GV300: Quantitative Political Analysis

Problem Set 5

Due Thursday, November 22, 9.45am on Faser

1. (30 marks) Read in the data from the 2002 Swedish election ("riksdagsval.csv"). This is a set of returns from different precincts ("Distriktskod") in an election with the number of votes for the different parties (labeled *M, C, FP, KD, S, V, MP, Other*) and the district vote totals (*Total*).

   (a) (5 marks) Create a variable with the last digits of each parties's votes and create a meaningful plot of those variables. What do you see? Would you say the Swedish election of 2002 was a fair election?

   (b) (5 marks) How would you figure out the probability of seeing as few 5s as you do? How would you judge whether the number of 5s you see is an extreme value? Also, why is calculating the exact value of 5s nearly infeasible?

   (c) (20 marks) Look only at the socialist party (variable S) in districts with under 100 voters. What is the probability of seeing so few 4s in a fair election? In `Stata`, the function `comb(n,k)` (which returns the number of ways of choosing k unique items out of n total) may help and `expand.grid` in `R`.

2. (14 marks) Simulate 10000 observations of a variable following a binomial distribution with success probability .2 and 7 trials.

   (a) (6 marks) Write down a verbal definition of probability mass function (PMF), probability density function, and cumulative distribution function (CDF).

   (b) (8 marks) Plot the observed PMF and empirical CDF of the variable you created in (a).

3. (26 marks)

   (a) (6 marks) Generate 50 observations of 10 χ^2 (that's "chi square") distributed variables with 50 degrees of freedom. Create a new variable, which is the average of each of these 10 χ^2 distributed variables and create a histogram of this variable (this will be a sample of 10 observations).

   (b) (10 marks) Repeat the steps in (a) for each combination of 10 and 100 χ^2 distributed variables and for each of 50, 100, 1000, 10000 observations. Comment on what you observe in each histogram you create.

   (c) (10 marks) Finish the `Stata program` or `R function` sketched below so that once you call this program/function it

      i. generates 50 observations of 10 χ^2 distributed variables with 50 degrees of freedom
      ii. creates a new variable, which is the average of each of these 10 χ^2 distributed variables and
      iii. plots a histogram of that new variable.

      Show the output so I can tell whether your program/function works properly. You can write your own program/function if you do not like my setup, it just needs to generate the variables and plot I ask for.

      Programs in `Stata` and functions in `R` will come in very handy in any future statistical work you will be doing, just try, even if you have never written a program/function before.

      `Stata`:

```
program drop _all
program chi2histogram
set obs 50
generate mean = .
forvalue i = 1/10 {
generate chi2Variable‘i’ = rchi2(50)
sumarize chi2Variable‘i’
replace mean = r(mean) in ‘i’
CONTINUE HERE
```

R:

```
chi2histogram <- function(numObs,df,numVar){
mean <- rep(NA, numVar)
    for(i in 1:numVar) {
mean[i] <- mean(rchisq(numObs, df, ncp = 0))
CONTINUE HERE
```

4. (30 marks) Input the data on campaign spending in a U.S.Senatorial election below into your preferred statistical software

District	Incumbent	Money	Vote Share
1	Matt Salmon	362	65
2	Ed Pastor	418	68
3	Jim Kolbe	712	52
4	Bob Stump	346	65
5	John Shadegg	426	68
6	J.D. Hayworth	1839	53

   (a) (10 marks) Let's define the correlation between variables **M**oney and **V**ote Share as

$$corr(M,V) = \frac{cov(M,V)}{\sigma_M \sigma_V}$$

   Also, $cov(M,V) = 1/n \sum_{i=1}^{n}(m_i - \overline{m})(v_i - \overline{v})$. Compute $corr(M,V)$ by hand and show your computations.

   (b) (10 marks) Run a linear regression of V on M, what do the coefficient estimates for intercept and V tell you? Speculate about reasons why we see the relationship we see.

   (c) (10 marks) Can we reject the null hypothesis that there is no relationship between V and M? When you answer this question mention the numerical values representing your test statistic, the level of significance, p-value, and the critical value for this hypothesis test. What is the theoretical distribution of the test statistic in this hypothesis test?