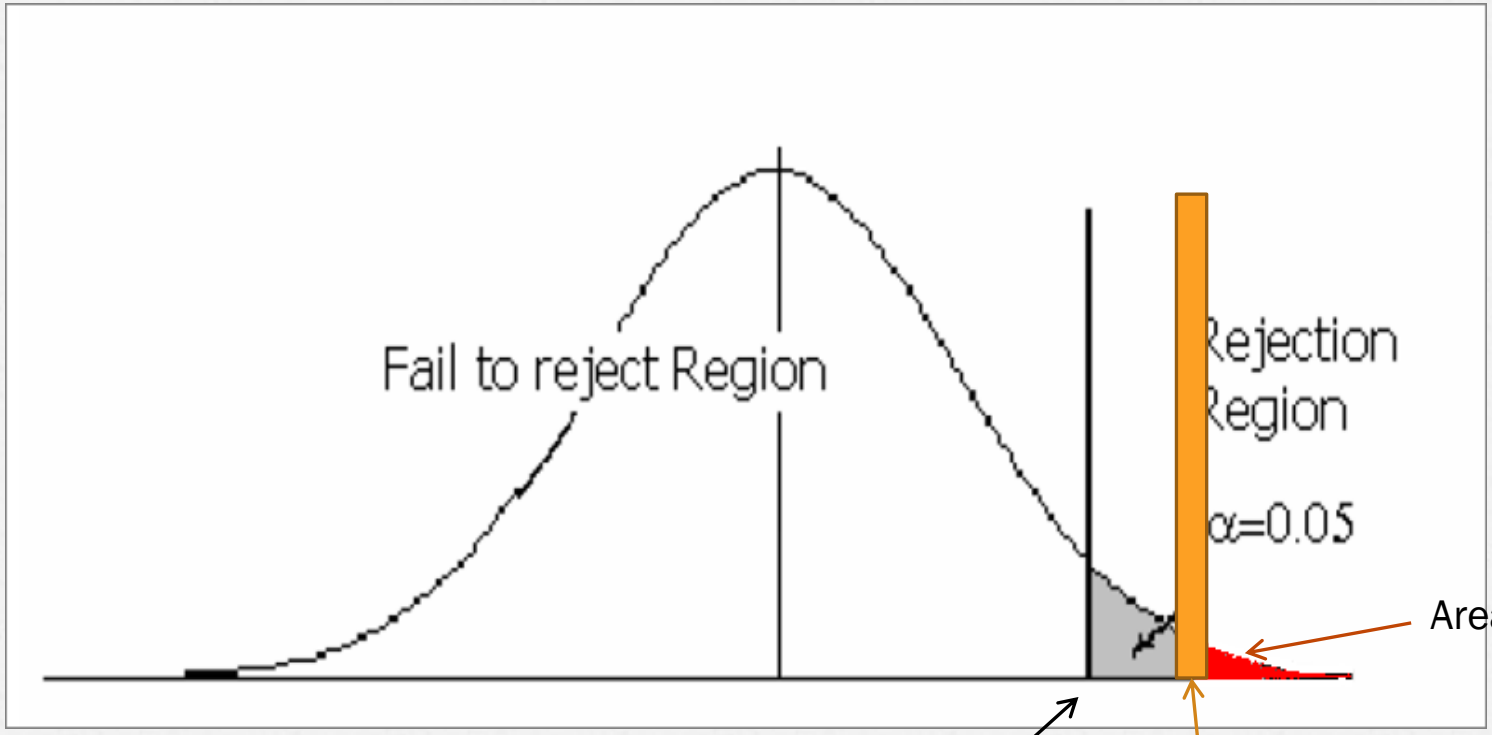
Error #1:Misuse of descriptive statistics

	Data 1	Data 2	Data3
value	1,2,3,4,5	1,2,3,4,500	-97, -47, 15, 47, 97
Mean	3	102	3
Median	3	3	15
SD	1.58	222.49	76.51
Range	4	499	194
Interquartile Range	2	2	94

Error #2: Misuse of p-value

- p=0.02 is more significant than p=0.03?
- p=0.02, so we proof that Ha is true?
- Absence of proof is not proof of absence（無效或 無法證明有效?）

Rejection Region



- A is significantly larger than C and (A>C)
  B is insignificantly different from C (B=C),
  then A is significantly larger than B(A>B)?

- Average change of blood pressure for 3 drugs

Type	Average Change	n	p	CI
A	-20	18	0.02	(-30, -1)
B	-0.2	18	0.03	(-3, -1)
C	-20	18	0.07	(-50,10)

Error #3: Fail to include the CI along with the estimate

- Is there a significant change on patients' compliance?

Mean	95% CL Mean		Std Dev	t Value	Pr > |t|
3.0000	0.0111	4.0999	5.0332	4.37	0.0222

- Although the p-value is <0.05, the average difference could be as low as 0.011

Error #4: Statistical significance against clinical importance

- Statistical significance v.s. practical significance
- Statistical insignificance v.s. practical insignificance
- Power and sample size

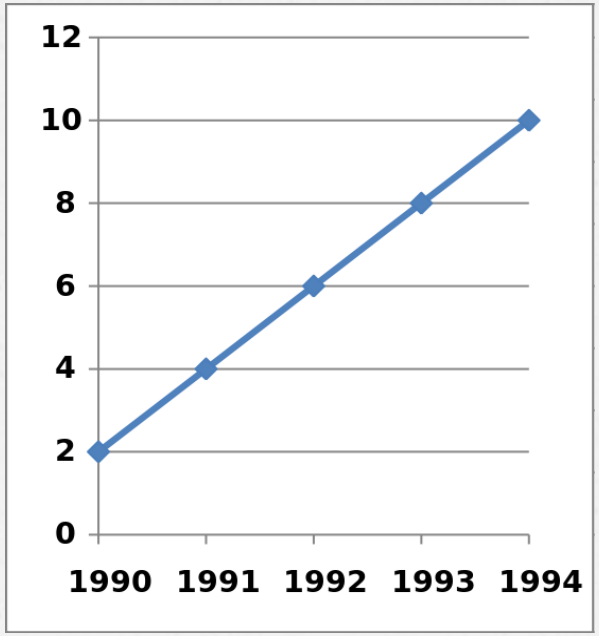
Expected difference (p_1-p_2)	Total sample size required*
5%	1450-3200
10%	440-820
20%	140-210
30%	80-100
40%	50-60

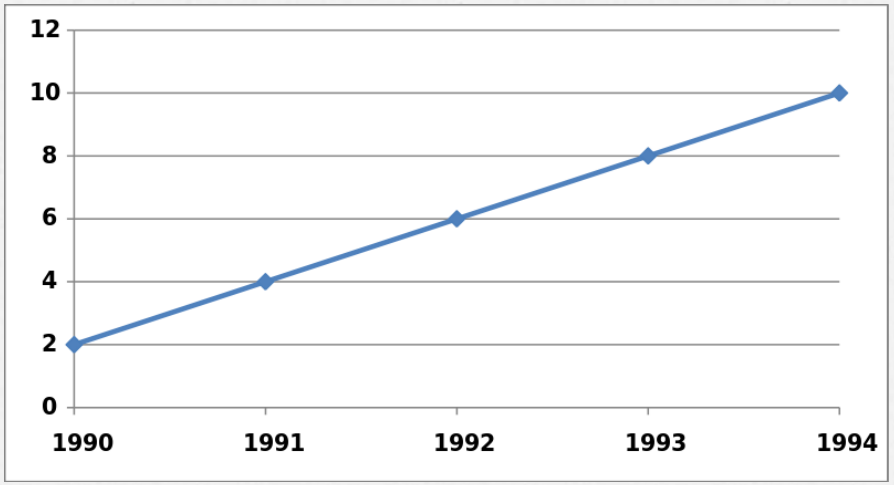
* 5% significance level, 80% power. Smaller numbers may be justified for rare outcomes (p_1 <.1)

