STAT 8670: Assignment 4

Fall 2026 | Computational Methods in Statistics

Chi-Kuang Yeh

STAT 8670 | Fall 2026
Due date: Friday, November 20, 2026
Total: 30 points
Coverage: Chapter 5: Monte Carlo Simulation and Variance Reduction, with a guided introduction to Chapter 6: Cross Validation. The additional practice also reviews Chapter 4: Generating Random Variables.

Submit **Questions 1-4 only (30 points)**. Questions 1-3 cover Chapter 5 [24 points]; Question 4 introduces Chapter 6 with a supplied reading guide [6 points]. P1-P10 are additional practice: **not graded and not required for submission**. Optional and supplementary sections of the notes are not required.

Exam preparation: We recommend working through this assignment before **Exam 2 on November 16**, together with Assignment 3 and the main material in Chapters 4-6. The submission deadline is **November 20, after the exam**. These exercises support review; they are not an exhaustive list of exam topics or a promise of identical questions.

Question	Topic / format	Points
1	Monte Carlo concepts (8 multiple-choice items)	8
2	Variance reduction (8 multiple-choice items)	8
3	Precision and interpretation (4 short-answer parts)	8
4	Guided preview: cross-validation (3 short parts)	6
Total		**30**
P1-P10	Additional Exam 2 practice	Not graded

Assignment PDF | Response template (.qmd)

Submission instructions

Use **Quarto** or **R Markdown**. Upload **both** files to the **Assignment 4** folder on **iCollege**:

1. Your rendered **PDF**, named `A4_LastName_FirstName.pdf`.
2. Its matching source, named `A4_LastName_FirstName.qmd` or `A4_LastName_FirstName.Rmd`.

The PDF must be generated from the submitted source. **HTML alone does not satisfy the submission requirement.**

- Put your name and course number at the beginning, and label every question and subpart. Use the response template or follow its labeling format; you do not need to copy a table.
- For Questions 1-2, enter exactly one letter for each item. For Question 3, show a brief formula or substitution with each calculation, and add one sentence of interpretation in part (b). For Question 4, follow the requested response type; part (b) needs one sentence. Equivalent correct reasoning is accepted.
- Round numerical answers to **three decimal places**, retaining full precision during calculations. Report counts as integers and leave exact symbolic quantities in symbolic form.
- No programming or plots are required. If you use R, include executable chunks after your answers and set `set.seed(8670)` once before each independent simulation; do not reset it within repetitions.
- Your source must render from beginning to end without errors. Use LaTeX for mathematics; do not submit screenshots of answers or code.

Question 1: Monte Carlo Concepts [8]

Select **one best answer** for each item. Each item is worth **1 point**. Assume the expectations and variances used below are finite. Here s_Y denotes the sample standard deviation of the outputs Y_i.

(a) The target [1]

Let $U \sim \text{Uniform}(0, 1)$, $Y = g(U)$, and $\theta = E[Y]$. What does the average of many independent $g(U_i)$ values estimate?

A. The variance of U.
B. The expectation $E[g(U)]$.
C. The largest possible value of $g(U)$.
D. The value $g(E[U])$, which always equals $E[g(U)]$.

(b) Output spread and simulation precision [1]

An estimate averages $N > 1$ independent, identically distributed outputs $Y_i = g(U_i)$ with nonzero variance. Which estimates its Monte Carlo standard error (MCSE)?

A. $s_U/\sqrt{N}$.
B. s_Y.
C. s_Y/N.
D. $s_Y/\sqrt{N}$.

(c) A larger simulation [1]

With the same output distribution, approximately how many independent draws are needed to reduce MCSE from 0.03 to 0.01 if the current run uses 2,000 draws?

A. 18,000.
B. 6,000.
C. 2,000.
D. 200.

(d) An integral [1]

To estimate $I = \int_2^5 g(x)\,dx$, generate independent $X_i \sim \text{Uniform}(2, 5)$. Which estimator is correct?

A. $\text{mean}(g(X_i))/3$.
B. $\text{mean}(X_i)$.
C. $3\,\text{mean}(g(X_i))$.
D. $\text{mean}(g(X_i))$.

(e) A probability [1]

Independent draws $X_1, \ldots, X_N$ follow the specified model. Let $I_i = 1$ if $X_i > 8$ and $I_i = 0$ otherwise. Which expression estimates $P(X > 8)$?

