Author's response to reviews

Title: Smoking differences between employees in faculties of the University of Tartu, Estonia, and changes during the country's transition

Authors:

Rauno Heikkinen (rauno.heikkinen@oulu.fi)
Jana Kivastik (jana.kivastik@ut.ee)
Peet-Henn Kingisepp (peet-henn.kingisepp@ut.ee)
Simo Näyhä (simo.nayha@oulu.fi)

Version: 3 Date: 9 February 2011

Author's response to reviews: see over
Dear Sir/Madam

Please find enclosed the revised version of the above manuscript which we have corrected according to your referee’s advice. We appreciated her criticism which did much to improve the quality of the paper. The referee was mainly concerned about statistical tests and about what she calls statistical significance. In our detailed response (attached) we explain that we have tested our results the way she requires, and we have done even more than that. We have also modified the text to make everything clear to the reader.

Please note that the order of the authors’ names has been changed. There will be a series of papers on this topic and we thought it is appropriate to circulate the authors’ names. This change has been approved by all authors.

The revisions are marked in blue.

Looking forward to your answer,

Yours sincerely,

Simo Näyhä
Professor, M.D., Ph.D.
Department of Health Sciences
University of Oulu
P.O. Box 5000
90014 University of Oulu
Finland

Phone +358-8-5375 671
Mobile +358-40-5646 159
Fax +358-8-5375 661

simo.nayha@oulu.fi
Reviewer's report
Title: Smoking differences between employees in faculties of the University of Tartu, Estonia, and changes during the country's transition

Version: 2 Date: 29 November 2010
Reviewer: Frida Eek

Reviewer's report:

Major compulsory review

Regarding my request for results to be presented and discussed in relation to their significance (or non-significance): I must say I’m a little bit puzzled by your response. I did not require p-values; if you prefer to use p-values or CI’s is up to you (actually, I also think that CI’s are preferable, since they provide more information than a single p-value). However, no matter whether you use CI or p-value: the aim is to judge whether your results are “significant” or not, i.e., how likely is it that the difference you observe (your point estimate) reflects a “true” difference, or could be a difference that happened just by chance. Since you haven’t included the total target population, and hence would be interested in making statistical inference to the population (in this case the university) this is a highly relevant question. Also, you set up hypotheses and, despite ‘the growing criticism against testing’ you do perform analytical tests, and you do present CI’s. So then, why are you “reluctant to use significance testing”!? Quote: We were not looking for “significances”. We were looking for difference estimates accurate enough to differ from the baseline with a reasonable certainty.”

Isn’t that exactly what significance testing is about!? Again, whether you use p-values or CI doesn’t matter, what you do evaluate is if you can, or can not, say that there is a difference with a “reasonable certainty”; namely 95% certainty, which is the generally accepted limit for “reasonable certainty”, and which you also appear to accept by presenting 95% CI’s.

Answer
Thank you for your criticism. We are sorry if our reply caused misunderstanding. The problem seems to be in how terminology is used. In our vocabulary, “significance” refers to the Fisher and Neyman-Pearson approaches to assess random variation, i.e., p values and testing hypotheses $H_0/H_1$, respectively (see e.g. Rothman and Greenland 1998, chapter 12). As confidence intervals (CI) are also based on hypothesis testing (in the “frequentist” sense), the referee is right in saying that in one way we did hypothesis testing. However, the CIs are more than just an alternative way to express statistical significance: they give the precision of parameter estimates which p values and significance do not. There is a major difference between testing and CIs, to the benefit of the latter. Poole (2001) gives good examples of instances where p values are small (“significant”) but random variation is very large, and vice versa. In cases where the difference from the baseline is “not significant at 5% level”, the data may still be more compatible with a difference than with no difference (we use this phrasing in the text where appropriate). In fact, non-significant results may be compatible with substantial effects (see e.g. Rothman and Greenland 1998, especially pages 191-194). In such cases, a reliance on statistical significance alone could lead to serious misinterpretations.
Quantifying the random variation is really the “highly relevant question” the referee is asking for, and this is exactly what we have done, but we have done more than merely “analytical tests”. While doing this, we follow the guidelines given by leading medical journals and recommendations. However, in the revised version, we have also referred to “significance” where appropriate, in the hope that this would help readers not accustomed to think in terms of CIs.