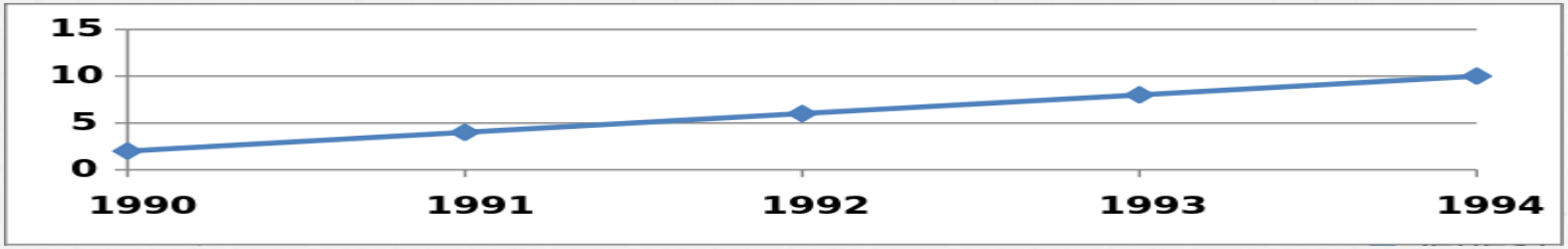
Error #5:Unnecesarily precise report

- The budget increased from 19,942 to 64,347
- The budget increased from 19,900 to 64,300
- The budget increased from about 20,000 to 64,000
- Group A is significantly lower than Group B (p=0.024578321)
- Group A is significantly different from 0 (p=0.000000000)

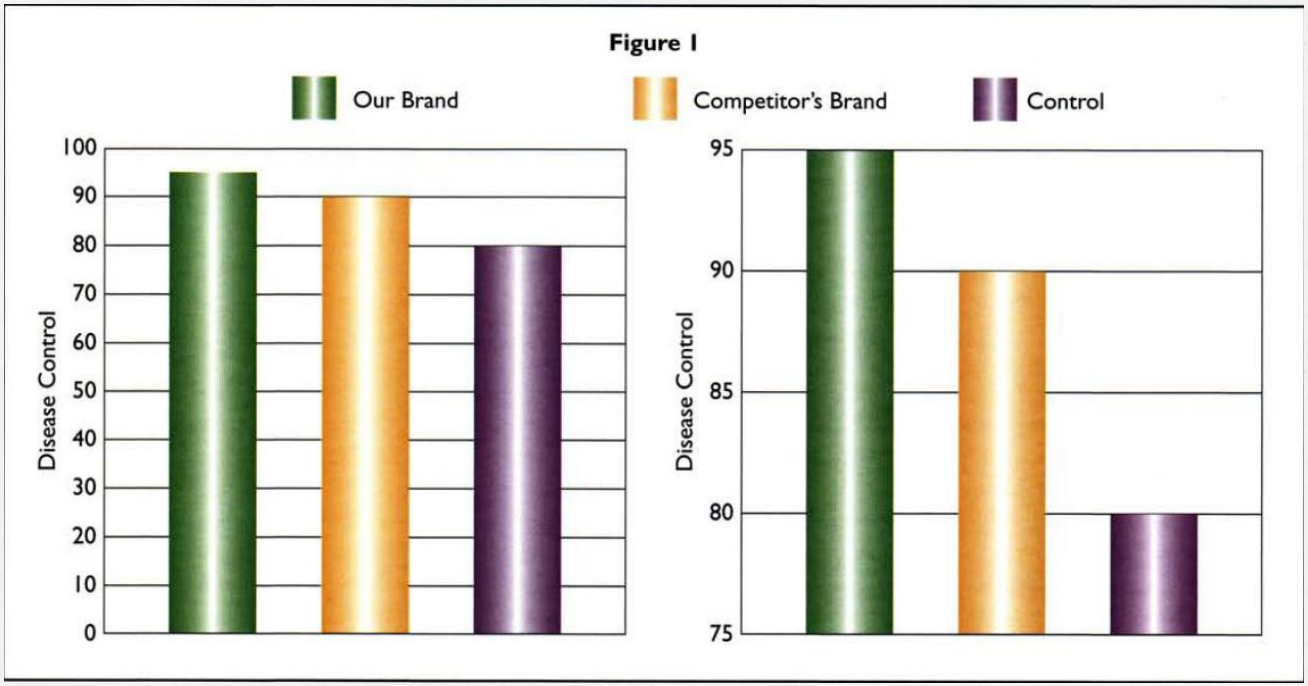
Error #6: Graphics that does not support the message of the data





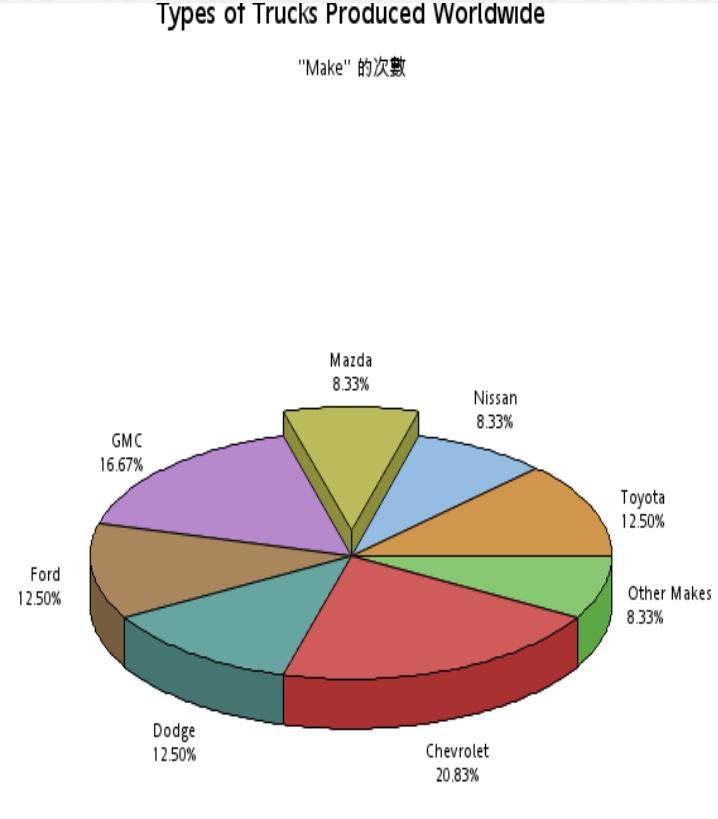


Axis changes



http://turf.unl.edu/extpresentationspdf/BairdStats.pdf

Misleading 3D charts



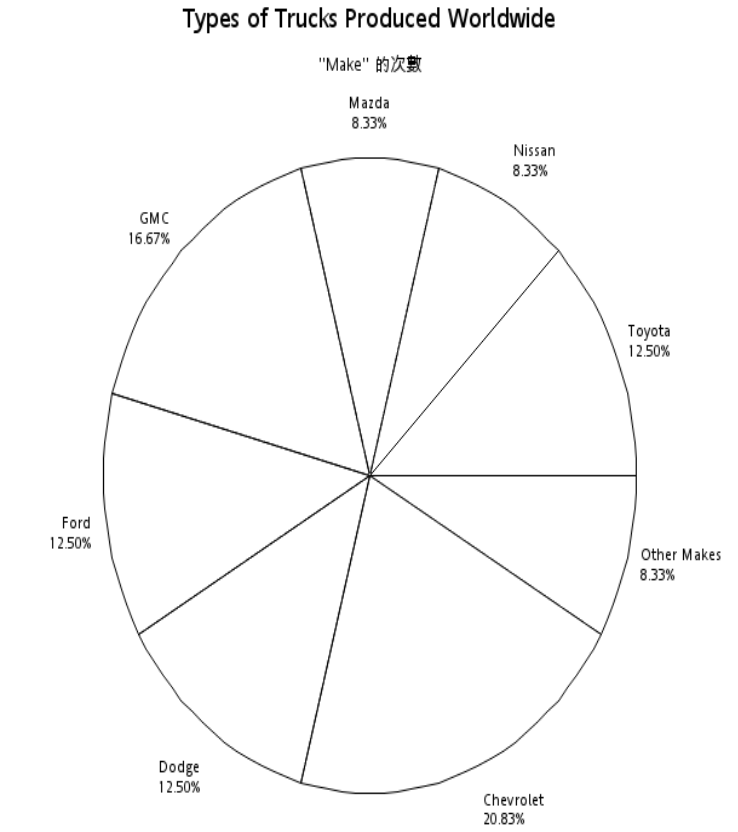


Figure 5: 3D Bar Chart. Needless Complexity.

