Semester 1 2012 Lecturer: Michael Stewart

Computer Exercise Week 7

Parametric bootstrap (i.e. simulation-from-best-guess) standard errors

This week we illustrate an alternative, computer-based method of obtaining standard errors (or rather approximations to them) which does not require any understanding of theoretical MSEs or large-sample approximations. Recall the data from the tutorial: if gene frequencies are in Hardy-Weinberg equilibrium, the genotypes AA, Aa and aa occur with $Bin(2, \theta)$ probabilities. Suppose that genotypes are determined for a sample of 190 people, with observed frequencies of each as shown below:

AA	Aa	aa
10	68	112

Letting X_i equal the number of a's in the genotype of individual i, we model the *original data* $x_1, x_2, \ldots, x_n$ as values taken by $n = 190$ independent $Bin(2, \theta)$ random variables. The table above is a summary of the original data.

1. Compute three estimates of θ, based on the counts above:
   (a) $\hat{\theta}_1 = \frac{1}{2}\bar{x}$ where $\bar{x}$ is the mean of the *original data*;
   (b) $\hat{\theta}_0 = 1 - \sqrt{n_0/n}$ where n_0 is the observed number of zeroes (i.e. of AA's);
   (c) $\hat{\theta}_2 = \sqrt{n_2/n}$ where n_2 is the observed number of twos (i.e. of aa's).
2. Obtain standard errors for each estimate obtained above (see Lecture 20).
3. Now, suppose we didn't know about theoretical variances or approximate, large-sample mean-squared errors. Can we simply use the computer (and our estimates) to come up with a standard error (i.e. likely size of error) in each case? For any given, fixed value of θ we could use the following procedure:
   (a) simulate a sample of n $Bin(2, \theta)$ pseudo-random numbers;
   (b) obtain an estimate $\hat{p}$ and resulting error $e = \hat{\theta} - \theta$;
   (c) repeat this B times to obtain B simulated errors $e_1, e_2, \ldots, e_B$;
   (d) compute the root-mean-squared (RMS) error: $\sqrt{\frac{1}{B}\sum_{i=1}^{B} e_i^2}$.

   This last quantity is a *simulated approximation* to the theoretical RMS error when the true value is θ. However, we don't know what the true "underlying" value of the binomial success parameter θ is.

   The answer is the use a **best guess**. We have an *estimate* $\hat{\theta}$ which we can use as a best guess; it is not exactly equal to the "true, underlying" binomial parameter, but it should not be far from it. Thus the resulting simulated approximation to the RMS error (i.e. the approximate standard error) we obtain using it should be "close" to the theoretical true RMS error.
4. Here we outline how one would obtain a simulation-from-best-guess standard error (also called a *parametric bootstrap standard error*) for the estimate $\hat{\theta}_1 = \frac{1}{2}\bar{x}$:

```
f=c(10,68,112)            # summary of original counts; frequenices of 0's, 1's and 2's
x=c(0,1,2)
n=sum(f)
m=sum(f*x)/n
m
est=m/2
est
simerr=0
B=10000
for(i in 1:B){
  simdat=rbinom(n,2,est)           # simulated data; actual 0's, 1's and 2's (not freqs)
  simerr[i]=(mean(simdat)/2) - est    # est regarded as "true value" for simulation
}
se=sqrt(mean(simerr^2))
se
```

   Compare the `se` you obtain here with the value obtained in the tutorial exercise.
5. Repeat for the other two estimation methods (last week's exercise may help). Comment.