Simulation

Berlin Chen

Department of Computer Science & Information Engineering

National Taiwan Normal University

Reference:

1. W. Navidi. *Statistics for Engineering and Scientists*. Section 4.11 & Teaching Material

Simulation

- Simulation refers to the process of generating random numbers and treating them as if they were data generated by an actual scientific distribution
  - The data so generated are called simulated or synthetic data (i.e., computer-generated numbers)

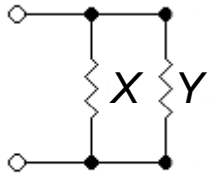
- E.g., the sexes of fraternal twins
  - Assume each twin is equally likely to be a boy or a girl and the sexes of twins are determined independently
  - What's the probability that both twins are boys? (Assume we don't know the multiplication rule $P(T_1 = B, T_2 = B) = P(T_1 = B)P(T_2 = B) = 0.5 \cdot 0.5 = 0.25$)

  Answer

  - Estimate the probability that two fair coins both lands heads
    - Head → Boy; Tail → Girl (a kind of Bernoulli trail)
    - Compute the portion of tosses in which both coins landed heads

Using Simulation to Estimate a Probability (1/3)

- Given that two resistors X and Y labeled with 100 Ω and 25 Ω are connected in parallel, and the actual resistances of X and Y may differ from the labeled values, $X \sim N(100, 10^2)$ and $Y \sim N(25, 2.5^2)$
  - What is the probability that the total resistances of the assembly $R=XY/(X+Y)$ is in the range of $19<R<21$?



Answer

- First take a sample of N resistors labeled 100 Ω whose actual resistances are $X_1, X_2, \ldots, X_N$
- Then independently take a equal size sample of resistors labeled25 Ω whose actual resistances are $Y_1, Y_2, \ldots, Y_N$
- Construct N assembles with resistances $R_1=X_1Y_1/(X_1+Y_1)$, $R_2=X_2Y_2/(X_2+Y_2), \ldots, R_N=X_NY_N/(X_N+Y_N)$, where the values $R_1, R_2, .., R_N$ can be viewed as a random sample from the population of all possible values of the total resistance
- Compute the portion of $R_1, R_2, .., R_N$ falling between 19 and 21

Using Simulation to Estimate a Probability (2/3)

– 48 values out of the sample of 100 are determined to fall in the range between 19 and 21 ($P(19<R<21) \sim 0.48$)

TABLE 4.2 Simulated data for resistances in a parallel circuit

	X^*	Y^*	R^*		X^*	Y^*	R^*		X^*	Y^*	R^*		X^*	Y^*	R^*
1	102.63	24.30	19.65	26	115.94	24.93	20.52	51	94.20	23.68	18.92	76	113.32	22.54	18.80
2	96.83	21.42	17.54	27	100.65	28.36	22.13	52	82.62	27.82	20.81	77	90.82	23.79	18.85
3	96.46	26.34	20.69	28	89.71	23.00	18.31	53	119.49	22.88	19.20	78	102.88	25.99	20.75
4	88.39	22.10	17.68	29	104.93	24.10	19.60	54	99.43	28.03	21.87	79	93.59	23.04	18.49
5	113.07	29.17	23.19	30	93.74	23.68	18.91	55	108.03	21.69	18.06	80	89.19	27.05	20.76
6	117.66	27.09	22.02	31	104.20	24.02	19.52	56	95.32	20.60	16.94	81	95.04	23.76	19.01
7	108.04	18.20	15.58	32	123.43	26.66	21.93	57	80.70	30.36	22.06	82	109.72	30.25	23.71
8	101.13	28.30	22.11	33	101.38	22.19	18.21	58	91.13	20.38	16.66	83	107.35	27.01	21.58
9	105.43	23.51	19.22	34	88.52	26.85	20.60	59	111.35	27.09	21.79	84	89.59	18.55	15.37
10	103.41	24.64	19.90	35	101.23	26.88	21.24	60	118.75	23.92	19.91	85	101.72	24.65	19.84
11	104.89	22.59	18.58	36	86.96	25.66	19.81	61	103.33	23.99	19.47	86	101.24	25.92	20.64
12	94.91	27.86	21.54	37	95.92	26.16	20.55	62	107.77	18.08	15.48	87	109.67	26.61	21.41
13	92.91	27.06	20.96	38	95.97	26.05	20.49	63	104.86	24.64	19.95	88	100.74	26.18	20.78
14	95.86	24.82	19.71	39	93.76	24.71	19.56	64	84.39	25.52	19.60	89	98.44	23.63	19.06
15	100.06	23.65	19.13	40	113.89	22.79	18.99	65	94.26	25.61	20.14	90	101.05	28.81	22.42
16	90.34	23.75	18.81	41	109.37	26.19	21.13	66	82.16	27.49	20.60	91	88.13	28.43	21.49
17	116.74	24.38	20.17	42	91.13	24.93	19.58	67	108.37	27.35	21.84	92	113.94	29.45	23.40
18	90.45	25.30	19.77	43	101.60	28.66	22.36	68	86.16	21.46	17.18	93	97.42	23.78	19.11
19	97.58	23.05	18.65	44	102.69	21.37	17.69	69	105.97	23.59	19.30	94	109.05	23.04	19.02
20	101.19	23.60	19.14	45	108.50	25.34	20.54	70	92.69	23.88	18.99	95	100.65	26.63	21.06
21	101.77	31.42	24.01	46	80.86	27.55	20.55	71	97.48	25.43	20.17	96	105.64	21.57	17.91
22	100.53	24.93	19.98	47	85.80	24.80	19.24	72	110.45	20.70	17.44	97	78.82	23.25	17.95
23	98.00	27.57	21.52	48	105.96	23.20	19.03	73	89.92	27.23	20.90	98	112.31	22.77	18.93
24	108.10	27.51	21.93	49	103.98	21.78	18.01	74	103.78	25.67	20.58	99	100.14	24.95	19.97
25	91.07	23.38	18.60	50	97.97	23.13	18.71	75	95.53	25.55	20.16	100	88.78	25.87	20.03