I suggest that, if you are reluctant to significance testing and not at all looking for statistically significant results or to make statistical inference, you may consider presenting the results as a descriptive report rather than as a research article, and then NOT present CI’s, or perform the analytical tests. If you do want to test hypotheses and perform analytical tests, you should use the information given and presented, and discuss the result by including the issue of statistical significance. Now, it is a mixture where you present CI’s but totally ignore the interpretation of them, or using vague expressions as “imprecise” or “tended to”. Also, expressions like ‘Women’s smoking declined in only four faculties “with any certainty”’ sound a little odd. Why not explicitly use the conventional probability levels?

You may well discuss the lack of significant changes or differences in relation to lack of power/small samle sizes, but you still need to be clear about the difference between significant and non-significant results (although I am well aware that the 5% limit is not equal to the difference between true or false). Again, I am NOT asking for p-values, you may well use the presented CI’s!

**Answer**

As explained above, we include statistical analysis of random variation, although we use statistical estimation (CIs) instead of “testing” (CIs include “significance”).

While comparing the 95% CIs of parameter estimates with the baseline, we do apply the “conventional probability level” (5%), but we deliberately avoid using phrasings which only give “yes/no” or “true/false” answers (this is what tests do), rather we use parameter estimates which express the size of the difference together with its precision (CI). As Stang et al. (2010) point out, we should move away from the dichotomy of accepting / rejecting statistical hypotheses (significant vs non-significant) to assessing how much the point estimate is affected by change.

Once again, we have certainly done what the referee is asking for, and we have done more than that. To make this clear, “significance” is now mentioned where needed, along with considerations of CI.

Do I understand the tables right, that some of the point estimates consist of one single smoker in the group? Eg Exercise and sport science 2003, 4% among 25 persons? Or women, 3% among 29? Or 4% among 28? (survey II)? And then you put these percentage point estimates into a regression model, where you find a significant association? I think that the major uncertainty in the point estimates that is put in the regression model at least needs to be discussed! Or have I misunderstood the method?

**Answer**

It was not the crude empirical percentages (e.g. Sports 4%) which were put into binomial regression but the individual “successes” vs “failures” (smoker/non-smoker, coded as 1/0) which follow the Bernoulli distribution. And in Figure 3, the question is about predicted (and adjusted) differences vs the medical faculty. Thus
e.g. in survey II, the crude difference Sports – Med was 14% – 0% = 14%, but when adjusted for age and occupation it was 11% (from Fig. 3, survey I). It is true that major uncertainty exists regarding the smallest faculties, but appropriate caveats are given to the reader in Discussion (4. paragraph), and we also suggest that pooling of data across other similar communities would be useful.

Additional comments:

I don’t think that your second hypothesis is possible to confirm or reject in an objective way, by the current methods. What does “similar enough to justify a conclusion that…” actually mean?

**Answer**

We agree that the referee’s concern is justified. We have merged the hypotheses 1) and 2) to a single hypothesis: it is just a question of whether there is any degree of similarity in the between-faculty differences immediately after the disruption of communism and one decade later. A further difficulty in testing this hypothesis could be that we compared two cross-sectional surveys and were unable to carry out an actual follow-up on an individual basis. However, this method is likely to work at aggregate level, i.e. at the level of faculties.

Minor essential revisions

Survey: First you write that questionnaires were identical, then you state that education was asked for only in the first survey?

**Answer**

Our expression was slightly inaccurate. Only the questions on smoking, occupation and personal details were identical. The text has been revised.

Results: response rates etc belongs, in my opinion, to methods/description of participants/study sample.

**Answer**

The referee’s suggestion is very good. The paragraph describing the participants has been moved to the Methods section.

P 8, smoking difference between work places

“the adjusted figures for daily smoking in the faculties…” Isn’t it rather “adjusted figures for difference in daily smoking between medical faculty and other faculties…”?

**Answer**

The referee is right. We have revised the text according to her advice.

Discretionary revisions

Data analysis: Why was age introduced as 3rd degree polynomial?
Answer
Age was asked to an accuracy of one year, and we treated age as a continuous variable to retain statistical power. Since the age trend may not be linear, it was necessary to include higher-degree terms. The first degree term (x) reflects the overall increase of smoking by age, the second degree term (x^2) allows for a monotonous curvilinear change and the third degree term (x^3) makes the regression still more flexible (this was needed to allow for non-monotonous changes at higher ages).

Level of interest: An article of importance in its field

Quality of written English: Needs some language corrections before being Published

Answer
The language (including the revisions) was revised by Mr. Malcolm Hicks, MA (Cantab).

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests

We thank the referees for their useful comments which did much to improve the quality of our paper.

References
Poole C: Low p-values or narrow confidence intervals: which are more durable? Epidemiology 2001; 12(3): 291-294.