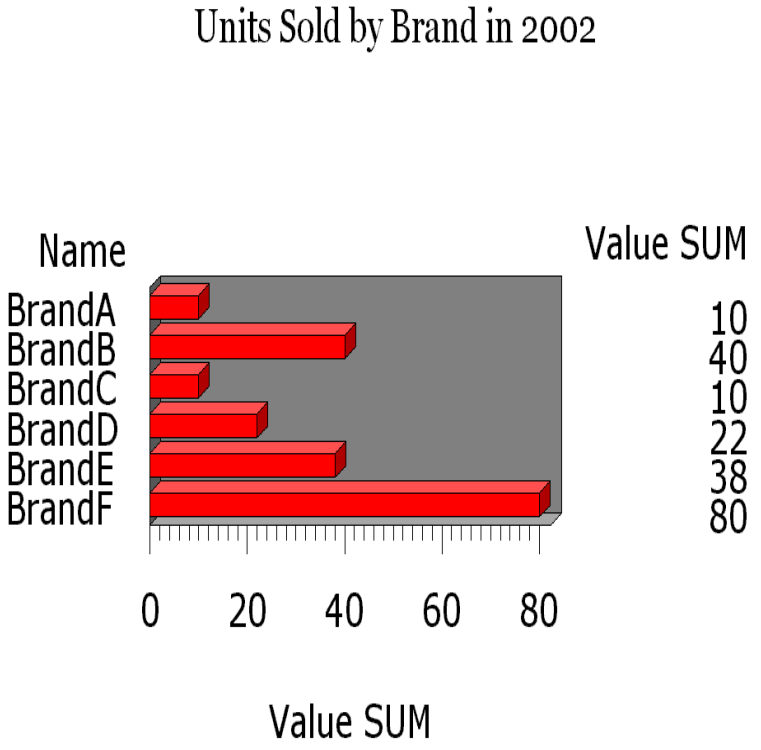
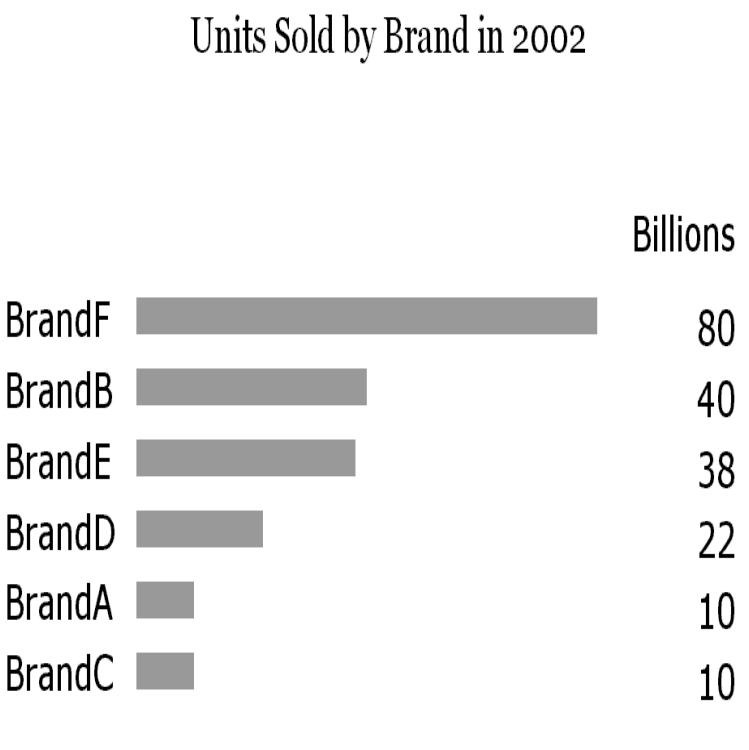


Figure 6: Best Horizontal Bar Chart.



Bessler. L. (sugi 28) Easy, Elegant, and Effective SAS® Graphs: Inform and Influence with Your Data

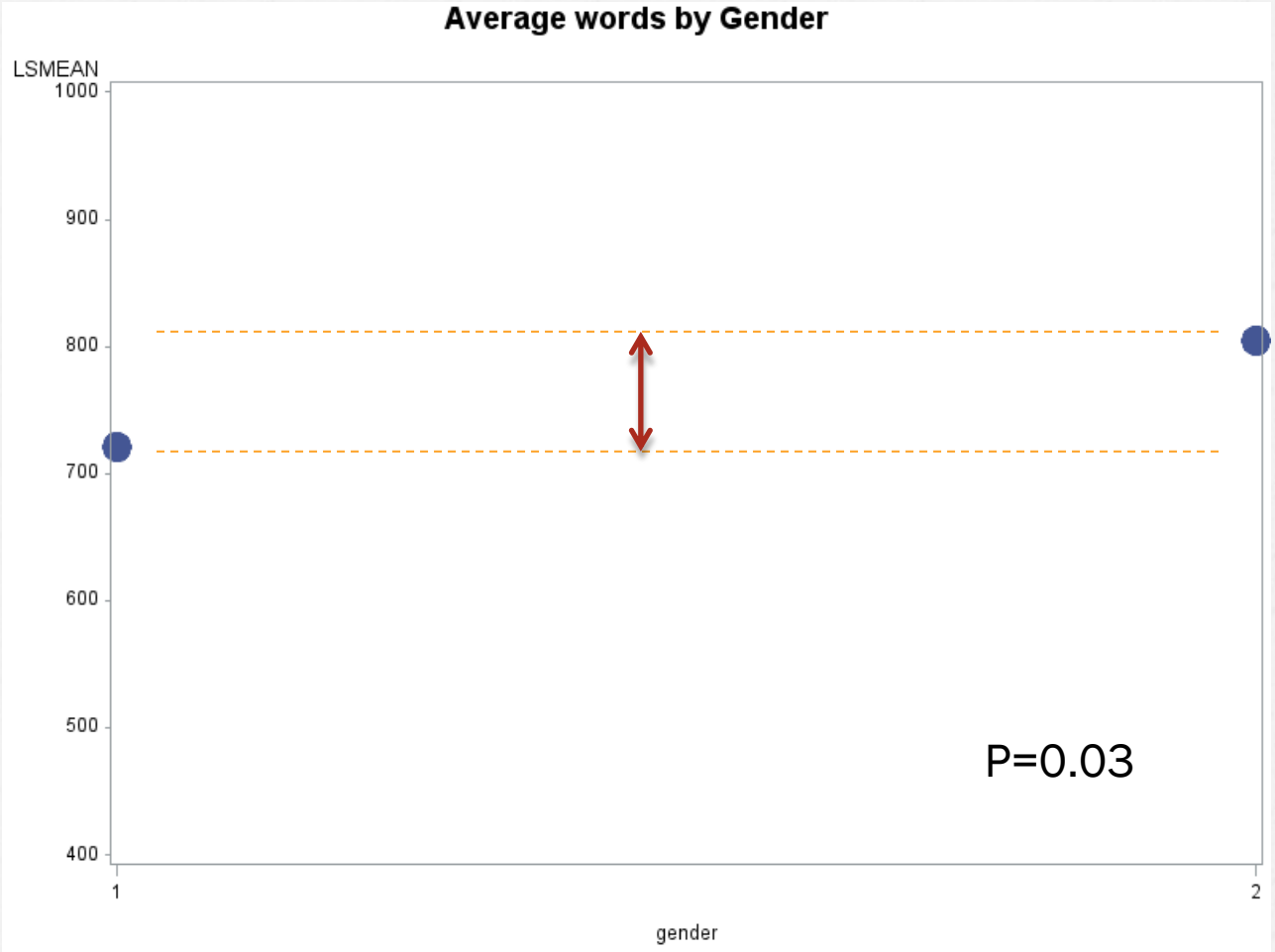
Error #7: Fail to interpret interaction correctly

