HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

Project co-financed by the European Union under the European Social Fund

The Interdisciplinary Ph.D. Studies: *SOCIETY – ENVIRONMENT – TECHNOLOGY*

Research design, experiments and GLM

by Paweł Koteja

EXAMPLES OF EXAM QUESTIONS

BASE:

1. Suppose you have a set of pairs of measurements of variables (characters, traits) Y and X . For example, in biological research it could be body length and leg length in a group of animals; in social sciences it could be the level of happiness (measured on a quantitative scale) and monthly income in a group of people; in physical/technical sciences it could be compression resistance and density measured in a sample of assorted rocks. Assume that a relation between Y and X is linear (if at all present). Suppose that the analysis of the relation between Y and X should serve two distinct purposes:
A) fitting a line describing a linear functional interrelation between the Y and X variable;
B) fitting a line that allows the best prediction of the Y value based X value.
Draw a graph with a few hypothetical data points and two hypothetical lines representing the two linear relationships, and show graphically and explain shortly the methods appropriate for fitting the two lines.

A complete answer to the question is in my lecture presentation and in chapter 5.3 (especially 5.3.2 and 5.3.14) of Quinn and Keough 2002 handbook.

2. Imagine a two-factor experiment performed in a "factorial" design: Factor A has 3 "levels" (groups) and factor B has 4 "levels" (groups). In each of the 12 subgroups (combinations of factors A and B) there are 5 observations of a dependent variable Y.

1.1. Present a linear ANOVA model relevant for this experimental design. Explain, which part of the model describes the expected value and which represents unexplained part of variation.
1.2. Calculate the number of degrees of freedom for each of the terms in the model.
1.3. Explain how would you perform the test of significance of each of the effects in the model, i.e. explain which "mean squares" (MS) should be used to calculate appropriate *F* statistics, when:
a) both A and B are fixed factors;
b) A is fixed and B is random (assume a restricted version of the mixed ANOVA);
c) both A and B are random.
Note: the answers should have a form of "$F_G = MS_G/MS_E$", where G indicates appropriate effect, and E indicates appropriate error term.
1.4. What are the assumptions of an ANOVA applied to results from such an experiment?

A complete answer to the question is in chapters 9.2 (factorial designs, two-way ANOVA) and 8.3 (assumptions of ANOVA) of Quinn and Keough 2002 handbook.

3. Imagine a 3-level experiment with a hierarchical (nested) design: The highest level fixed factor A has 2 groups, the second level random factor B has 3 groups nested within each group of A, and the third level random factor C has 4 groups nested within each group of B. In each of the C groups there are 5 observations of a dependent variable Y.

1.1. Present a linear model relevant for this experimental design. Explain, which part of the model describes expected value and which represents unexplained part of variance.
1.2. Calculate the number of degrees of freedom for each of the terms in the model.
1.3. Explain how would you perform the test of significance of each of the effects in the model, i.e. which "mean squares" should be used to calculate appropriate *F* statistics. Note: the answers should have a form of "$F_G = MS_G/MS_E$", where G indicates appropriate effect (A or B) and E indicates appropriate error term.
1.4. What are the assumptions of an ANOVA applied to results from such an experiment?

A complete answer to the question is in chapters 9.1 (nested designs, three-level) and 8.3 (assumptions of ANOVA) of Quinn and Keough 2002 handbook.

ADVANCED:

4. Suppose you want to learn whether the level of happiness (measured quantitatively, based on answers to a set of questions asked by researcher)

1) differs between inhabitants of villages and inhabitants of small towns, of comparable population size,
2) differs between men and women, and
3) whether the (hypothetical) difference in happiness between inhabitants of villages and cities depends on gender.

4.1. Design a scheme of data sampling suitable for such a research. The data gathering should be done be one researcher in 12 working days. Assume that 12 interviews per day can be obtained from responders living in the same area. Describe verbally very shortly the scheme of the study: how would you organize the work? Do NOT describe how the "happiness" is measured – just assume that you somehow can get the 12 values per day from people in one city or village. If you fill unhappy with measuring happiness, suppose that the problem concerns body mass, and you just need to take body mass of each individual included in the study. Focus of the scheme of the data sampling, i.e., the experimental DESIGN.

4.2. Present statistical model (ANOVA) suitable for analyzing results from this study. Do not forget about additional effects that may have to be included in the model, in addition to those corresponding to the main questions.

4.3. Describe symbols representing effects on the model, explain which of the effects are fixed and which are random, and explain which terms of the models allow to test hypothesis associated with the three main questions asked in the research project.

4.4. Calculate the number of degrees of freedom for each of the terms in the model.

4.5. Explain how would you perform the test of significance of the effects in the model (the answers should have a form of "$F_G = MS_G/MS_E$", where G indicates appropriate effect (A or B) and E indicates appropriate error term.

Hint: think about the faults in the badly designed project presented during the first lecture and avoid the faults. Appropriate designs are described in Quinn and Keough chapter 11. Also, I think that on each of the seminar groups we discussed at least one project, in which a similar design has been applied.

Notes concerning additional sources of information.

Especially for students from social sciences: Information about theoretical background of the analysis of regression and analysis of variance, as well as about factorial, nested, and combined (cross-nested = partly-nested = split-plot) ANOVA designs is also clearly presented in a popular statistical handbook Ferguson and Takane "Statistical Analysis in Psychology and Education" (Polish edition: "Analiza statystyczna w psychologii i pedagogice", PWN, 1997). The only topic missing in this book is distinction between model I and model II regression, and the alternative criteria for fitting a regression line – but this has been explained carefully enough in lecture presentations.

Especially for students from life sciences: theoretical background of the analysis of regression (model I and II) and of ANOVA factorial and nested designs is presented in a simple way in A. Łomnicki "Wprowadzenie do statystyki dla przyrodników" (PWN, use the edition from 2003 or newer; older ones have important errors).

For students representing "hard" physical sciences: unfortunately, I do not have an additional source of information I could offer especially for you, but I have heard the students of mathematics, physics, astronomy, chemistry etc. are the smartest ones in the entire university (or perhaps the entire Universe), so I am confident you will easily understand the texts concerning mathematical methods addressed to biologists or psychologists!

Paweł Koteja