Problem 1.

I. Which of the following is a *random* variable?

- a) The population variance
- ✓ b) The sample mean
- c) The variance of the sample mean
- ✓ d) The smallest value in the sample
- ✓ e) The estimated variance of the sample mean
- f) The sample size

II. True or False?

- True g) The center of a 95% confidence interval for the population mean is a random variable
- False h) A 95% confidence interval for μ contains the sample mean with probability 0.95
- False i) A 95% confidence interval contains 95% of the population
- False j) Out of one thousand 95% confidence intervals for μ, 950 will contain μ.

a) The population variance σ^2 is not random ;
it is an unknown constant

b) The sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is random since it depends on random variables $X_1 \dots X_n$

c) The variance of the sample mean, $\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$
is a constant, not a random variable

d) The smallest value in the sample $X_{(1)} = \min\{X_1 \dots X_n\}$
is random since depends on $X_1 \dots X_n$

e) The estimated variance of the sample mean

$$s^2_{\overline{X}_n} = \frac{Nn-n}{Nn-N}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{n}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)$$ is random

f) The sample size n is a constant

g) Confidence interval is random
$\Rightarrow$ its center is also random

$$\boxed{\begin{array}{l} \text{a) d) e)} \\ \text{g) is true} \\ \text{h) i) j) are false} \end{array}}$$

h) i) j) are False, since
a $100(1-\alpha)\%$ confidence interval contains μ with probability $(1-\alpha)$

Problem 2. Consider a population of size N=5, the members of which have values x_1, x_2, x_3, x_4, x_5.

1) If simple random sampling were used, how many samples of size n=3 are there?

2) Suppose that rather than simple random sampling, the following sampling scheme is used. The possible samples of size n=3 are

$$\{x_1,x_2,x_3\},\{x_2,x_3,x_4\},\{x_3,x_4,x_5\},\{x_4,x_5,x_1\},\{x_5,x_1,x_2\}$$

and the sampling is done in such a way that each of these 5 possible samples is equally likely. Is the sample mean unbiased?

1) How many times we can pick 3 objects out of 5 ?

We can simply list all of them

$$\left.\begin{matrix} x_1\ x_2\ x_3 \\ x_1\ x_2\ x_4 \\ x_1\ x_2\ x_5 \\ x_1\ x_3\ x_4 \\ x_1\ x_3\ x_5 \\ x_1\ x_4\ x_5 \\ x_2\ x_3\ x_4 \\ x_2\ x_3\ x_5 \\ x_2\ x_4\ x_5 \\ x_3\ x_4\ x_5 \end{matrix}\right\}$$

$$\# = \binom{5}{3} = \frac{5!}{3!\,2!} = \frac{1\cdot 2\cdot 3\cdot 4\cdot 5}{1\cdot 2\cdot 3\cdot 1\cdot 2} = 10$$

or, we can use "n choose k" formula

2) Let us compute the expected value of the sample mean :

$$\mathbb{E}[\bar{X}_n] = \frac{x_1+x_2+x_3}{3}\cdot \overbrace{\mathbb{P}(\{x_1 x_2 x_3\} \text{ is chosen})}^{1/5} +$$

$$+ \frac{x_2+x_3+x_4}{3}\overbrace{\mathbb{P}(\{x_2,x_3,x_4\} \text{ is chosen})}^{1/5} + \dots + $$

$$+ \frac{x_5+x_1+x_2}{3}\cdot \overbrace{\mathbb{P}(\{x_5,x_1,x_2\} \text{ is chosen})}^{1/5} =$$

$$= \frac{1}{5}\,\frac{1}{3}\left([x_1+x_2+x_3]+[x_2+x_3+x_4]+\dots+[x_5+x_1+x_2]\right) =$$

$$= \frac{1}{5}\cdot\frac{1}{3}\cdot 3\,(x_1+x_2+x_3+x_4+x_5)$$

for every i, there 3 terms x_i in the sum

$$= \frac{x_1+x_2+x_3+x_4+x_5}{5} = \mu$$

$$\boxed{\begin{array}{l} 1)\ \# = 10 \\ 2)\ \text{Yes} \end{array}}$$

Problem 3. Suppose X is a random variable with mean μ and variance σ^2. Use the delta method to approximate mean and variance of $Y = \log(X)$.

Hint: To approximate $E[Y]$, use the Taylor series expansion of $g(x) = \log(x)$ at $x = \mu$ to the second order; to approximate $Var[Y]$, use the Taylor series expansion of $g(x) = \log(x)$ at $x = \mu$ to the first order.