- In two-way ANOVA, Factor A and Factor B may affect the outcome individually and collectively.
- Reading Speed Data:

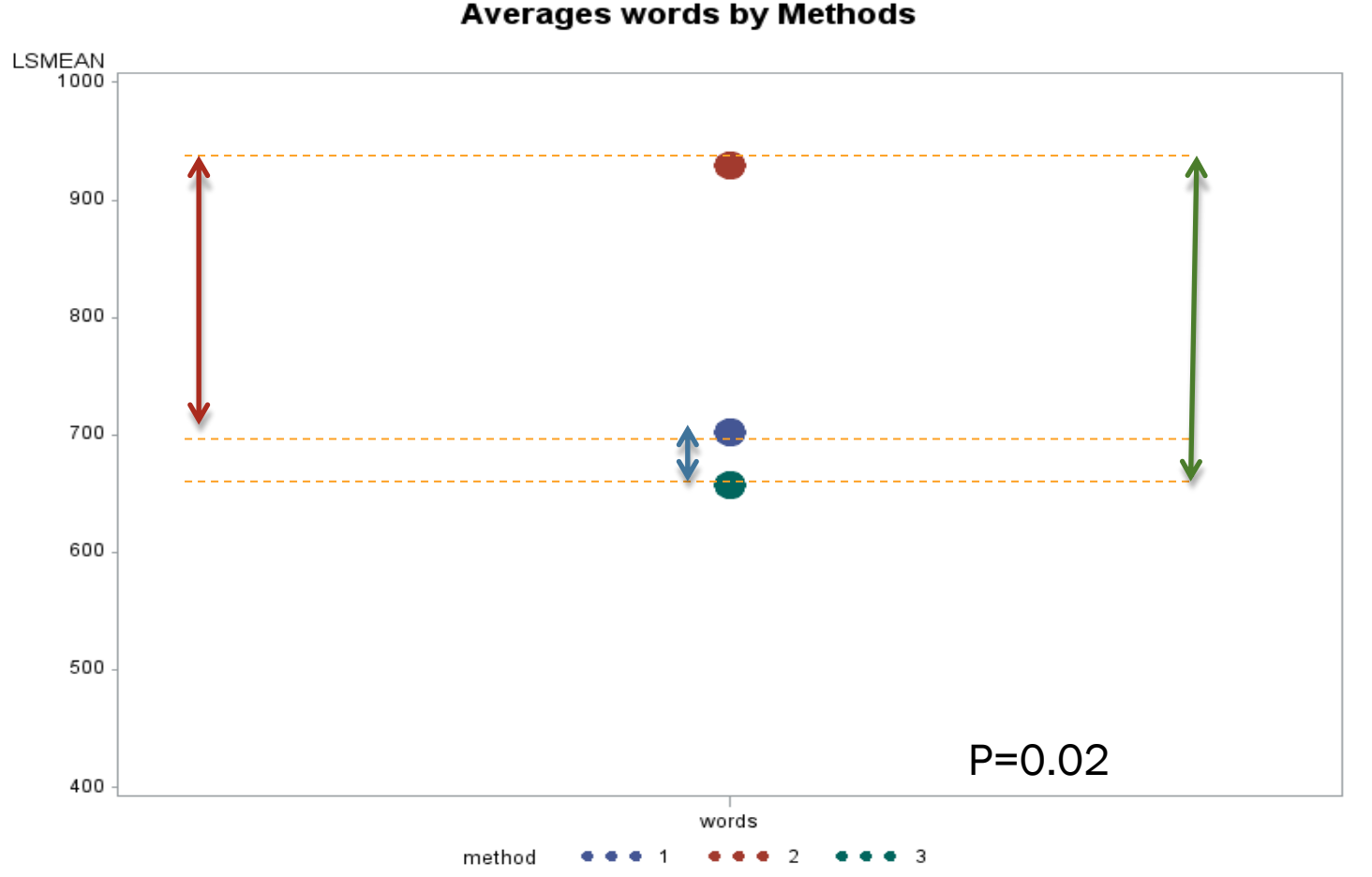
Gender	Method	Words
1	1	461
1	1	653
1	2	900
1	2	815
1	3	696
1	3	799
2	1	1000
2	1	693
2	2	1000
2	2	1000
2	3	560
2	3	576

- **Three hypothesis tests are:**
  - **Is there a Gender effect?**
  - **Is there a Method effect?**
  - **If there a Gender*Method effect?**

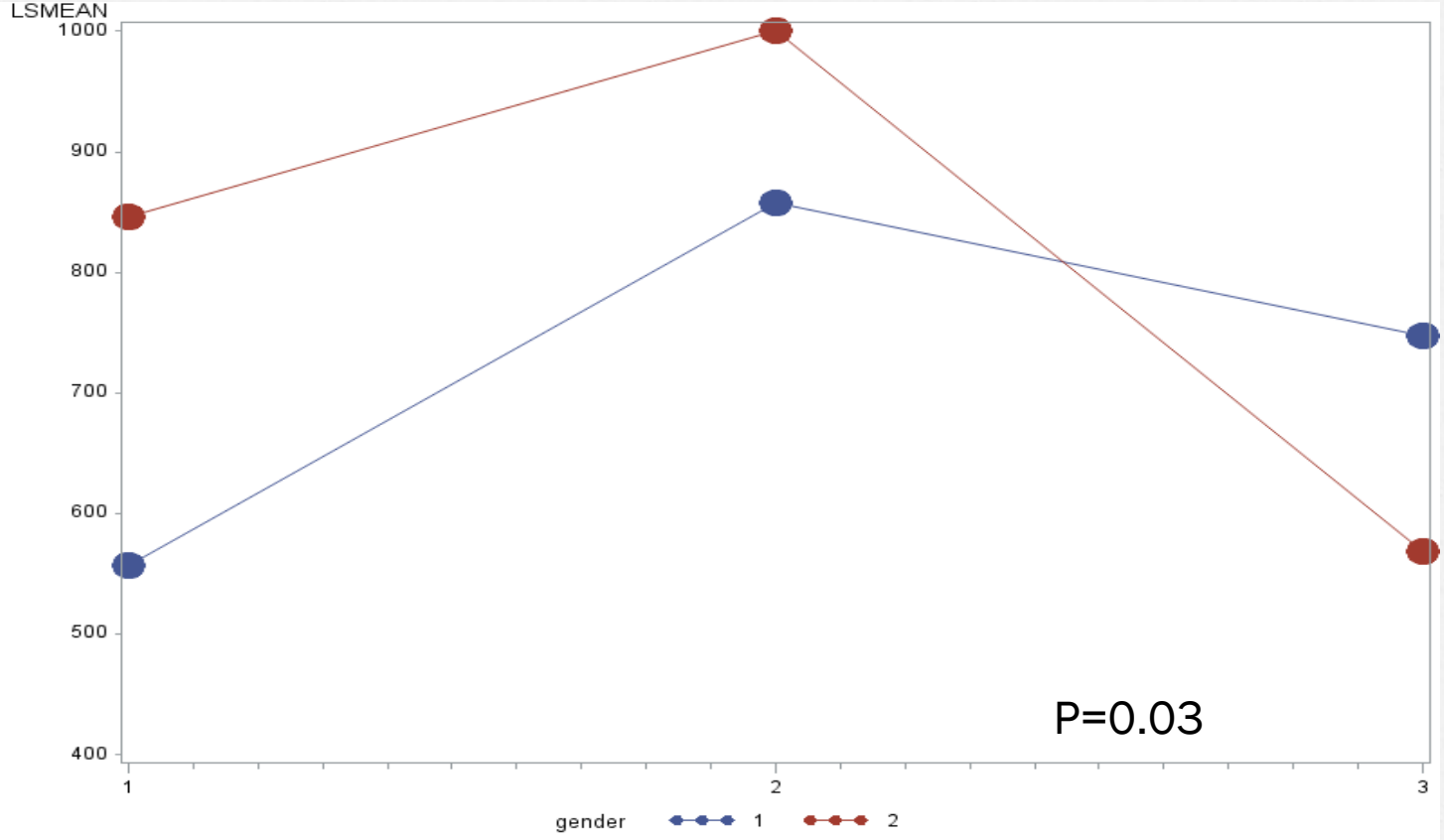
Is there a Gender effect?