Using Simulation to Estimate a Probability (3/3)

- Simulation using MATLAB

```
a = 100+10.*randn(100,1);
b = 25+2.5.*randn(100,1);
c = (a.*b)./(a+b);
d = find(c>19 & c<21);
disp(size(d,1)/size(c,1));
```

Example 4.70 (1/2)

- An engineer has to choose between two types of cooling fans to install in a computer. The lifetimes, in months, of fans of type A are exponentially distributed with mean 50 months, and the lifetime of fans of type B are exponentially distributed with mean 30 months.
  - Since type A fans are more expensive, the engineer decides that she will choose type A fans if the probability that a type A fan will last more than twice as long as a type B fan is greater than 0.5 ($P(A>2B) > 0.5$?). Estimate this probability.

TABLE 4.3 Simulated data for Example 4.70

	A^*	B^*	$A^* > 2B^*$
1	25.554	12.083	1
2	66.711	11.384	1
3	61.189	15.191	1
4	9.153	119.150	0
5	98.794	45.258	1
6	14.577	139.149	0
7	65.126	9.877	1
8	13.205	12.106	0
9	20.535	21.613	0
10	62.278	13.289	1
⋮	⋮	⋮	⋮
1000	19.705	12.873	0

Answer:

- Among the 1000 simulated pairs, there are 460 for which $A^*>2B^*$.

Therefore, the estimated probability

$P(A>2B) = 0.46$

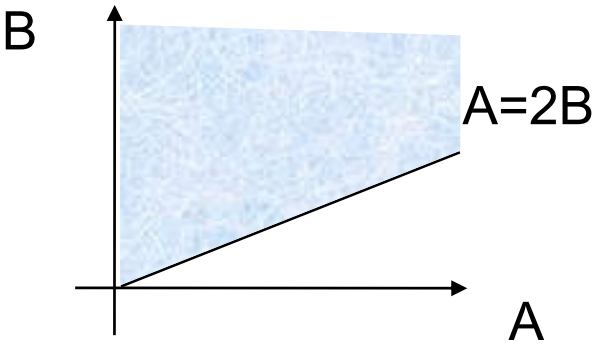
Example 4.70 (2/2)

- However, the exact probability $P(A>2B)$ is 5/11=0.4545

$$f_A(a)=\lambda_A e^{-\lambda_A a} \quad \left(\lambda_A=\frac{1}{50}\right)$$

$$f_B(b)=\lambda_B e^{-\lambda_B b} \quad \left(\lambda_B=\frac{1}{30}\right)$$

$$F_A(a)=\begin{cases}1-e^{-\lambda_A a}, & a\geq 0\\ 0, & \text{otherwise}\end{cases}$$



$$P(A>2B)=\int_0^\infty\int_{2b}^\infty f_{A,B}(a,b)\,da\,db$$

$$=\int_0^\infty\left[\int_{2b}^\infty f_A(a)\,da\right]f_B(b)\,db$$

$$=\int_0^\infty e^{-2\lambda_A b}\left(\lambda_B e^{-\lambda_B b}\right)db$$

$$=\frac{\lambda_B}{2\lambda_A+\lambda_B}\int_0^\infty(2\lambda_A+\lambda_B)e^{-(2\lambda_A+\lambda_B)b}\,db$$

$$=\frac{\lambda_B}{2\lambda_A+\lambda_B}=\frac{1/30}{(2/50)+(1/30)}=5/11$$

(the integral equals 1)

