Basics

- Simulation studies are commonly done to evaluate the performance of a frequentist statistical procedure, or to compare the performance of two or more different procedures for the same problem.
- Enable us to see what happens “when many many samples of the same size are drawn from the same population.”
- Properties of estimators that are often evaluated by simulation:
  - Bias
  - Mean squared error
  - Coverage of confidence intervals
- Properties of hypothesis tests also can be evaluated by simulation studies:
  - Size
  - Power
- Simulation studies are experiments, and the things you know about experimental design and sample size calculation apply.

Terminology

- Simulation: a numerical technique for conducting experiments on the computer.
- Monte Carlo simulation: a computer experiment involving random sampling from probability distributions
  - What statisticians usually mean by “simulations.”

Rationale

- Properties of statistical methods must be established before the methods can safely be used in practice.
- But exact analytical derivations of properties are rarely possible.
- Large sample approximations to properties are often possible
  - Evaluation of the relevance of the approximation to (finite) sample sizes likely to be encountered in practice is needed.
- Analytical results may require assumptions such as normality.
  - What happens when these assumptions are violated? Analytical results, even large sample ones, may not be possible.
Questions to be addressed regarding an estimator or testing procedure

- Is an estimator biased in finite samples? What is its sampling variance?
- How does it compare to competing estimators on the basis of bias, precision, etc.?
- Does a hypothesis testing procedure attain the claimed level or size?
- If so, what power is possible against different alternatives to the null hypothesis? Do different test procedures deliver different power?

Role of Monte Carlo simulation

- Goal is to evaluate sampling distribution of an estimator under a particular set of conditions (sample size, error distribution, etc.)
- Analytic derivation of exact sampling distribution is not feasible
- Solution: Approximate the sampling distribution through simulation
  - Generate $S$ independent data sets under the conditions of interest
  - Compute the numerical value of the estimator/test statistic $T(data)$ for each data set, yielding $T_1, \ldots, T_S$
- If $S$ is large enough, summary statistics across $T_1, \ldots, T_S$ should be good approximations to the true sampling properties of the estimator/test statistic under the conditions of interest

Simulation for properties of estimators

Simple example: Compare three estimators for the mean $\mu$ of a distribution based on i.i.d. draws $Y_1, \ldots, Y_n$

- Sample mean $T^{(1)}$
- Sample 20% trimmed mean $T^{(2)}$
- Sample median $T^{(3)}$

Remarks:

- If the distribution of the data is symmetric, all three estimators indeed estimate the mean
- If the distribution is skewed, they do not

Simulation procedure

For a particular choice of $\mu$, $n$, and true underlying distribution

- Generate independent draws $Y_1, \ldots, Y_n$ from the distribution
- Compute $T^{(1)}$, $T^{(2)}$, $T^{(3)}$
- Repeat $S$ times $\Rightarrow T^{(1)}_1, \ldots, T^{(1)}_S; T^{(2)}_1, \ldots, T^{(2)}_S; T^{(3)}_1, \ldots, T^{(3)}_S$
- Compute for $k = 1, 2, 3$

$$\hat{\text{mean}} = S^{-1} \sum_{s=1}^{S} T^{(k)}_s = T^{(k)}, \quad \hat{\text{bias}} = T^{(k)} - \mu$$

$$\hat{\text{SD}} = \sqrt{(S-1)^{-1} \sum_{s=1}^{S} (T^{(k)}_s - T^{(k)})^2}$$

$$\hat{\text{MSE}} = S^{-1} \sum_{s=1}^{S} (T^{(k)}_s - \mu)^2 \approx \hat{\text{SD}}^2 + \hat{\text{bias}}^2$$
Relative efficiency

For a particular choice of $\mu$,

Relative efficiency: For any estimators for which $E(T^{(1)}) = E(T^{(2)}) = \mu$

$$RE = \frac{\text{var}(T^{(1)})}{\text{var}(T^{(2)})}$$

is the relative efficiency of estimator 2 to estimator 1

- When the estimators are not unbiased it is standard to compute

$$RE = \frac{\text{MSE}(T^{(1)})}{\text{MSE}(T^{(2)})}$$

- In either case $RE < 1$ means estimator 1 is preferred (estimator 2 is inefficient relative to estimator 1 in this sense)

Normal data:

```r
> set.seed(3)
> S <- 1000
> n <- 15
> trimmean <- function(Y){mean(Y,0.2)}
> mu <- 1
> sigma <- sqrt(5/3)
> out <- generate.normal(S,n,mu,sigma)
> outsampmean <- apply(out$dat,1,mean)
> outtrimmean <- apply(out$dat,1,trimmean)
> outmedian <- apply(out$dat,1,median)
> summary.sim <- data.frame(mean=outsampmean,trim=outtrimmean,
+ median=outmedian)
> results <- simsum(summary.sim,mu)
```

First 5 rows

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>trim</th>
<th>median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7539</td>
<td>0.7132</td>
<td>1.0389</td>
</tr>
<tr>
<td>2</td>
<td>0.6439</td>
<td>0.4580</td>
<td>0.3746</td>
</tr>
<tr>
<td>3</td>
<td>1.5553</td>
<td>1.6710</td>
<td>1.9395</td>
</tr>
<tr>
<td>4</td>
<td>0.5171</td>
<td>0.4827</td>
<td>0.4119</td>
</tr>
<tr>
<td>5</td>
<td>1.3603</td>
<td>1.4621</td>
<td>1.3452</td>
</tr>
</tbody>
</table>
Performance of estimates of uncertainty

How well do estimated standard errors represent the true sampling variation?