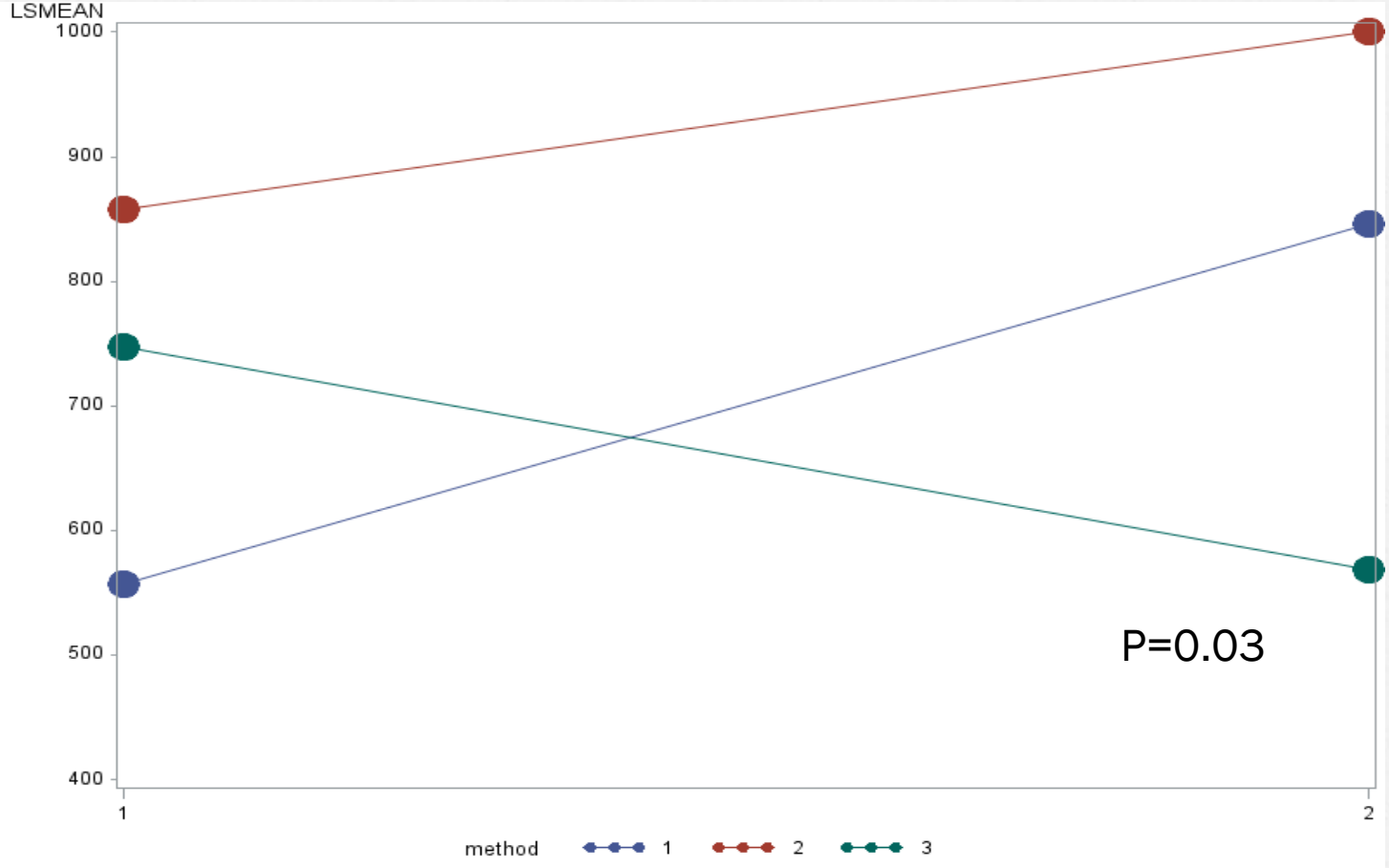
Is there a Method effect?



If there a Gender*Method effect?



Is there a Gender*Method effect?



例

目的：探討教師口語回饋對高低能力學生籃球精熟學習的影響。本研究以隨機抽取籃球選修課大學生二個班共80人為研究對象。方法：為實驗研究法，是以2×2的隨機因子設計交互分析。結果：學生在教師口語回饋的精熟學習情境下，

Result:

