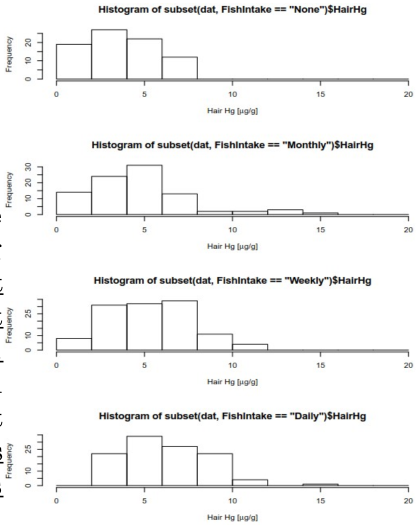
Evidence based healthcare special lecture I Final mini-test   18th July 2018, 18:30-20:00, D101
(Answer)

1. Please specify the wrong points in explanation and/or method to analyze and suggest how to improve it (if no wrong point, answer so) for each issues underlined below.

(1) In R island with population size of about 4,000 in a developing country. When we see there with Google Earth, almost equal-sized 40 villages are scattered. Recently the people with high fish and whale intake are reported to suffer from neural damage symptoms. Due to the possibility of mercury poisoning, (A) 4 villages were randomly selected from 40 villages, and in those 4 villages, all 400 residents were recruited to join the survey. They were interviewed about the frequency of fish/whale intake, such as [1] none or rarely, [2] monthly, [3] weekly, [4] daily, and the hair mercury concentrations of them were measured. The results (of which raw data are available from http://minato.sip21c.org/fish-Hg-2018.txt; variables are PID as personal ID number, HairHg as hair mercury concentration, HighHairHg is 1 if HairHg>=5, otherwise 0, and FishIL is one of 4 categories shown above) were summarized below.

Eating fish or whale	N	Median	Mean ± SD	High Hg
frequency		(Hg μg/g hair)	(Hg μg/g hair)	≥5 μg/g hair
1. None	80	3.44	3.55 ± 2.02	27
2. Monthly	90	4.34	4.72 ± 2.89	36
3. Weekly	120	5.45	5.36 ± 2.30	69
4. Daily	110	5.97	6.20 ± 2.37	72



There are two approaches to analyze the relationships between fish/whale intake and mercury exposure. First, the independence between High Hg and eating fish/whale frequency can be analyzed. (B) Fisher's exact test resulted in $p = 1.34 \times 10^{-5}$ and the null hypothesis was rejected, so that the relationship between the two variables is statistically significant. Second, the effect of fish/whale intake on hair mercury concentration can be analyzed. Welch's one-way ANOVA resulted in F-value of 24.273, first d.f. of 3, second d.f. of 210.39, and p<0.001. (C) Then pairwise comparisons of hair Hg levels between all pairs among 4 fish/whale intake frequencies can be conducted by repeated use of Welch's t-test.

(A) Correct. In such situation, cluster sampling is appropriate.

(B) Correct

(C) Due to the problem of multiple comparison causing excess alpha error, we must not repeat t-test.

For this purpose, adjustment for multiple comparison by Holm or FDR method, or Tukey's HSD is needed. By any of those, except Monthly and Weekly, all pairwise differences are significant (p<0.05).

(2) Based on the result of large scale cohort study, brown rice intake is proved to significantly reduce the body weight and white rice intake is proved to significantly increase the body weight. Then 5 obese patients changed their stable food from white rice to brown rice. The changes of body weight (kg) between the 2 timings (before intervention and after 6 months) were 70→65, 140→135, 95→85, 95→90, 85→80. The result of paired t-test was p=0.0008, so that brown rice intake can be judged to have a significant weight reduction effect as the result the intervention study.

* There are 2 level answers to this question. (Level 1) p=0.003883, but the conclusion is same. (Level 2) Meals are not only composed of rice. Forced change of staple food may lead to change the amount of meals, menus of side-dishes, and those may be true cause of weight loss. Thus the design is inappropriate.

(3) The gold standard method A can measure the concentration of biochemical marker for disease X, where the concentration exceeds a specific threshold value. A cheaper and more rapid new method B was developed. Validity of B can be confirmed by showing the fact that the Pearson's correlation coefficient between the measurements by A and B for the sufficient number of X patients and healthy volunteers is more than 0.8 and that is statistically significant.

Even highly correlated, absolute levels may differ, or systematically biased (eg. in lower ranges, A<B but in higher ranges, A>B). To check such possibility, Bland-Altman plot is needed.

(4) The 44 chronic hepatitis patients were randomly divided into 2 groups. Treatment group was treated by prednisolone, the other (control group) was just observed. At the end of the study, 11 patients lived in treatment group, 6 lived in control group, but the result of Fisher's exact test was not significant (p=0.215). The months until death or censoring (lived at the end of the study) were recorded in http://minato.sip21c.org/hepatitis-2018.txt as **time**, with **flag** (1 if died, 0 if still lived) and **group** (1 for treatment group, 2 for control group). For all patients (ignoring alive/dead), the mean survival months (109.5 months in treatment group and 64.7 months in control group) were compared by Welch's t-test, then p=0.013, so that prednisolone has a significant effect of lengthen survival for chronic hepatitis patients.

