Exam: Principles of statistical data analysis

S. Vansteelandt

Academic year 2009-2010

---

1. Answer the following questions and motivate your answer clearly.

   (a) In a scientific paper, a one-way anova test is performed. The following results are reported: $F = 4.1, df_1 = 3, df_2 = 36$. How many groups were compared and how many data were analyzed? Make sure your reasoning is clear!

   .....

   .....

   .....

   .....

   (b) What would be the conclusion of the one-way anova test at the 5% significance level?

   .....

   .....

   .....

   (c) A scientific paper describes the results of a randomized comparison of 2 blood pressure lowering medications on diastolic blood pressure. The following results are published on the basis of the one-way anova test: average difference 10.6 mmHg (SE 4.9 mmHg) and within MSE 600.25. If you know that both treatment groups were equally large, what was then the number of data per group?

   .....

   .....

   .....

   .....

   .....

   .....

   (d) Calculate a 95% confidence interval for the average difference in diastolic blood pressure between both medications.

   .....

   .....

   .....

   .....

   .....

   .....

(e) Would the one-way anova test be appropriate for the comparison of 2 treatment groups in a randomized experiment?

(f) In a scientific paper, authors write: 'Embryo transfer in combination with acupuncture is associated with a significant and clinically relevant improvement in clinical pregnancy (relative risk 1.65, 95% confidence interval 1.27 tot 2.14)'. Is it indeed possible to confirm on the basis of this information that the improvement is statistically significant? If yes, why? If no, why not and what additional information would you need?

(g) Is it indeed possible to confirm on the basis of this information that the improvement is clinically / scientifically significant? If yes, why? If no, why not and what additional information would you need?

(h) Explain clearly whether the following definition of a p-value is correct and why/why not: 'probability that, if you reject the null hypothesis, that this is an error'.

(i) Explain clearly whether the following definition of the standard error of an estimator is correct and why/why not: 'standard deviation of the observations divided by the square root of the number of observations'.

.....

.....

.....

.....

.....

(j) In July 2008, London was shocked by the BBC report on four fatal stabbings in 1 day. How likely is it to have four fatal stabbings in 1 day, knowing that - based on historic data - there are on average 68 fatal stabbings per year in London?

.....

.....

.....

.....

.....

(k) The BBC further reported that the number of murders between the start of 2008 and the end of July 2008 had reached 90. Assuming that - based on historic data - the number of murders in London is 167 per year and further assuming that the occurrence of murders is constant through time, explain how you could judge whether having 90 murders after 7/12's of a year is exceptional[1].

.....

.....

.....

.....

.....

.....

.....

.....

.....

2. The enclosed paper 'Commercial honey bees (Apis mellifera) reduce the fecundity of an Australian native bee (Hylaeus alcyoneus)' by Paini and Roberts reports the results of an experiment run over two years on the impact of commercial honey bees on the fecundity of a solitary native bee, Hylaeus alcyoneus. Registered apiary sites were used as treatment sites (with honey bees) while control sites (without honey bees) were interspersed between. Please answer the following questions. There is no need to read the article.

[1] Take into account the random nature of the data, and thus the uncertainty in the data!

(a) In the statistical analysis section, the authors write 'Heterogeneity[2] increases the risk of a type I error which is only relevant when a significant difference between treatment and control is found (Underwood, 1999). Therefore the ANOVA was still performed on data that failed the test for homogeneity[3] and if there was no difference between treatment and control ($\alpha = 0.05$), the result was accepted.' Do you agree with this strategy? What is your opinion?

.....

.....

.....

.....

.....

.....

(b) In Section 3.1, the authors argue that 'The honey bee density data from week 24 (15 August 2000) failed Cochran's test for homogeneity[4], even after transformation. The heterogeneity was caused by a large number of zero values in control sites invalidating the use of analysis of variance models.' Do you see alternative possibilities for testing? Explain.

.....

.....

.....

.....

.....

.....

(c) Suppose that in view of the previous problem, the authors make the assumption that the bee density data are Poisson distributed with mean μ in the treatment group and ν in the control group. For such data, the statistic

$$T_n = 2n\left(\sqrt{\bar{X}_n} - \sqrt{\bar{Y}_n}\right)^2,$$

with $\bar{X}_n$ and $\bar{Y}_n$ the mean densities in the treatment and control group, respectively, is *asymptotically chisquare distributed with 1 degree of freedom*. Explain in simple words what the phrase in italics refers to.

.....

.....

.....

.....

.....

[2] i.e., heteroscedasticity.
[3] i.e., homoscedasticity or constancy of variance.
[4] i.e., an anlog to the Levene's test.

(d) On the basis of the previous result, show how you would perform a test of the hypothesis that $\mu = \nu$.

(e) In Figure 3, the authors have used ANOVA tests to compare the mean ShannonWeiner index between the treatment sites and control sites, for each of the displayed weeks. What would the statement 'Week 12 (1999) was not significant after sequential Bonferroni adjustment.' mean and what would be the reason for doing Bonferroni adjustment?

(f) Do you see a way for comparing the mean ShannonWeiner index between the treatment and control sites in a way that avoids Bonferroni adjustment? If so, would your suggested approach be advantageous to that of the authors? Explain.

(g) In Section 3.4.1, the authors write 'The mean number of nests produced per measurement interval over both seasons was 23% less in treatment sites than control sites (treatment = 0.53 ± 0.20 SE; control = 0.85 ± 0.23 SE).' The quantitative information in this sentence could be made much more informative about the comparison of treatment and control sites by providing a 95% confidence interval for the estimate '23%'. How could you obtain a 95% confidence interval for this estimate (or for a similar ratio of central location measures of the number of nests in treatment versus control sites)? Explain.

(h) In Table 2, the authors report the results of a power calculation to detect differences in the number of eggs per nest between control and treatment sites. Explain how you could reproduce their results.

(i) Suppose that the authors considered the use of nonparametric tests. Then give detailed R code (or pseudo code if you are not familiar with the R code) showing how would you calculate the power of such a test via simulation, assuming that the data follow a Poisson distribution[5].

Good luck!

[5] `rpois` is the R function for generating Poisson data.