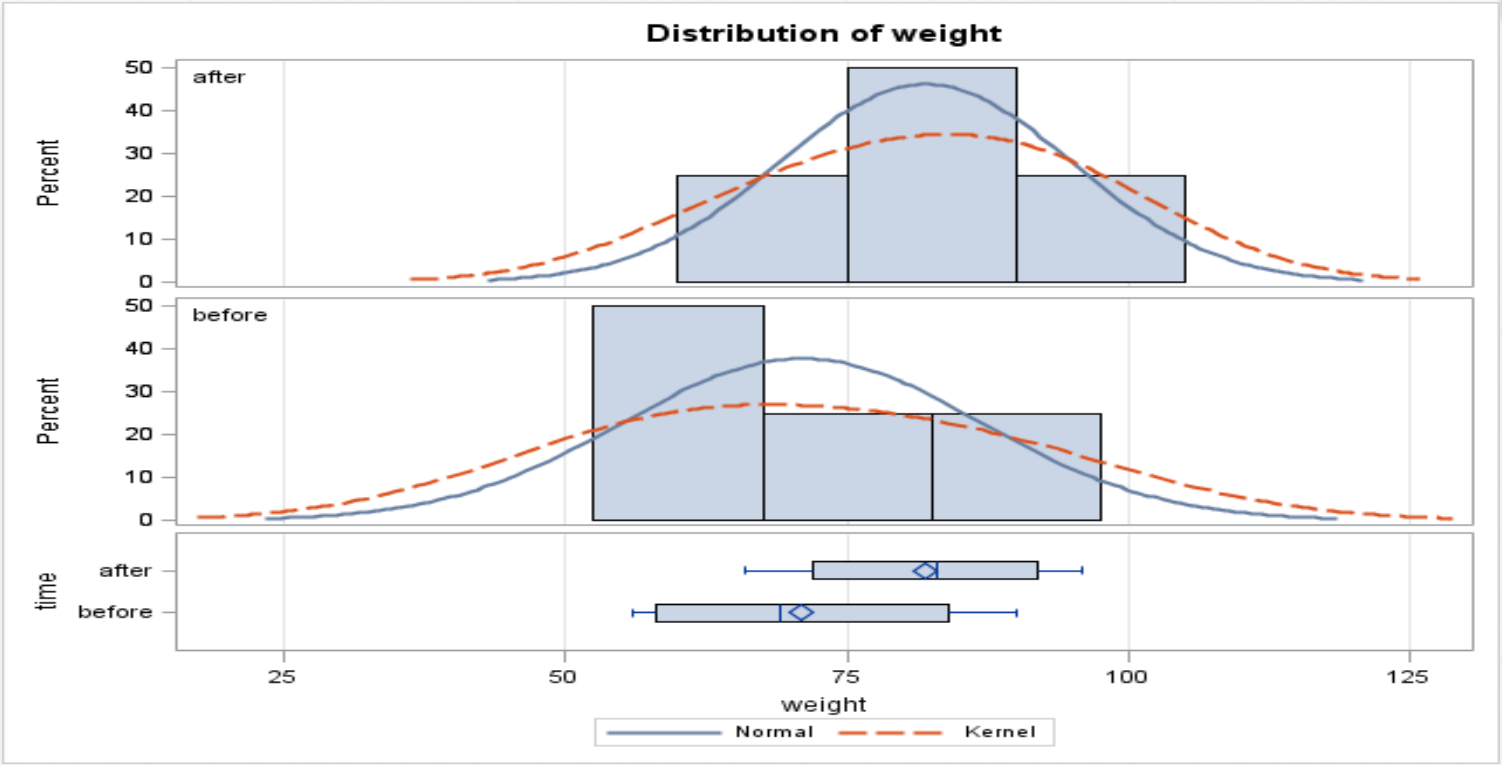
	口語回饋&精熟學習	精熟學習	總結
高能力	38.45	34.52	36.48
低能力	34.8	33.6	33.7
總結	36.88	33.64	35.26

結論1: 加入口語回饋較單純精熟學習顯得有效率
結論2: 高能力顯著超越低能力
結論3: 高能力且接收口語回饋有顯著高於低能力，但高能力且接受精熟學習與低能力無差別

Error #8:Fail to incorporate the correlation between repeated measures

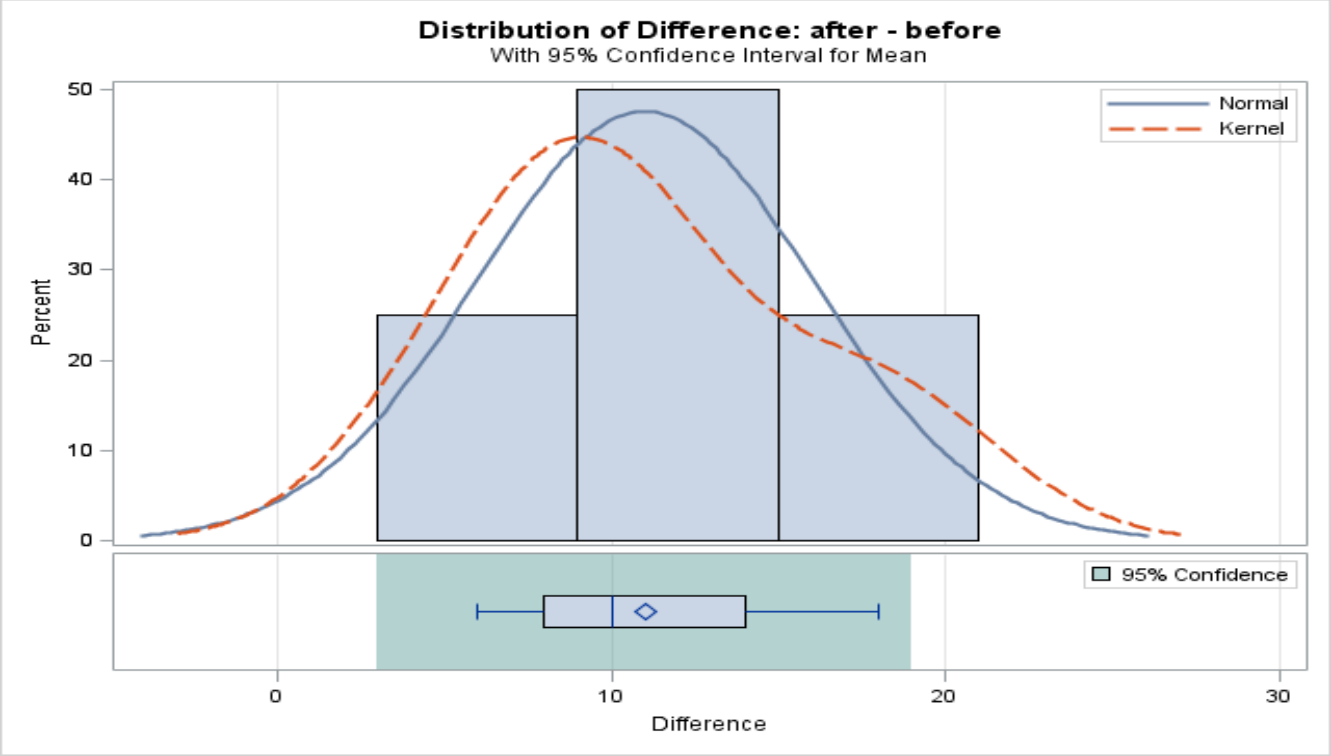
- Weight before/after diet data:

Patient	before	After	difference
1	60	78	18
2	56	66	10
3	90	96	6
4	78	88	10
	71	82	



Independent Sample T-Test

time	*Method*	*Mean*	*95% CL Mean*		*Method*	*Pr > |t|*
after		82.00	61.37	102.6		
before		71.00	45.74	96.25		
Diff (1-2)	Pooled	11.00	-14.07	36.07	Pooled	0.3243

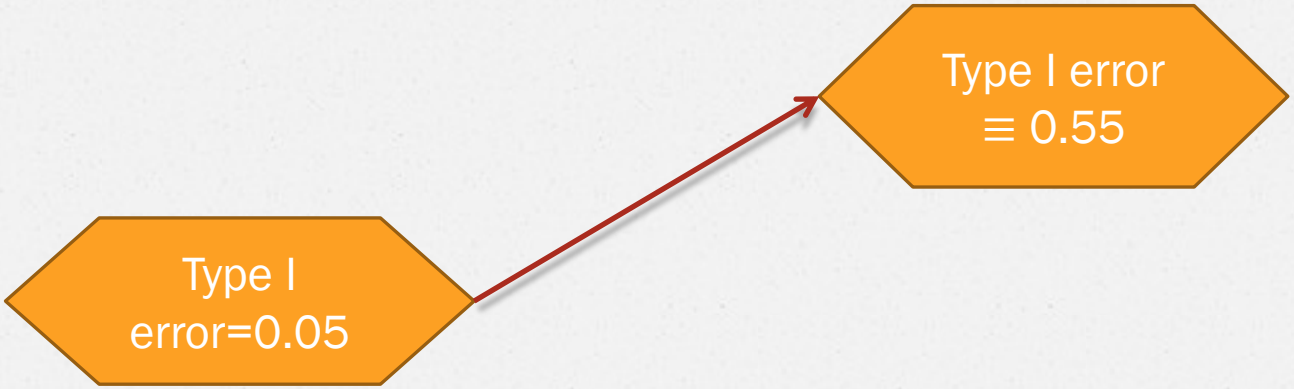


Pair T-Test

Mean	*95% CL Mean*		*Std Dev*	*DF*	*t Value*	*Pr > |t|*
11.00	2.99	19.00	5.02	3	4.37	0.0222

Error #9: unlimited multiple hypothesis tests?

