Demo 5: Categorical associations, nonparametric tests, summary of the course

Experimental and Statistical Methods in Biological Sciences I

October 13, 2014

1 Categorical associations and nonparametric tests

In this first part, we will use data from a study that measured the experience of subjective symptoms caused by a cellphone: dizziness, headache, tiredness, itchiness, blushing, and warmth. These variables were measured 3 times each: before the study, when phone is on, and when phone is off.

The data also includes three categorical variables: sex, age group, and rfjarj. Sex is coded as "male" and "female". Age group is coded as 1 = younger than 22, 2 = age 22-24, 3 = older than 25. Rfjarj codes the order of the experiment, with 1 = "cellphone on first", 2 = "cellphone off first".

1.1 Load in and prepare data

Load in cellphone.csv from `http://becs.aalto.fi/~heikkih3/cellphone.csv`. Header information is on the first row.

Examine the data using `summary, head`. Describe what you see.

Check that factors are coded as factors. Correct if necessary.

Check for missing values. Are NAs already coded in the data? Correct if necessary.

1.2 Normality tests

Test for normality in all time points for variables head ache, dizziness, tiredness, and itchiness. Use histograms and explicit tests for normality.

```
# Histograms
layout(matrix(c(1,2,3,4),2,2))
hist(headache1)
```

```
hist(dizziness1)
hist(tiredness1)
hist(itching1)
# Kolmogorov-Smirnov tests
ks.test(headache1, "pnorm", mean=mean(headache1, na.rm=T), sd=sd(headache1, na.rm=T))
# run this for the other variables too
```

1.3 χ^2 tests

Does the sex distribution in our sample represent the sex distribution of the population?

```
table(sex)
chisq.test(table(sex), p=c(0.5,0.5))
```

Are there differences between men and women in age group?

```
table(sex, agegroup)
chisq.test(sex, agegroup)
```

1.4 Two independent samples

Two data samples are independent if they come from distinct populations and the samples do not affect each other. Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

The null hypothesis is that the numeric data from both groups are from identical populations.

Plot a boxplot of headache when phone is on for males and females separately. How do the distributions look like?

```
boxplot(headache_on ~ sex)
```

Compare sex differences in headache when phone is on using independent samples t-test. Do the same with Mann-Whitney-Wilxocon test. Compare results.

```
# repeated samples t-test
t.test(headache_on ~ sex)
# Wilcoxon test
wilcox.test(headache_on ~ sex)
```

Question 1 Compare sex differences in dizziness when phone is off using independent samples t-test. Do the same with Mann-Whitney-Wilxocon test. Compare results. Get the plot too.

1.5 Repeated samples

Two data samples are matched if they come from repeated observations of the same subject. Using the Wilcoxon Signed-Rank Test, we can decide whether the corresponding data population distributions are identical without assuming them to follow the normal distribution.

The null hypothesis is that the two samples are from identical populations.

Examine whether the experienced headache changes when you turn phone on. Compare headache before experiment with headache when phone is on. Run comparison using repeated samples t-test and Wilcoxon test. Compare results.

```
# repeated samples t-test
t.test(headache1, headache_on, paired=T)
# Wilcoxon test
wilcox.test(headache1, headache_on, paired=T)
```

Question 2 Run similar analysis for tiredness.

1.6 Multiple independent samples

A collection of data samples are independent if they come from unrelated populations and the samples do not affect each other. Using the Kruskal-Wallis Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

The null hypothesis is that the numeric data from all groups are from identical populations.

Examine age group differences in experienced tiredness when phone is on. Run comparisons using descriptive statistics, boxplots, independent samples ANOVA and Kruskall- Wallis test.

```
# take table of means:
tapply(tiredness_on, agegroup, mean)
# take boxplot:
boxplot(tiredness_on~agegroup)
# ANOVA:
model.anova <- aov(tiredness_on~agegroup)
summary(model.anova)
# Kruskall-Wallis:
kruskal.test(tiredness_on~agegroup)
```

Question 3 Run similar analysis for headache.

