SIMULATION STUDIES IN STATISTICS

- What is a (Monte Carlo) simulation study, and why do one?
- Simulations for properties of estimators
- Simulations for properties of hypothesis tests
- Simulation study principles
- Presenting results

WHAT IS A SIMULATION STUDY, AND WHY DO ONE?

Simulation: A numerical technique for conducting experiments on the computer

Monte Carlo simulation: Computer experiment involving random sampling from probability distributions

- Invaluable in statistics…
- Usually, when statisticians talk about "simulations," they mean "Monte Carlo simulations"

Rationale: In statistics

- Properties of statistical methods must be established so that the methods may be used with confidence
- Exact analytical derivations of properties are rarely possible
- Large sample approximations to properties are often possible, however…
- …evaluation of the relevance of the approximation to (finite) sample sizes likely to be encountered in practice is needed
- Moreover, analytical results may require assumptions (e.g., normality)
- But what happens when these assumptions are violated? Analytical results, even large sample ones, may not be possible

Usual issues: Under various conditions

- Is an estimator biased in finite samples? Is it still consistent under departures from assumptions? What is its sampling variance?
- How does it compare to competing estimators on the basis of bias, precision, etc.?
- Does a procedure for constructing a confidence interval for a parameter achieve the advertised nominal level of coverage?
- Does a hypothesis testing procedure attain the advertised level or size?
- If it does, what power is possible against different alternatives to the null hypothesis? Do different test procedures deliver different power?

Usual issues: Under various conditions

- Is an estimator biased in finite samples? Is it still consistent under departures from assumptions? What is its sampling variance?
- How does it compare to competing estimators on the basis of bias, precision, etc.?
- Does a procedure for constructing a confidence interval for a parameter achieve the advertised nominal level of coverage?
- Does a hypothesis testing procedure attain the advertised level or size?
- If it does, what power is possible against different alternatives to the null hypothesis? Do different test procedures deliver different power?

How to answer these questions in the absence of analytical results?

Monte Carlo simulation to the rescue:

- An estimator or test statistic has a true sampling distribution under a particular set of conditions (finite sample size, true distribution of the data, etc.)
- Ideally, we would want to know this true sampling distribution in order to address the issues on the previous slide
- But derivation of the true sampling distribution is not tractable
- $\Rightarrow$ Approximate the sampling distribution of an estimator or test statistic under a particular set of conditions

How to approximate: A typical Monte Carlo simulation involves the following

- Generate S independent data sets under the conditions of interest
- Compute the numerical value of the estimator/test statistic $T(\text{data})$ for each data set $\Rightarrow T_1, \ldots, T_S$
- 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

How to approximate: A typical Monte Carlo simulation involves the following

- Generate S independent data sets under the conditions of interest
- Compute the numerical value of the estimator/test statistic $T(\text{data})$ for each data set $\Rightarrow T_1, \ldots, T_S$
- 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

E.g., for an estimator for a parameter θ: T_s is the value of T from the sth data set, $s = 1, \ldots, S$

- The sample mean over S data sets is an estimate of the true mean of the sampling distribution of the estimator

SIMULATIONS FOR PROPERTIES OF ESTIMATORS

Simple example: Compare three estimators for the mean μ 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 $T^{(3)}$

SIMULATIONS FOR PROPERTIES OF ESTIMATORS

Simple example: Compare three estimators for the mean μ 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 μ, 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 $\Rightarrow$

$$T_1^{(1)}, \ldots, T_S^{(1)}; \quad T_1^{(2)}, \ldots, T_S^{(2)}; \quad T_1^{(3)}, \ldots, T_S^{(3)}$$

- Compute for $k = 1, 2, 3$

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

$$\widehat{\text{SD}} = \sqrt{(S-1)^{-1} \sum_{s=1}^{S} (T_s^{(k)} - \overline{T}^{(k)})^2}, \quad \widehat{\text{MSE}} = S^{-1} \sum_{s=1}^{S} (T_s^{(k)} - \mu)^2 \approx \widehat{\text{SD}}^2 + \widehat{\text{bias}}^2$$