1. Taylor expansion of $g(x) = \log(x)$ at μ :

$$g(x) \approx g(\mu) + (x-\mu) g'(\mu) + \frac{1}{2}(x-\mu)^2 g''(\mu)$$

$$g'(x) = \frac{1}{x}$$

$$g''(x) = -\frac{1}{x^2}$$

$$\boxed{g(x) \approx \log \mu + (x-\mu)\frac{1}{\mu} - \frac{1}{2}(x-\mu)^2 \frac{1}{\mu^2}}$$

2. Mean

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] \approx \mathbb{E}\left[\log\mu + (X-\mu)\frac{1}{\mu} - \frac{1}{2\mu^2}(X-\mu)^2\right]$$

$$= \log\mu - \frac{1}{2\mu^2}\underbrace{\mathbb{E}\left[(X-\mu)^2\right]}_{\sigma^2} = \log\mu - \frac{\sigma^2}{2\mu^2}$$

3. Variance

$$\mathbb{V}[Y] = \mathbb{V}[g(X)] \approx \mathbb{V}\left[\log\mu + (X-\mu)\frac{1}{\mu}\right]$$

$$= \mathbb{V}\left[\frac{X}{\mu} - 1\right] = \frac{1}{\mu^2}\mathbb{V}[X] = \frac{\sigma^2}{\mu^2}$$

$$\boxed{\begin{aligned} \mathbb{E}[Y] &\approx \log\mu - \frac{\sigma^2}{2\mu^2} \\ \mathbb{V}[Y] &\approx \sigma^2/\mu^2 \end{aligned}}$$

Problem 4. An investigator quantifies her uncertainty about the estimate of a population mean by reporting $\overline{X}_n \pm 2s_{\overline{X}_n}$. What size confidence interval is this?

Hint: In may help to know the following value $\Phi(2) \approx 0.977$, where Φ is of the standard normal CDF.

An approximate $100(1-\alpha)\%$ confidence interval for μ is

$$\overline{X}_n \pm z_{\frac{\alpha}{2}} \sigma_{\overline{X}_n}$$

$s^2_{\overline{X}_n}$ is an unbiased estimate of $\sigma^2_{\overline{X}_n}$.

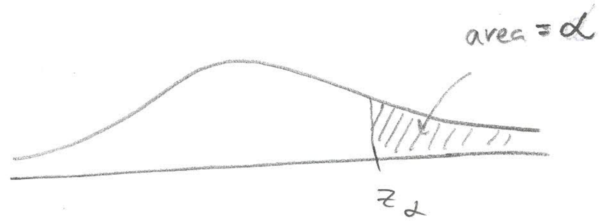
Thus, in practice, $s_{\overline{X}_n}$ is used instead of $\sigma_{\overline{X}_n}$ (since $\sigma_{\overline{X}_n}$ is typically unknown)

The reported interval is

$$\overline{X}_n \pm 2\, s_{\overline{X}_n}$$

To find α, we have to solve: $z_{\frac{\alpha}{2}} = 2$ for α.

This equation is equivalent to



$$\Phi(z_{\frac{\alpha}{2}}) = \Phi(2)$$

$$1 - \frac{\alpha}{2} = \Phi(2)$$

$$\alpha = 2 \cdot (1 - \Phi(2)) \approx 0.046$$

$1 - \Phi(z_\alpha) = \alpha$

$\Rightarrow$ $\overline{X}_n \pm 2\, s_{\overline{X}_n}$ is an approximate 95.4% confidence interval

$$\boxed{\text{size} = 95.4\%}$$

Problem 5. The following table (Cochran 1977) shows the stratification of all farms in a county by farm size and the mean and standard deviations of the number of acres of corn in each stratum.

Farm Size	N_k	μ_k	σ_k
0-40	394	5.4	8.3
41-80	461	16.3	13.3
81-120	391	24.3	15.1
121-160	334	34.5	19.8
161-200	169	42.1	24.5
201-240	113	50.1	26.0
241 +	148	63.8	35.2

a) For a sample size of n=100 farms, compute the sample sizes from each stratum for proportional and optimal allocations.
b) Calculate the variances of the sample mean for each allocation and compare them to each other

a) For proportional allocation : $\tilde{n}_K = n \cdot w_K = n \cdot \frac{N_K}{N} = n \cdot \frac{N_K}{\sum_{K=1}^{7} N_K}$

For optimal allocation : $\hat{n}_K = n \frac{w_K \sigma_K}{\sum_{j=1}^{7} w_j \sigma_j}$

Using values from the table we obtain :

k	1	2	3	4	5	6	7
$\tilde{n}_K$	$19.6 \approx \boxed{20}$	$22.9 \approx \boxed{23}$	$19.45 \approx \boxed{19}$	$16.6 \approx \boxed{17}$	$8.4 \approx \boxed{8}$	$5.6 \approx \boxed{6}$	$7.36 \approx \boxed{7}$
$\hat{n}_K$	$9.56 \approx \boxed{10}$	$17.9 \approx \boxed{18}$	$17.26 \approx \boxed{17}$	$19.3 \approx \boxed{19}$	$12.1 \approx \boxed{12}$	$8.6 \approx \boxed{9}$	$15.2 \approx \boxed{15}$

b) $$\mathbb{V}[\bar{X}^*_{n,opt}] = \frac{1}{n}\left(\sum_{K=1}^{7} w_K \sigma_K\right)^2 \approx \underline{2.9}$$

$$\mathbb{V}[\bar{X}^*_{n,p}] = \frac{1}{n}\sum_{K=1}^{7} w_K \sigma_K^2 \approx \underline{3.4}$$

$$\mathbb{V}[\bar{X}^*_{n,p}] > \mathbb{V}[\bar{X}^*_{n,opt}]$$

$\tilde{n}_K = [20, 23, 19, 17, 8, 6, 7]$

$\hat{n}_K = [10, 18, 17, 19, 12, 9, 15]$

$\mathbb{V}[\bar{X}^*_{n,p}] = 3.4$ $\quad$ $\mathbb{V}[\bar{X}^*_{n,opt}] = 2.9$