Estimating Means and Variances

- **Example 4.71**: Use the simulated values R_i^* in Table 4.2 to estimate the mean μ_R and standard deviation σ_R of the total resistance R
  - The values $R_1^*, R_2^*, \ldots, R_{100}^*$ can be treated as if they were a random sample of actual total resistances
  - Estimate μ_R with sample mean $\overline{R}^*$ and σ_R with the sample standard deviation s_{R^*}

$$\mu_R \approx \overline{R}^* \text{ (sample mean)} = 19.856$$
$$\sigma_R \approx s_{R^*} \text{ (sample standard deviation)} = 1.6926$$

Comparison with Propagation of Error (1/3)

- Recall that the method of propagation of error (c.f. Section 3.4) can also be used to approximate the mean and variance of a function of random variables, such as $U = U(X_1, \cdots, X_n)$
  - It has to require that the standard deviations of X_i be small due to the Taylor series approximation
  - It doesn't need to know the distributions of X_i and also can pinpoint which of the X_i contributes most to the uncertainty in U
- However, simulation can do things that propagation of error cannot do, such as
  - Estimate probability
  - Determine whether a given function of random variables is normally distributed
  - It has not to require that the standard deviations of X_i be small

Comparison with Propagation of Error (2/3)

- Use the method of propagation of error to estimate the mean μ_R and standard deviation σ_R of the total resistance R ($R = U(X,Y) = XY/(X+Y)$) in Example 4.71
  - For X and Y have small standard deviations (are close to their means μ_X and μ_Y , respectively), we have the following (first-order) Taylor series approximation

$$U(X,Y) - U(\mu_X,\mu_Y) \approx \left(\left.\frac{\partial U}{\partial X}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}}\right)(X-\mu_X) + \left(\left.\frac{\partial U}{\partial Y}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}}\right)(Y-\mu_Y)$$

$$U(X,Y) \approx \left(U(\mu_X,\mu_Y) - \left(\left.\frac{\partial U}{\partial X}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}}\right)\mu_X - \left(\left.\frac{\partial U}{\partial Y}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}}\right)\mu_Y\right) + \left(\left.\frac{\partial U}{\partial X}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}}\right)X + \left(\left.\frac{\partial U}{\partial Y}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}}\right)Y$$

$$U(\mu_X,\mu_Y) = \frac{100*25}{100+25} = 20$$

$$\left(\left.\frac{\partial U}{\partial X}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}}\right) = \frac{\partial(XY/(X+Y))}{\partial X} = \left.\frac{Y(X+Y)-XY}{(X+Y)^2}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}} = \frac{25(100+25)-100\cdot 25}{(100+25)^2} = 0.04$$

$$\left(\left.\frac{\partial U}{\partial Y}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}}\right) = \frac{\partial(XY/(X+Y))}{\partial Y} = \left.\frac{X(X+Y)-XY}{(X+Y)^2}\right|_{\substack{X=\mu_X \\ Y=\mu_Y}} = \frac{100(100+25)-100\cdot 25}{(100+25)^2} = 0.64$$

Comparison with Propagation of Error (3/3)