Relative efficiency: For any estimators for which

$$E(T^{(1)}) = E(T^{(2)}) = \mu \quad \Rightarrow \quad RE = \frac{\text{var}(T^{(1)})}{\text{var}(T^{(2)})}$$

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)

In R: See class website for program

```
> set.seed(3)

> S <- 1000

> n <- 15

> trimmean <- function(Y){mean(Y,0.2)}

> mu <- 1

> sigma <- sqrt(5/3)
```


Normal data:

```
> 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)
```



```
> view(round(summary.sim,4),5)
First 5 rows

    mean   trim median
1 0.7539 0.7132 1.0389
2 0.6439 0.4580 0.3746
3 1.5553 1.6710 1.9395
4 0.5171 0.4827 0.4119
5 1.3603 1.4621 1.3452
```



```
> results
                      Sample mean Trimmed mean   Median
true value                  1.000        1.000    1.000
# sims                   1000.000     1000.000 1000.000
MC mean                     0.985        0.987    0.992
MC bias                    -0.015       -0.013   -0.008
MC relative bias           -0.015       -0.013   -0.008
MC standard deviation       0.331        0.348    0.398
MC MSE                      0.110        0.121    0.158
MC relative efficiency      1.000        0.905    0.694
```

Performance of estimates of uncertainty: How well do estimated standard errors represent the true sampling variation?

- E.g., For sample mean $T^{(1)}(Y_1,\ldots,Y_n) = \overline{Y}$

$$SE(\overline{Y}) = \frac{s}{\sqrt{n}}, \qquad s^2 = (n-1)^{-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2$$

- MC standard deviation approximates the true sampling variation
- $\Rightarrow$ Compare average of estimated standard errors to MC standard deviation

For sample mean: MC standard deviation 0.331

```
> 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-α)% confidence interval for μ: Based on sample mean

$$\left[\, \overline{Y} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{Y} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \,\right]$$

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

```
> 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

Real simple example: Size and power of the usual t-test for the mean

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq mu_0$$

- To evaluate whether size/level of test achieves advertised α 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

