Introduction to Statistics
lab 7

Johan A. Elkink
jos.elkink@ucd.ie

24 October 2016

Data

For this lab we will continue to make use of the Irish National Election Study (INES) data from the 2007 general election. See Lab 6 for download and subset instructions.

One-sample t-test

1. Test whether voters on average think Bertie Ahern is more dishonest than honest (**v0480** is less than 5):
   SPSS: `T-TEST TESTVAL 5 /VARIABLE = v0480.`
   R: `t.test(ines$v0480, mu = 5)` [1]
   Stata: `ttest v0480 == 5`
2. Formulate the appropriate null and alternative hypotheses.
3. Evaluate the p-value of the test. What do you conclude?
4. What do you conclude overall from this test, in terms of voters' evaluation of Bertie Ahern's honesty?

[1] You might have to recode the variable first to make sure it is numeric. Safe code would use the `recode()` function in the `car` package, but a shortcut would be: `ahern <- as.integer(ines$v0480) - 1`, followed by `ahern[ines$v0480 == "dont know"] <- NA`. Make sure you test whether it correctly translate the variable: `table(ahern, ines$v0480)`.

Testing for independence

When producing a cross-table of two categorical variables, we can use the χ^2-test to test whether the two variables are independent of each other or not. This will be discussed in more detail in Lecture 8. We will look at whether those that think it doesn't matter who is in government (low political efficacy) also are less likely to vote in the general elections.

1. Recode **v0291b** into a new variable **efficacy**,[2] whereby you combine "strongly disagree" and "disagree" into "disagree", and "strongly agree" and "agree" into "agree" (while "neither agree nor disagree" just remains the same).
2. Recode **v0072** into a new variable **turnout**, whereby you combine the three categories that represent not voting.
3. Produce a cross-table of **turnout** by **efficacy**, with the appropriate percentages.
4. Perform a χ^2-test to test for the independent between the two variables:
   SPSS: `CROSSTABS turnout BY efficacy /STATISTICS = CHISQ.`
   R: `chisq.test(table(efficacy, turnout))`
   Stata: `tab efficacy turnout, chi2`
5. The null hypothesis of a χ^2-test is that there is no dependence between the two variables. A high χ^2 value means the two variables are dependent on each other. Try to interpret the p-value with this in mind. What does the p-value represent? (Figure 1 might help.)
6. Based on the table and the test, what do you substantively conclude about political efficacy and turnout?

Bootstrapping

This is a somewhat advanced topic, beyond the scope of this course, but these days relatively easy to implement practically, and can come in very handy.[3] We normally calculate standard errors by calculating the standard error based on reasonable assumptions about the distribution of a particular statistic across hypothetical samples. For example, we know that for the mean the sampling distribution will approximate the normal distribution as the number of samples increases. For some other statistics, however, we do not know the mathematical distribution and cannot calculate standard errors in the same manner. For example, for the median, this is not straightforwardly done.[4] Bootstrapping is a method that can assist in this kind of situation.

[2] Use the codebook to know the meaning of the variables!

[3] Unfortunately, I am not certain bootstrapping is available under the UCD license to SPSS. If not, you can skip the parts of this exercise that involve the bootstrap.

[4] Which is one reason why, despite the median being less sensitive to outliers than the mean, the mean is usually preferred.

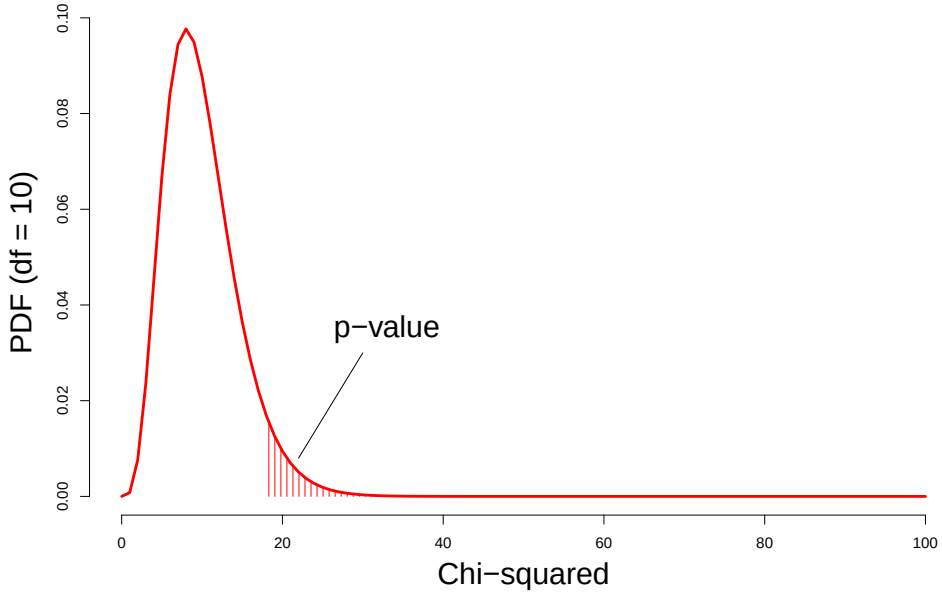

Figure 1: Sampling distribution and the p-value for a χ^2-test

The intuition behind bootstrapping is remarkable but also straightforward: when one takes different samples of the sample at hand, whereby the sample size remains the same, sampling with replacement, then the distribution of the statistic calculated across those samples will approximate the sampling distribution of the statistic. For example, if we have a sample of eight individuals $i = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{bmatrix}'$, we might calculate the median for the following ten bootstrap samples:

Sample 1	1	2	2	2	2	3	5	8
Sample 2	2	3	3	6	6	6	7	8
Sample 3	1	3	3	5	6	7	8	8
Sample 4	1	2	2	3	4	7	7	7
Sample 5	1	3	4	5	5	6	6	7
Sample 6	1	3	3	7	8	8	8	8
Sample 7	1	4	4	4	5	8	8	8
Sample 8	2	4	5	5	6	6	8	8
Sample 9	1	2	2	3	3	3	6	7
Sample 10	1	2	3	6	6	6	7	8

We then calculate the statistic (e.g. the median) for each bootstrap sample, and the distribution of the statistic is an approximation of the sampling distribution.

1. Construct a new variable **age** which is calculated on the basis of **v0906**. Remember that the year of the survey is 2007. SPSS users will need the `COMPUTE` command – cf. how you calculated the logarithmically transformed variables (see Lab 2). Stata users will use `gen`.
2. Calculate the mean and standard error of the mean of **age** (see Lab 6).

3. Calculate the median age.

4. Generate a bootstrapped standard error of the median using 100 bootstrap samples:
   SPSS: `BOOTSTRAP /VARIABLES INPUT = age /CRITERIA NSAMPLES = 100.`[5]
   `DESCRIPTIVES age /STATISTICS = median.`
   R: Install and open the `bootstrap` library, then:
   ```
   res <- bootstrap(age, 100, median, na.rm = TRUE)$thetastar
   sd(res)
   ```
   Stata: `bootstrap r(c_1), reps(100): centile age, centile(50)`[6]

5. Generate a bootstrapped standard error of the mean using 400 bootstrap samples. Compare with the parametrically derived standard error above.

[5] Just `BOOTSTRAP.` should do the trick, but defaults to 1,000 samples. Which is, in fact, more reasonable than 100.

[6] The `centile` function can be used to calculate any percentile, including the 50% percentile, which is the median. See `http://www.stata.com/manuals13/rcentile.pdf`.