- How many p's are you going to report in one study?
- Ex. For seven groups, 21 sets pairwise comparisons are possible for testing.



- Multiple testing occurs when
  - testing each of several baseline characteristics for differences between groups
  - multiple pair-wise comparisons
  - testing multiple endpoints
  - performing secondary analyses of relationships observed during the study;

例

目的：探討動態恢復對漸增性衰竭運動後血液中淋巴細胞數目及亞群比例變化情形。方法：本實驗對象爲 20 位高中田徑隊男性選手，其中 10 名爲動態恢復組（平均年齡：16.2 ± 1.1 歲，最大攝氧量：57.3 ± 8.7 ml・min-1・kg-1）而 10 名爲安靜休息組（平均年齡：16.6 ± 1.0 歲，最大攝氧量：58.5 ± 9.0 ml・min^{-1}・kg^{-1}）。兩組在跑步機上進行漸增衰竭運動，動態恢復組在衰竭運動後，再接著 20 分鐘低強度 (35% VO_2max) 運動，而安靜休息組則採坐姿休息。於運動前、運動後立即、運動後 20 分鐘及運動後 2 小時進行採血，以便分析血液中淋巴細胞數目和 T 淋巴亞群、B 細胞及 NK 細胞比例。所得資料以混合設計二因子變異數進行分析。結果：動態恢復組

	運動前	運動後	運動後20分鐘	運動後120分鐘
AR				
RR				

依變項:淋巴細胞數目、T淋巴細胞亞群、B細胞及NK細胞比例

Error #10 Correlation v.s. Causation

- Shoe size v.s. scores on a reading exam
- Ice cream sales v.s. drowned visitors
- Sleep with light on v.s. myopia
- lurking variable
- How can you claim causality?

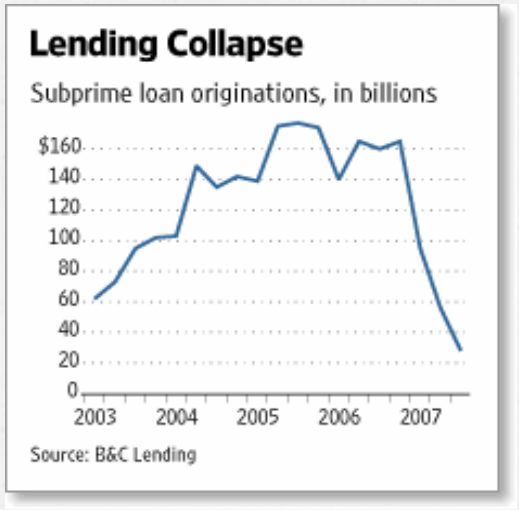
例

- A study published in 2010 showed that city dwellers have a 21 % risk of developing anxiety disorders and a 39% higher risk of developing mood disorders than those who live in the country. A follow-up study published in 2011 used brain scans of city dwellers and country dweller. The brain scans showed very different levels of activity in stress centers of the brain, with the urban dwellers having greater brain activity than rural dwellers in areas that react to stress.
- So, living in a city increases a person's likelihood of developing a anxiety or mood disorder?

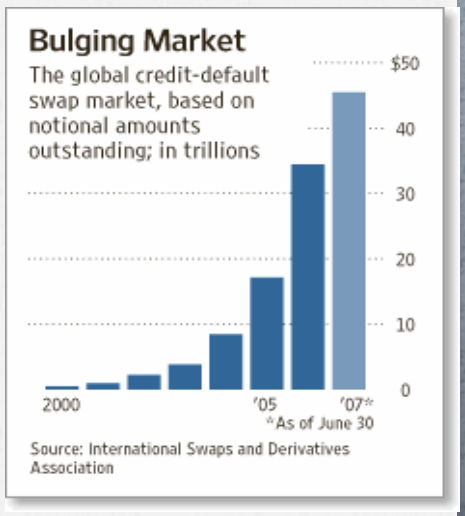
Error #11 Extrapolation

- Extrapolating to a different population
  - Sampling issue
- Extrapolating to an undiscovered data range

US Home Price Index

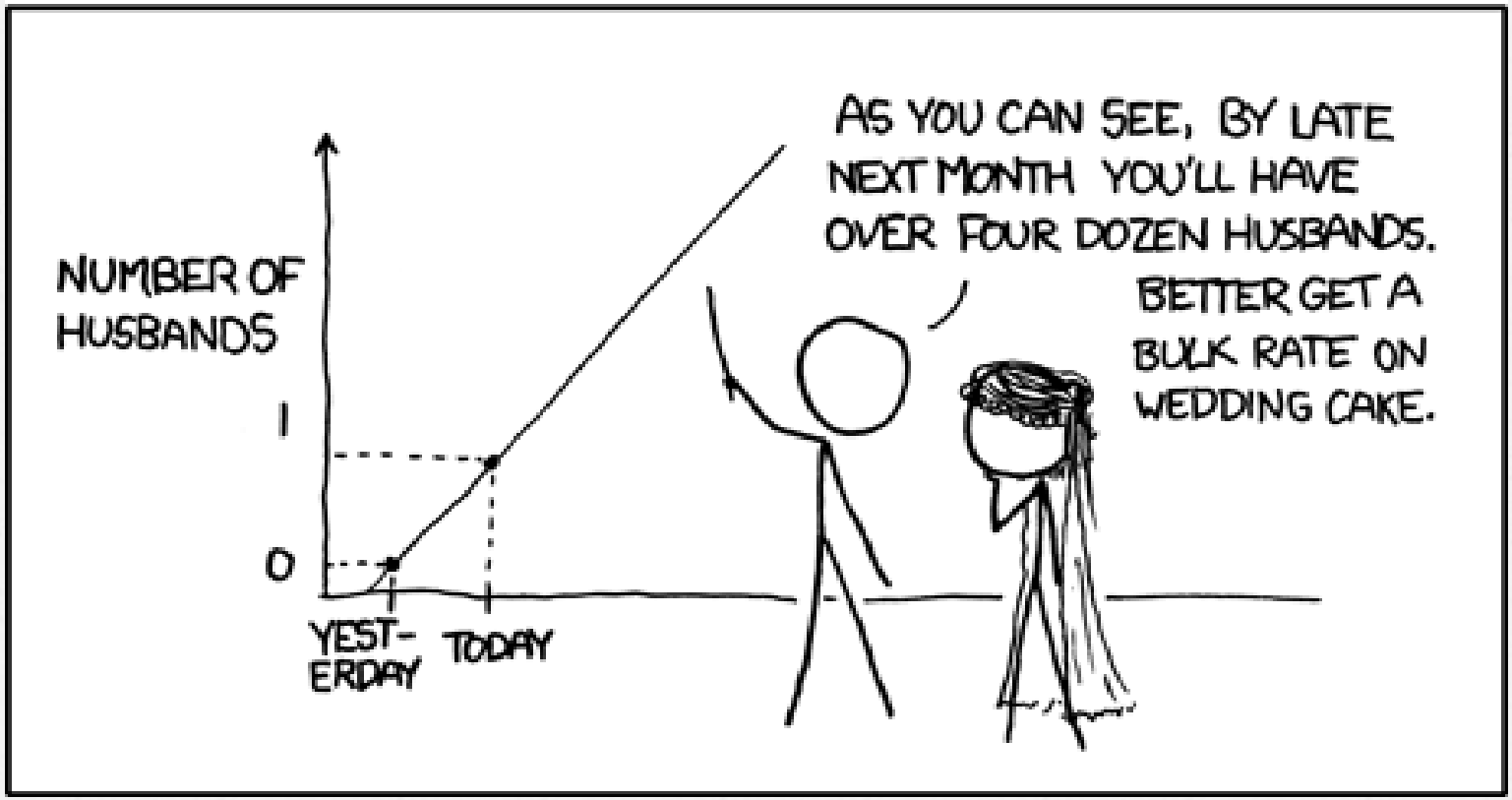


Subprime loan origination



Exponential Market

MY HOBBY: EXTRAPOLATING



Error #12 Representativeness

- The 1936 Literary Digest Poll
  - A sample size of 2.4 million
  - Over ten million voters being asked, including telephone directories, club membership lists, magazine subscribers.
  - Predicted Alfred Landon(Republican) : Franklin Roosevelt(Democratic)=57:43
  - Results: Alfred Landon: Franklin Roosevelt=38:62
  - Selection bias & nonresponse bias

Error #13 Analyzing your data with inappropriate procedure.