Size/level of test:

```
> 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(5/3)

> 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" (number of data sets 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 $\Rightarrow$ time consuming
- Can use experimental design principles
- Results must be recorded and saved in a systematic, sensible way
- Don't only choose factors favorable to a method you have developed!
- "Sample size S (number of data sets) must deliver acceptable precision...

Choosing S: Estimator for θ (true value θ_0)

- Estimation of mean of sampling distribution/bias:

$$\sqrt{\mathrm{var}(\overline{T}-\theta_0)} = \sqrt{\mathrm{var}(\overline{T})} = \sqrt{\mathrm{var}\left(S^{-1}\sum_{s=1}^{S} T_s\right)} = \frac{\mathrm{SD}(T_s)}{\sqrt{S}} = d$$

where d is the acceptable error

$$\Rightarrow \quad S = \frac{\{\mathrm{SD}(T_s)\}^2}{d^2}$$

- Can "guess" $\mathrm{SD}(T_s)$ from asymptotic theory, preliminary runs

Choosing S: Coverage probabilities, size, power

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

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

- 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!!!

PRESENTING THE RESULTS

Key principle: Your simulation is useless unless other people can clearly and unambigously understand what you did and why you did it, and what it means!

What did you do and why? Before giving results, you must first give a reader enough information to appreciate them!

- State the objectives – Why do this simulation? What specific questions are you trying to answer?
- State the rationale for choice of factors studied, assumptions made
- Review all methods under study – be precise and detailed
- Describe exactly how you generated data for each choice of factors – enough detail should be given so that a reader could write his/her own program to reproduce your results!

Results: Must be presented in a form that

- Clearly answers the questions
- Makes it easy to appreciate the main conclusions

Results: Must be presented in a form that

- Clearly answers the questions
- Makes it easy to appreciate the main conclusions

Some basic principles:

- Only present a subset of results ("Results were qualitatively similar for all other scenarios we tried.")
- Only present information that is interesting ("Relative biases for all estimators were less than 2% under all scenarios and hence are not shown in the table.")
- The mode of presentation should be friendly. . .

Tables: An obvious way to present results, however, some caveats

- Avoid zillions of numbers jam-packed into a table!
- Place things to be compared adjacent to one another so that comparison is easy
- Rounding. . .

Rounding: Three reasons (Wainer, 1993)

- Humans cannot understand more than two digits very easily
- More than two digits can almost never be statistically justified
- We almost never care about accuracy of more than two digits

Wainer, H. (1993) Visual Revelations, *Chance Magazine*

Understanding/who cares?

- "This year's school budget is \$27,329,681.32" or "This year's school budget is about 27 million dollars"
- "Mean life expectancy of Australian males is 67.14 years" or "Mean life expectancy of Australian males is 67 years"

Statistical justification: We are statisticians! For example

- Reporting Monte Carlo power – how many digits?
- Design the study to achieve the desired accuracy and only report what we can justify as accurate
- The program yields 0.56273
- If we wish to report 0.56 (two digits) need the standard error of this estimated proportion to be ≤ 0.005 so we can tell the difference between 0.56 and 0.57 or 0.58 $(1.96 \times 0.005 \approx 0.01)$
- $d = 0.005 = 1/\sqrt{4S}$ gives $S = 10000!$

Statistical justification: We are statisticians! For example

- Reporting Monte Carlo power – how many digits?
- Design the study to achieve the desired accuracy and only report what we can justify as accurate
- The program yields 0.56273
- If we wish to report 0.56 (two digits) need the standard error of this estimated proportion to be ≤ 0.005 so we can tell the difference between 0.56 and 0.57 or 0.58 $(1.96 \times 0.005 \approx 0.01)$
- $d = 0.005 = 1/\sqrt{4S}$ gives $S = 10000!$

Always report the standard error of entries in the table so a reader can gauge the accuracy!

Bad table: Digits, "apples and oranges"

	Sample mean		Trimmed mean		Median	
	Normal	t_5	Normal	t_5	Normal	t_5
Mean	0.98515	0.98304	0.98690	0.98499	0.99173	0.98474
Bias	−0.01485	−0.01696	−0.01310	-0.01501	-0.00827	-0.01526
SD	0.33088	0.33067	0.34800	0.31198	0.39763	0.35016
MSE	0.10959	0.10952	0.12116	0.09746	0.15802	0.12273
Rel. Eff.	1.00000	1.00000	0.90456	1.12370	0.69356	0.89238

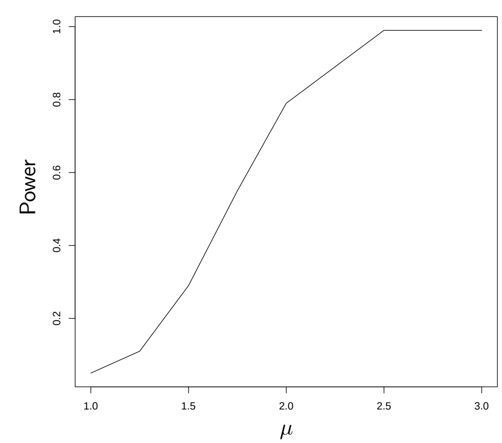
Good table: Digits, "apples with apples"

	Normal			t_5		
	Sample mean	Trim mean	Median	Sample mean	Trim mean	Median
Mean	0.99	0.99	0.99	0.98	0.98	0.98
Bias	−0.01	−0.01	−0.01	−0.02	−0.02	−0.02
SD	0.33	0.35	0.40	0.33	0.31	0.35
MSE	0.11	0.12	0.16	0.11	0.10	0.12
Rel. Eff.	1.00	0.90	0.69	1.00	1.12	0.89

Graphs: Often a more effective strategy than tables!

Example: Power of the t-test for $H_0 : \mu = 1.0$ vs. $H_1 : \mu \neq 1.0$ for normal data ($S = 10000$, $n = 15$)

μ	1.0	1.25	1.50	1.75	2.00	2.50	3.00
power	0.05	0.11	0.29	0.55	0.79	0.99	0.99



Must reading: Available on the class web page

Gelman, A., Pasarica, C., and Dodhia, R. (2002). Let's practice what preach: Turning tables into graphs. *The American Statistician*