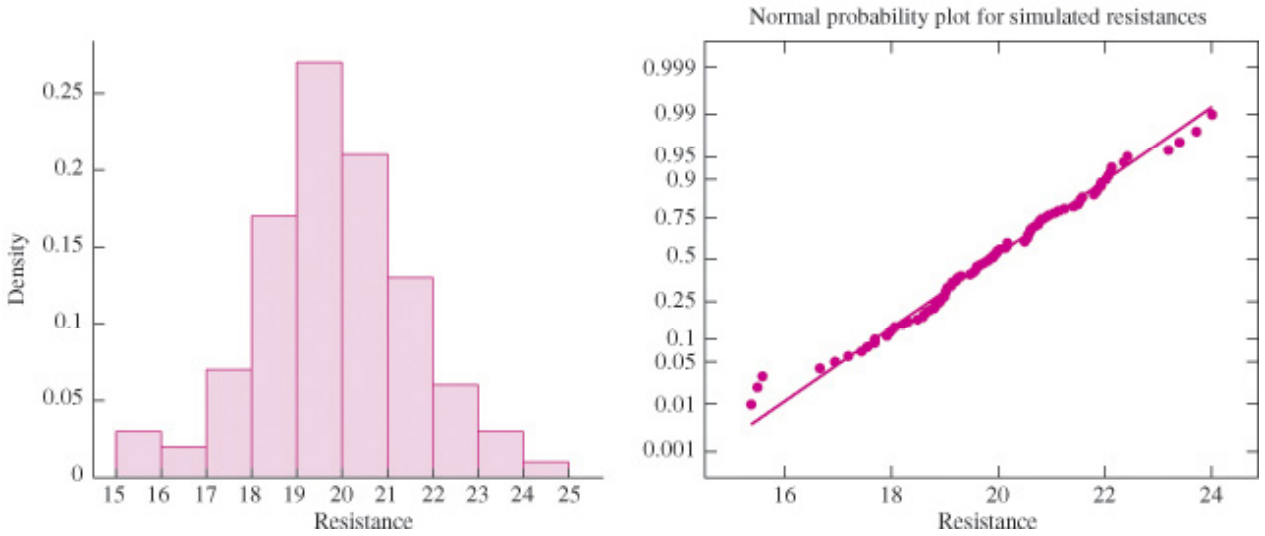
$$U(X,Y) \approx 20 - 0.04 \cdot 100 - 0.64 \cdot 25 + 0.04X + 0.64Y$$
$$= 0.04X + 0.64Y$$
$$\therefore \mathbf{E}[U(X,Y)] = 0.04 \cdot 100 + 0.64 \cdot 25 = 20$$
$$\text{var}(U(X,Y)) = 0.04^2 \,\text{var}(X) + 0.64^2 \,\text{var}(Y)$$
$$= 0.04^2 \cdot 10^2 + 0.64^2 \cdot 2.5^2$$
$$= 2.72$$
$$\therefore \sigma_{U(X,Y)} = 1.6492$$

$$\text{Recall that}: \sigma_U \approx \sqrt{\left(\frac{\partial U}{\partial X_1}\right)^2 \sigma_{X_1}^2 + \left(\frac{\partial U}{\partial X_2}\right)^2 \sigma_{X_2}^2 + \cdots + \left(\frac{\partial U}{\partial X_n}\right)^2 \sigma_{X_n}^2}$$

if $X_1, X_2, \ldots, X_n$ are independent.

Using Simulation to Determine Whether a Population is Approximately Normal

- Construct a histogram and a normal probability plot of the simulated sample to see if the data approximately normal
- **Example 4.72**
  - For the simulated sample of total resistance in Table 4.2



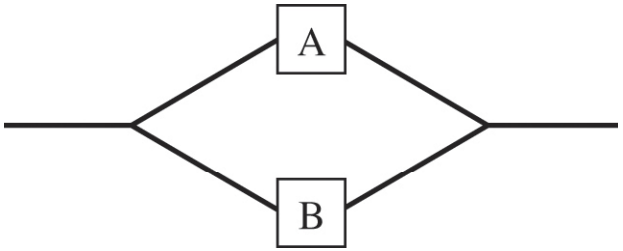
The distribution appears to be approximately normal !

Using Simulation in Reliability Analysis (1/3)

- A system is made up of components, each of which has a lifetime that is random (The lifetime of the system is hence also random). Reliability engineers want to determine the lifetime probability distribution of the system given that the lifetime probability distributions of the components are approximately known
  - It can be very difficult to calculate the distribution of the system lifetime directly from the distribution of the component lifetimes
  - If the lifetimes of the components are independent, it can often be done easily with simulation

Using Simulation in Reliability Analysis (2/3)

- **Example 4.74:** A system consists of components A and B connected in parallel
  - The lifetime in month of A is distributed Exp(1), while that of B is Exp(0.5)



  - The system will fail if both A and B fail
  - Estimate the mean lifetime of the system (in months), the probability that the system functions for less than 1 month, and the 10th percentile of the system lifetime

TABLE 4.5 Simulated data for Example 4.74

	A^*	B^*	L^*
1	0.0245	0.5747	0.5747
2	0.3623	0.3998	0.3998
3	0.8858	1.7028	1.7028
4	0.1106	14.2252	14.2252
5	0.1903	0.4665	0.4665
6	2.2259	1.4138	2.2259
7	0.8881	0.9120	0.9120
8	3.3471	3.2134	3.3471
9	2.5475	1.3240	2.5475
10	0.3614	0.8383	0.8383
⋮	⋮	⋮	⋮
1000	0.3619	1.8799	1.8799

