Data Simulation Using R

Angel An

Department of Biostatistics, Vanderbilt University School of Medicine

Continuing Education Workshop  October, 2007
Simulation is a useful technique in the statistical research.
- Form and develop a hypothesis
- Study and plot a distribution
- Sensitivity and robustness analysis

Three steps involved in data simulation
- Assumption: specify model assumptions and input variables
- Derivation: derive other variables based on your assumptions
- Implementation: use random number generating techniques to simulate the data accordingly
Step 1: Assumption – Specify model assumptions and input variables

- Assumption 1: Sample sizes $n_0$ and $n_1$ are known, but not the cell numbers.

<table>
<thead>
<tr>
<th>genotype</th>
<th>Disease status</th>
<th></th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Case</td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>$n_{10}$</td>
<td>$n_{11}$</td>
<td>$n_1.$</td>
</tr>
<tr>
<td>Aa</td>
<td>$n_{20}$</td>
<td>$n_{21}$</td>
<td>$n_2.$</td>
</tr>
<tr>
<td>aa</td>
<td>$n_{30}$</td>
<td>$n_{31}$</td>
<td>$n_3.$</td>
</tr>
<tr>
<td></td>
<td>$n_0$</td>
<td>$n_1$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

- Assumption 2: The population disease prevalence rate is known: $k$.

- Assumption 3: Population allele frequencies are known: $\text{prob}(A) = p; \text{prob}(a) = q$. 
Step 1: Assumption – Specify model assumptions and input variables (Cont.)

- Assumption 4: The genotypic proportions, \( \text{prob}(AA) = p^2; \text{prob}(Aa) = 2pq; \text{prob}(aa) = q^2 \), satisfies the following relationship (known as the Hardy-Weinburg equilibrium):
  \[ p + q = 1 \text{ and } p^2 + 2pq + q^2 = 1. \]

- Assumption 5: The genetic model is known (expect for \( \beta_1 \)): genotypes affect the disease outcome in the following manner.
  \[ \text{logit}(\text{disease} = 1) = \beta_1 + \beta_2 I_{Aa} + \beta_3 I_{aa} + \epsilon \]
**Step 2: Derivation – Derive the other parameters**

- Use their internal relationships (constraints) to calculate the unknown parameters:
  - To simulate the whole dataset, we need the following genotype frequencies: $P(AA|\text{case})$, $P(Aa|\text{case})$, $P(aa|\text{case})$, $P(AA|\text{control})$, $P(Aa|\text{control})$, $P(aa|\text{control})$
  - To derive them, use the rule of conditional probabilities:

$$P(AA|\text{case}) = \frac{P(AA)P(\text{case}|AA)}{P(\text{case})} = \frac{p^2 \alpha}{k} \quad (1)$$

where $\alpha$ is the baseline disease rate (disease rate among patients with AA genotype). It can be solved through the following equation with the `uniroot()` function:

$$k = p^2 \alpha + 2p(1-p) \frac{\alpha e^{\beta_2}}{1 - \alpha + \alpha e^{\beta_2}} + (1-p)^2 \frac{\alpha e^{\beta_3}}{1 - \alpha + \alpha e^{\beta_3}} \quad (2)$$
Step 3: Implementation – Simulate the data accordingly

- Generate the outcome variable according to: $n_0$ and $n_1$
- Generate the genotypes for each observation according to:
  $P(AA|case)$, $P(Aa|case)$, $P(aa|case)$, $P(AA|control)$, $P(Aa|control)$, $P(aa|control)$
- Lastly, wrap everything into a complete and independent function for later use.
Resample technique is used for bootstrap and data expansion. The key function is: `sample(base, n, replace=TRUE)`

For example: we can expand the above simulated data in a way that:

- \((\text{number of cases}) / (\text{number of cases and controls}) = k\)
- Therefore, keep the cases untouched, and resample from the controls s.t.: \(n_0 = (1/k - 1) \times n_1\)
Summary

- When data are unavailable: three steps for simulation
  - Assumption
  - Derivation
  - Implementation

- When data are small: use sample() for resampling
Acknowledgement

- Professor Chang Yu
- Professor Chuan Zhou
- Tebeb Gebretsadik
- Svetlana Eden
- Peggy Schuyler