- How do I know if it's a inappropriate procedure?
- Fail to validate the test assumptions
- Assumption of "LINEAR" Regression
- Assumption of ANOVA
- Consequence for violation of assumptions

Ten Ways to Cheat on Statistical Tests
BMJ 1997 315:422-5

1. Throw all data into a computer and report as significant any relation where $P<0.05$
2. If baseline differences between the groups favor the intervention group, remember not to adjust for them
3. Do not test for normal distribution. If you do, you might get stuck with non-itemmetric tests.
4. Ignore all drop outs and non-responders, so the analysis only concerns subjects who fully complied with treatment

5. Assume that you can calculate "*r* value" on set of data against another and that a "significant" *r* value proves causation
6. If outliers are messing up calculations, just rub them out. But if outliers are helping the case, even if they seem to be spurious results, leave them in
7. If the confidence intervals overlap zero difference between the groups, leave them out of report. Better still, mention them briefly in the text but don't draw them in on the graph—and ignore them when drawing your conclusions

8. If the difference between two groups becomes significant 4.5 months into a 6 month trial, stop the trial and start writing up. Alternatively, if at 6 months the results are "nearly significant,“ extend the trial for another 3 weeks
9. If results prove uninteresting, see if any particular subgroups behaved differently
10. If analyzing data the way you plan to does not give the result you wanted, run the figures through a selection of other tests

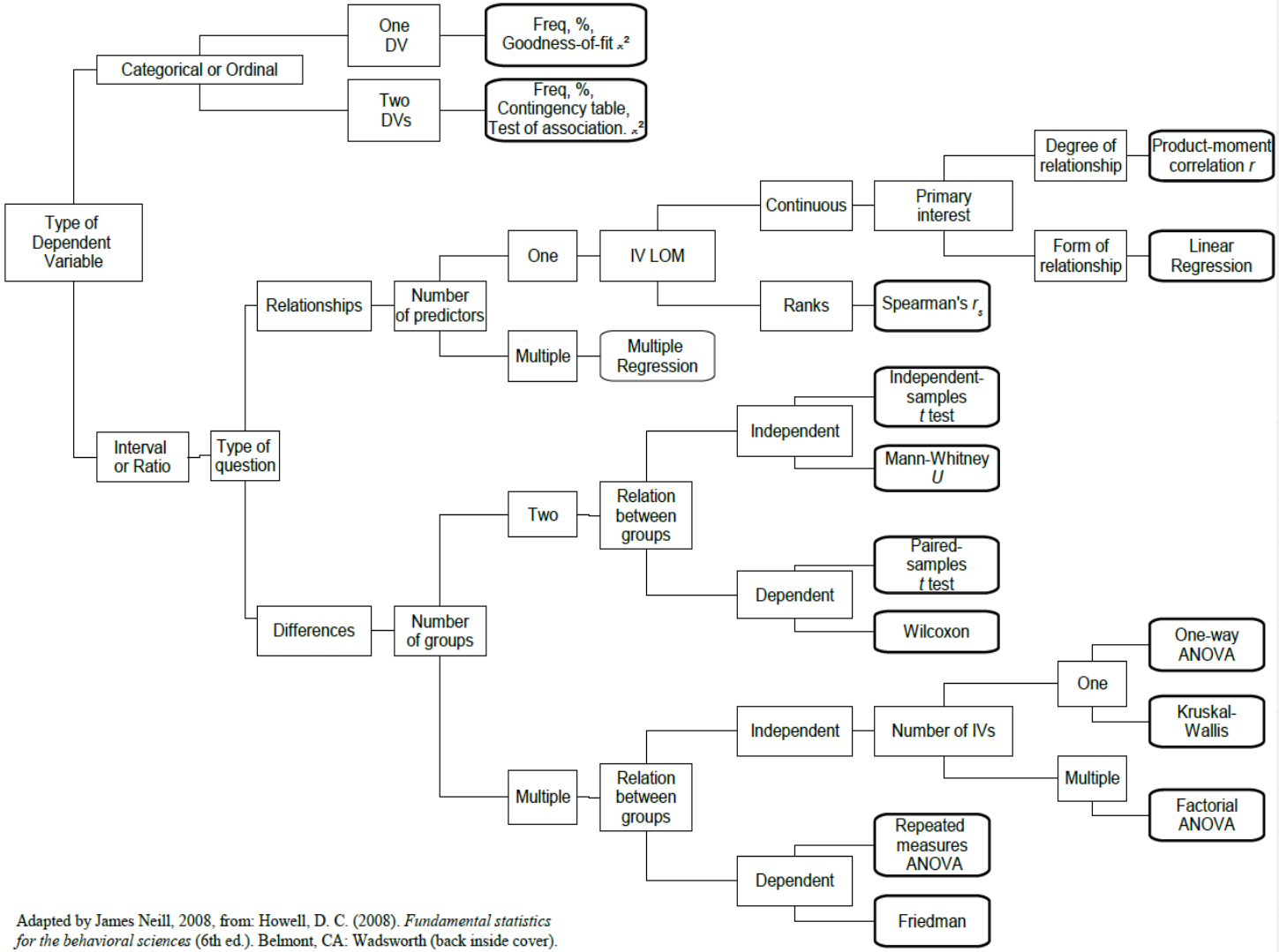
Biochemia medica

- *Table 1. The frequency of statistical errors in manuscripts submitted to Biochemia Medica during 2006-2009. Errors are*

Error	Error rate N (proportion)
Power analysis not provided	55/55 (1.0)
Incorrect use of statistical test for comparing three or more groups for differences	21/28 (0.75)
Incorrect presentation of P value	36/54 (0.66)
Incorrect choice of the statistical test	34/55 (0.62)
Incorrect interpretation of correlation analysis	11/20 (0.55)
Incorrect use or presentation of descriptive analysis	19/55 (0.35)
Incorrect interpretation of P value	12/54 (0.22)

NATURE statistical checklist

- **Type and applicability of test used**
  - Name of tests applied are clearly stated
  - Data meets all assumptions of tests applied
- **Details about the test**
  - Sample size calculation (or justification) is given
  - Alpha level is given for all statistical tests
  - Actual *P* values are given for primary analyses
- **Descriptive statistics summary**
  - A clearly labeled measure of center (e.g. mean or median) is given
  - All numbers following a ± sign are identified as standard errors (s.e.m.) or standard deviations (s.d.)
- **Within individual graphs**
  - Distortions
  - Clear labelling



Adapted by James Neill, 2008, from: Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Wadsworth (back inside cover).

Selected Reference

- *Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles*
- *Common Statistical Errors Even YOU Can Find*
- *Reviewer's quick guide to common statistical errors in scientific papers*
- *http://en.wikipedia.org/wiki/Misleading_graph*
- *台北醫學大學生物統計e報第一期*

Thanks for listening