Based on a smaple of 1000 simulated data

$\Rightarrow$ The mean lifetime of the system is approximately 2.29

The probability that the system falls within a month is 0.278

The 10th percentile of the system is 0.516

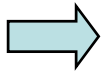
Estimated system life time $L^* = \max\left(A^*, B^*\right)$

Using Simulation in Reliability Analysis (3/3)

- Calculate the system lifetime distribution using the "Derived Distributions" method

$$
\begin{aligned}
F_L(l) &= P(L \le l) \\
&= P(\max(A,B) \le l) \\
&= P(A \le l, B \le l) \\
&= P(A \le l)P(B \le l) \\
&= \left(1 - e^{-\lambda_A l}\right)\left(1 - e^{-\lambda_B l}\right) \\
&= 1 - e^{-\lambda_A l} - e^{-\lambda_B l} + e^{-(\lambda_A + \lambda_B)l}
\end{aligned}
$$

$$
f_L(l) = \lambda_A e^{-\lambda_A l} + \lambda_B e^{-\lambda_B l} - (\lambda_A + \lambda_B) e^{-(\lambda_A + \lambda_B)l}
$$



The expected lifetime of the system

$$
\begin{aligned}
\mathbf{E}[l] &= \int_0^\infty l \cdot f_L(l)\,dl \\
&= \frac{1}{\lambda_A} + \frac{1}{\lambda_B} - \frac{1}{\lambda_A + \lambda_B} \\
&= \frac{1}{1} + \frac{1}{0.5} - \frac{1}{1.5} \\
&= 2.33
\end{aligned}
$$

$$
\begin{aligned}
P(L \le 1) = F_L(1) &= 1 - e^{-\lambda_A} - e^{-\lambda_B} + e^{-(\lambda_A + \lambda_B)} \\
&= 1 - e^{-1} - e^{-0.5} + e^{-1.5} \\
&= 0.2487
\end{aligned}
$$

Using Simulation to Estimate Bias

- The sample standard deviation s of a random sample $X_1,\ldots,X_n$ is used to estimate the population standard deviation σ
  - We know that s is a biased estimate
  - Can we use simulation to estimate the bias in s ?
- Example 4.75

TABLE 4.6 Simulated data for Example 4.75

	X_1^*	X_2^*	X_3^*	X_4^*	X_5^*	X_6^*	s^*
1	−0.4326	0.7160	−0.6028	0.8304	−0.1342	−0.3560	0.6160
2	−1.6656	1.5986	−0.9934	−0.0938	0.2873	−1.8924	1.3206
3	0.1253	−2.0647	1.1889	−0.4598	0.3694	0.4906	1.1190
4	−1.7580	0.1575	−0.8496	0.3291	−1.5780	−1.1100	0.8733
5	1.6867	0.3784	0.3809	0.4870	0.9454	−0.4602	0.7111
6	1.3626	0.7469	−2.1102	2.6734	−0.5311	1.1611	1.6629
7	−2.2424	−0.5719	−1.9659	0.1269	−0.2642	0.3721	1.0955
8	1.3765	−0.4187	−0.5014	1.9869	−0.0532	−0.7086	1.1228
9	−1.8045	0.5361	−0.9121	1.4059	−1.2156	−0.9619	1.2085
10	0.3165	0.6007	−0.5363	−0.2300	0.2626	0.0523	0.4092
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	0.3274	0.1787	0.2006	−1.1602	1.1020	0.3173	0.7328

$X_1^*,\ldots,X_1^*$ is a simple random sample from $N(0,1)$

1. s^* is the sample deviation of the simulated simple random sample

The bias in s^* can be expressed by $\mu_{s^*} - \sigma$

2. We can also veiw the values of $s_1^*,\ldots,s_{1000}^*$ a random sample from the population of all possible values of s^*

3. μ_{s^*} can be estimated by $\bar{s}^*$, then $\mu_{s^*} - \sigma$ is approximately estimated by $\bar{s}^* - \sigma = 0.9601 - 1 = -0.0399$

(Parametric) Bootstrap Methods (1/2)

- (Parametric) Bootstrap Methods: simulation methods in which the distribution to be sampled from is determined from the data (the distribution parameters are unknown in advance)
- **Example 4.76**
  - A sample, 5.23, 1.93, 5.66, 3.28, 5.93 and 6.21, is taken from a normal distribution whose mean and variance are unknown
  - The sample mean $\overline{X} = 4.7067$ and the sample standard deviation $s = 1.7137$. Estimate the bias in s .