2 Summary dataset

The dataset includes 600 randomly selected cases from the NHANES Epidemiological Follow-up Study (NHEFS), a large ongoing study involving well over 10,000 participants. The purpose of the study is to look at health. In this exercise, you will be focusing on the following variables:

- Neuroticism: a continuous Neuroticism measure
- Extraversion: a continuous Extraversion measure
- Openness: a continuous Openness measure
- CES_D: the Center for Epidemiologic Studies Depression Scale (CES-D), a continuous instrument used to assess the presence of depressive symptoms
- male: a factor indicating gender (0 = female, 1 = male)
- age: a continuous age variable
- dr_heart - dr_hypertension: 5 variables indicating the physician assessed presence of heart disease, cancer, stroke, diabetes, or hypertension (0 = absent, 1 = present)
- marital: factor indicating marital status (1 = married, 2 = widowed, 3 = divorced, 4 = separated, 5 = never married)

2.1 Load in, prepare and describe data

1. Retrieve the data from `http://becs.aalto.fi/~heikkih3/nhefs.csv`, and read it into a variable. The data is in .csv format with the variable names in the first row.
2. Examine the data. Note that the data frame includes a lot of extra variables in addition to the ones you are going to use (listed above). To simplify things, save into a new data frame only the variables you will need later.
3. Data checking:
   (a) Check that factors are coded as factors.
   (b) Find the number of missing values for each variable.
   (c) Replace missing values for Extraversion with mean. I.e., instead of coding the missing values as NAs, replace them with the mean of Extraversion.
4. Creating a new variable

(a) Create a new variable that codes the number of health problems. This should include the sum of the 5 variables indicating the presence of different health issues.

5. Plotting data:

   (a) Generate some preliminary plots of the data showing the distribution of and relationship among personality and depression variables. Tip: use histograms, boxplots and/or scatterplots.

   (b) Make a plot that compares the relationship between Neuroticism and CES-D for males and females. Tip: use xyplot/coplot (demo4).

6. Describing data:

   (a) Show the descriptive statistics for personality and depression variables. Are all the personality and depression variables normally distributed? What are the consequences of the result for future analyses?

   (b) Look at the age distribution. What is the age range in the study? Is age normally distributed?

2.2 Comparing groups

Use nonparametric tests when necessary! Report your results verbally (interpretation) and include test statistics.

1. χ^2 tests

   (a) Are there sex differences in marital status? Show a contingency table and run a χ^2 test.

   (b) Is the sex distribution of the sample representative for the population (assuming that 50% of the population are men and 50% women)?

   (c) Are there sex differences in the number of health problems?

2. T-tests

   (a) Are there sex differences in

      i. the three personality variables?

      ii. the depression variable?

3. ANOVA

   (a) Are there differences between classes of marital status in

i. the three personality variables?

ii. the depression variable?

(b) What about differences between classes of marital status depending on sex? (two-way anova, include interactions)

(c) Think of 4 orthogonal contrasts for the marital variable. Make a table of these contrasts. Describe the contrasts verbally as hypotheses, e.g., "Married individuals will be lower in depression than divorced individuals".

i. Create the contrast coding for the marital status.

ii. Create a model that tests your hypotheses.

2.3 Correlations

1. Show correlation matrix between all the personality and depression variables. Interpret the results.

2.4 Regression

1. Modelling: Neuroticism and depression

   (a) Specify a null hypothesis and experimental hypothesis about Neuroticism and depression.

   (b) Model the effect of Neuroticism on depression. Save the results to model_1.

   (c) Output the regression coefficients for the model, and use these to write the equation for your fitted model predicting CES-D from Neuroticism.

   (d) How much of the variance in CES-D is explained by Neuroticism?

   (e) Create another model that also estimates the effects of age and sex on depression. Save the model as model_2.

   (f) Compare the fit of model_2 to model_1.

2. Modelling: Extraversion

   (a) Take the best model from the previous question and add Extraversion as a predictor. Save the output as model_3.

   (b) Test whether adding Extraversion improved the model fit.

   (c) Create another model (model_4) that tests whether the effect of personality on depression differs for men and women. Examine the model coefficients and test the fit of this model against the previous model.

(d) Is there a sex difference in the effect of Neuroticism and Extraversion on depression?

3. Modelling: Health problems

(a) Starting from your best model from previous question, create a new model (model_5) to test whether the personality depression link is not real and that both are just a reflection of one's present health state.

(b) Create another model (model_6) that predicts depression from number of health problems but does not include the personality variables. Evaluate this model against the previous model.

(c) Formulate your final conclusions regarding the relationship between health, personality, and depression.