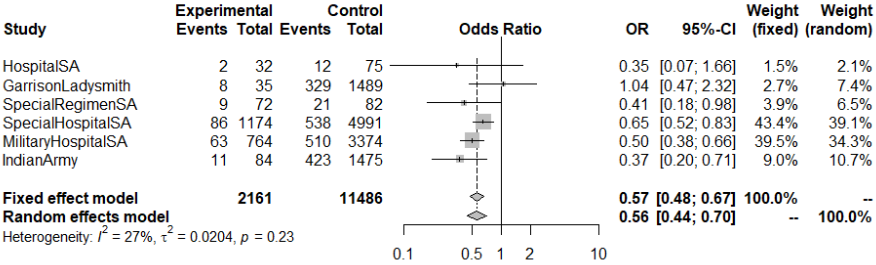
* There is 2 levels of answers. (Level 1) Appropriate (test statistics and conclusion is not wrong).

(Level 2) Survivors surely lived longer than the recorded time, so that this analysis is inappropriate. Insted, log-rank test should be used to test the difference of survival considering censoring, p=0.0309. Conclusion remains unchanged.

2. Please explain the incidence rate of a disease. What type of study design is needed has to be included.

Basically cohort study is needed. During observation, the total number of disease occurrence divided by person-times. It means the rate of disease occurrence per unit time.

3. Sir Wright developed typhus vaccine and it was used in 6 studies conducted around 1900. In 1904, Karl Pearson summarized the results by simply taking average of tetrachoric correlation coefficients of 6 studies. The result was 0.193, which is judged as not enough strong to recommend as vaccine by Karl Pearson. However, if we consider the death as outcome, and conduct the meta-analysis of odds ratios, integrated odds ratio is 0.57 (95%CI, 0.48-0.67), which means the vaccine having the significant effect to reduce mortality. Please explain why Pearson made a wrong judge. Why we must not take simple averages of correlation coefficients?



Study	Experimental Events	Experimental Total	Control Events	Control Total	OR	95%-CI	Weight (fixed)	Weight (random)
HospitalSA	2	32	12	75	0.35	[0.07; 1.66]	1.5%	2.1%
GarrisonLadysmith	8	35	329	1489	1.04	[0.47; 2.32]	2.7%	7.4%
SpecialRegimenSA	9	72	21	82	0.41	[0.18; 0.98]	3.9%	6.5%
SpecialHospitalSA	86	1174	538	4991	0.65	[0.52; 0.83]	43.4%	39.1%
MilitaryHospitalSA	63	764	510	3374	0.50	[0.38; 0.66]	39.5%	34.3%
IndianArmy	11	84	423	1475	0.37	[0.20; 0.71]	9.0%	10.7%
Fixed effect model		**2161**		**11486**	**0.57**	**[0.48; 0.67]**	**100.0%**	**--**
Random effects model					**0.56**	**[0.44; 0.70]**	**--**	**100.0%**

Heterogeneity: $I^2 = 27\%$, $\tau^2 = 0.0204$, $p = 0.23$

(This question is difficult) Pearson ignored the differences in sample sizes of 6 studies. One study with relatively small sample size of experimental group (GarrisonLadysmith) showed very low correlation, which biased the simple mean.

4. Whether the ability of flash memory is improved by glucose candy or not was investigated for 10 healthy volunteers. The result is shown below. Please test whether glucose candy improves flash memory or not. P-value is needed. You can use computer software or calculator, but manual calculation is possible if you use 97.5% point of t-distribution with d.f. 9 is 2.262 and either of √2=1.414, √3=1.732, or √5=2.236. ***** Note: The values are slightly different from Japanese version *****

Scores before glucose candy	8	6	3	7	6	7	5	8	7	6
Scores after glucose candy	9	7	5	8	7	9	6	9	8	7

Difference 1 1 2 1 1 2 1 1 1 1 mean 1.2 S.E.√(1.6/9)/√10=0.4/3

Then, t=1.2/(0.4/3)=9 > 2.262. Thus the difference between taking candy is statistically significant.

`t.test(c(9,7,5,8,7,9,6,9,8,7),c(8,6,3,7,6,7,5,8,7,6),paired=TRUE)` gives the same result, where $p = 8.5 \times 10^{-6}$

5. The RCT (Randomized Controlled Trial) to test new stretch method to improve body flexibility is to be conducted. Outcome measure is the change of standing-posture body anteflexion in centimeter. Based on previous studies, the conventional stretch method increased 3 cm (standard deviation 2 cm) of standing posture body anteflexion measurement. If the new stretch method can increase more than 4cm (as the difference from the conventional method, more than 1 cm) of standing posture body anteflexion, it can be judged as clinically valuable. Please calculate needed sample size for this RCT for 2-tailed t-test with 5% significance level and 80% power, and assuming the same size of 2 groups.

The result of `power.t.test(delta=1, sd=2, sig.level=0.05, power=0.8)` is n=63.76…, then sample size is 64 for each group (EZR, 63 each, PS, 64 each). ****(Level 2 answer)** This test assess the very short-term effect of intervention, the same subject can be used as control group after enough washout period. If paired comparison is done, using PS, the sample size is 33 (using EZR, 34).