TABLE 4.7 Simulated data for Example 4.76

	X_1^*	X_2^*	X_3^*	X_4^*	X_5^*	X_6^*	s^*
1	2.3995	4.8961	3.6221	6.9787	4.4311	4.5367	1.5157
2	2.6197	4.3102	3.2350	6.2619	4.4233	3.5903	1.2663
3	3.0114	5.2492	7.6990	6.0439	6.5965	3.7505	1.7652
4	3.9375	5.2217	1.9737	4.5434	3.0304	3.8632	1.1415
5	5.8829	5.3084	4.6003	2.6439	2.3589	2.3055	1.6054
6	7.8915	3.9731	5.1229	5.1749	3.5255	3.3330	1.6884
7	4.2737	5.5189	2.3314	5.1512	5.7752	4.0205	1.2705
8	5.8602	5.3280	5.5860	6.8256	7.5063	3.9393	1.2400
9	5.7813	4.9364	2.5893	3.7633	0.9065	3.8372	1.7260
10	3.3690	1.8618	2.7627	3.2837	3.9863	6.0382	1.4110
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	2.0496	6.3385	6.2414	5.1580	3.7213	8.4576	2.2364

1. Use the sample mean $\overline{X} = 4.7067$ and sample standard deviation $s = 1.7137$ to estimate the population mean μ and standard deviation σ $\left(\sigma^* = 1.7137\right)$, respectively
2. Generate a simulated random sample of $s_1^*, \ldots, s_{1000}^*$ as those done in Example 4.75
3. Calculate the sample mean $\overline{s}^*$ $(= 1.6188)$, the the bias in is approximately estimated by

$$\mu_{s^*} - \sigma \approx \overline{s}^* - \sigma^* = 1.6188 - 1.7137 = -0.0947$$

(Parametric) Bootstrap Methods (2/2)

- Bootstrap results can sometimes be used to adjust estimates to make them more accurate
- **Example 4.77**: in **Example 4.76**, a sample of size 6 was taken from an $N(\mu,\sigma^2)$ population. The sample standard deviation $s = 1.7137$ is an estimate of the unknown population standard deviation σ .
  - Use the bootstrap result in Example 4.76 to reduce the bias in this estimate

Answer

  - The bias in s is -0.049, which means that on average, the sample standard deviation computed from the $N(\mu,\sigma^2)$ population is less than the true standard deviation σ by about -0.0949
  - We hence can adjust for the bias by adding 0.0949 to the estimate to have bias-corrected estimate $s' = 1.7137 + 0.0949 \cong 1.81$

Nonparametric Bootstrap

- If we have a sample $X_1,\ldots,X_n$ from an unknown distribution, we will simulate samples $X_{1i}^*,\ldots,X_{ni}^*$ as follows:
  1. Image placing the values $X_1,\ldots,X_n$ in a box, and drawing out one value at random. Then replace the value and draw again. Continue until n draws have been made to form the first bootstrap sample $X_{11}^*,\ldots,X_{n1}^*$
     - It will probably contains some of the original sample items more than once, and others not at all
  2. Draw more bootstrap samples !

Law of Large Numbers

- Let $X_1,\ldots,X_n$ be a sequence of independent random variables with $\mathbf{E}[X_i]=\mu$ and $\text{var}(X_i)=\sigma^2$. Let $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$. Then, for any $\varepsilon>0$,

$$P\left(\left(\overline{X}-\mu\right)\geq\varepsilon\right)\leq\frac{\text{var}\left(\overline{X}\right)}{\varepsilon^2}=\frac{\sigma^2}{n\varepsilon^2}\to 0,\quad \text{as } n\to 0$$

$$\mathbf{E}\left[\overline{X}\right]=\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\mu$$

$$\text{var}\left(\overline{X}\right)=\frac{1}{n^2}\sum_{i=1}^{n}\text{var}\left(X_i\right)=\frac{\sigma^2}{n}\quad (\text{since } X_i \text{ are independent})$$

The desired result follows immediately from Chebyshev's inequality, which states that,

$$P\left(\left(X-\mu_X\right)\geq\varepsilon\right)\leq\frac{\sigma_X^2}{\varepsilon^2}\quad \text{for } \varepsilon>0$$