- E.g., For sample mean \( T^{(1)}(Y_1, \ldots, Y_n) = \bar{Y} \)
  \[ SE(\bar{Y}) = \frac{s}{\sqrt{n}}, \quad s^2 = (n-1)^{-1} \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \]
- MC standard deviation approximates the true sampling variation
- Compare average of estimated standard errors to MC standard deviation

For sample mean: MC standard deviation 0.331

```r
> outsampmean <- apply(out$dat,1,mean)
> sampmean.ses <- sqrt(apply(out$dat,1,var)/n)
> ave.sampmeanses <- mean(sampmean.ses)
> round(ave.sampmeanses,3)
[1] 0.329
```

Usual 100(1-\(\alpha\))% confidence interval for \( \mu \):

Based on sample mean

\[ [\bar{Y} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{Y} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}] \]

- Does the interval achieve the nominal level of coverage 1 - \(\alpha\)?
- E.g. \(\alpha = 0.05\)

```r
> t05 <- qt(0.975,n-1)
> coverage <- sum((outsampmean-t05n*sampmean.ses <= mu) & (outsampmean+t05n*sampmean.ses >= mu))/S
> coverage
[1] 0.949
```

Simulations for properties of hypothesis tests

Simple example: Size and power of the usual \(t\)-test for the mean

\( H_0 : \mu = \mu_0 \) vs. \( H_1 : \mu \neq \mu_0 \)

- To evaluate whether size/level of test achieves advertised \(\alpha\) generate data under \( \mu = \mu_0 \) and calculate proportion of rejections of \( H_0 \)
- Approximates the true probability of rejecting \( H_0 \) when it is true
- Proportion should \( \approx \alpha \)
- To evaluate power, generate data under some alternative \( \mu \neq \mu_0 \) and calculate proportion of rejections of \( H_0 \)
- Approximates the true probability of rejecting \( H_0 \) when the alternative is true (power)
- If actual size is > \(\alpha\), then evaluation of power is flawed

```r
> set.seed(3); S <- 1000; n <- 15; sigma <- sqrt(5/3)
> mu0 <- 1; mu <- 1
> out <- generate.normal(S,n,mu,sigma)
> ttests <- (apply(out$dat,1,mean)-mu0)/sqrt(apply(out$dat,1,var)/n)
> t05 <- qt(0.975,n-1)
> power <- sum(abs(ttests)>t05)/S
> power
[1] 0.051
```
Power of test:

> set.seed(3); S <- 1000; n <- 15; sigma <- sqrt(S/3)
> mu0 <- 1; mu <- 1.75
> out <- generate.normal(S,n,mu,sigma)
> ttests <- (apply(out$dat,1,mean)-mu0)/sqrt(apply(out$dat,1,var)/n)
> t05 <- qt(0.975,n-1)
> power <- sum(abs(ttests)>t05)/S
> power
[1] 0.534

Simulation study principles

Issue: How well do the Monte Carlo quantities approximate properties of the true sampling distribution of the estimator/test statistic?

- Is $S = 1000$ large enough to get a feel for the true sampling properties? How “believable” are the results?
- A simulation is just an experiment like any other, so use statistical principles!
- Each data set yields a draw from the true sampling distribution, so $S$ is the “sample size” on which estimates of mean, bias, SD, etc. of this distribution are based
- Select a “sample size” ($S$) that will achieve acceptable precision of the approximation in the usual way!

Principle 1: A Monte Carlo simulation is just like any other experiment

- Careful planning is required
- Factors that are of interest to vary in the experiment: sample size $n$, distribution of the data, magnitude of variation, ...
- Each combination of factors is a separate simulation, so that many factors can lead to very large number of combinations and thus number of simulations — time consuming
- Use experimental design principles
- Results must be recorded and saved in a systematic, sensible way
- Don’t choose only factors favorable to a method you have developed!
- “Sample size $S$ (number of data sets in each simulation) must deliver acceptable precision...
Choosing $S$: Coverage probabilities, size, power

- Estimating a proportion $p$ (= coverage probability, size, power) $\Rightarrow$ binomial sampling, e.g. for a hypothesis test

$$Z = \#\text{rejections} \sim \text{binomial}(S, p) \Rightarrow \sqrt{\frac{\text{var}(Z)}{S}} = \sqrt{\frac{p(1-p)}{S}}$$

- Worst case is at $p = 1/2$ $\Rightarrow 1/\sqrt{4S}$
- $d$ acceptable error $\Rightarrow S = 1/(4d^2)$; e.g., $d = 0.01$ yields $S = 2500$
- For coverage, size, $p = 0.05$

Principle 2: Save everything!

- Save the individual estimates in a file and then analyze (mean, bias, SD, etc) later
  - as opposed to computing these summaries and saving only them
- Critical if the simulation takes a long time to run!
- Advantage: can use software for summary statistics (e.g., SAS, R, etc.)

Principle 3: Keep $S$ small at first

- Test and refine code until you are sure everything is working correctly before carrying out final “production” runs
- Get an idea of how long it takes to process one data set

Principle 4: Set a different seed for each run and keep records

- Ensure simulation runs are independent
- Runs may be replicated if necessary

Principle 5: Document your code