A. $N^{-1} \sum_{i=1}^{N} I_i$.
B. The average of only those X_i that exceed 8.
C. $\sum_{i=1}^{N} I_i$.
D. The indicator that the sample mean of the X_i exceeds 8.

(f) An uncertainty interval [1]

A large simulation of independent, identically distributed outputs reports $\hat{\theta} \pm 1.96\widehat{\text{MCSE}}$. What uncertainty does this interval address?

A. The spread of individual simulated outputs.
B. All uncertainty in the scientific model and its assumptions.
C. Monte Carlo error in estimating the specified model expectation.
D. Only measurement error in the original dataset.

(g) Reproducibility [1]

A script sets `set.seed(8670)` once before a simulation. Which statement is correct?

A. The seed guarantees the estimate equals the target.
B. The seed makes the simulation error zero.
C. Every independently chosen seed gives the same estimate.
D. The seed supports reproducibility; MCSE is still needed to assess precision.

(h) A missing factor [1]

An estimate of $\int_2^5 g(x)\,dx$ uses $N^{-1} \sum_{i=1}^{N} g(X_i)$ for independent $X_i \sim \text{Uniform}(2, 5)$, omitting the interval-width factor. Assume the desired integral is nonzero. What does increasing N accomplish?

A. It automatically supplies the missing factor.
B. It estimates the wrong scaled target more precisely.
C. It makes the output distribution uniform.
D. It guarantees a larger estimated MCSE.

Question 2: Variance Reduction [8]

Select **one best answer** for each item. Each item is worth **1 point**. Use the conventions from Chapter 5. Assume finite second moments; s_A is the sample standard deviation of the K pair averages, with $K > 1$.

(a) Antithetic inputs [1]

For $U \sim \text{Uniform}(0, 1)$, which statement about $(U, 1 - U)$ is correct?

A. The two values are independent.
B. Their sum is random.
C. Each is uniform, but they are dependent.
D. Only the first value is uniform.

(b) The independent unit [1]

From K independent uniforms, form $A_k = \{g(U_k) + g(1 - U_k)\}/2$. Which estimates the MCSE of $\overline{A}$?

A. $s_A/\sqrt{K}$.
B. $s_A/\sqrt{2K}$.
C. The SD of all $2K$ outputs divided by $\sqrt{2K}$, treating them as independent.
D. $s_U/\sqrt{K}$.

(c) A fair cost comparison [1]

Each call to g has the same cost. An antithetic run uses 600 pairs. Which plain Monte Carlo run has the same number of calls to g?

A. 300 independent outputs.
B. 600 independent outputs.
C. 2,400 independent outputs.
D. 1,200 independent outputs.

(d) When reflection helps [1]

At the same number of calls to g, antithetic pairing reduces estimator variance relative to independent outputs when:

A. $g(U)$ and $g(1 - U)$ have positive covariance.
B. $g(U)$ and $g(1 - U)$ have negative covariance.
C. U is resampled until it exceeds 0.5.
D. g is any function, without further conditions.

(e) The sign of a control coefficient [1]

Use $Z = Y + c(C - \mu_C)$ with known $\mu_C = E[C]$ and $\mathrm{Var}(C) > 0$. If $\mathrm{Cov}(Y, C) > 0$, the variance-minimizing fixed coefficient c^* is:

A. Positive.
B. Zero.
C. Negative.
D. Always equal to 1.

(f) Preserving the target [1]

For a fixed coefficient c, why does $Z = Y + c(C - \mu_C)$ have the same expectation as Y?

A. $E[C - \mu_C] = 0$.
B. C is always zero.
C. C and Y must be independent.
D. Z must equal Y for every draw.

(g) Replacing a known mean [1]

A programmer uses $Z_i = Y_i + c(C_i - \overline{C})$ with the same sample mean $\overline{C}$ and the same fixed c in every row. What happens to $\overline{Z}$?

A. It always becomes zero.
B. Its variance must be smaller than that of the original sample mean.
C. It becomes the known control mean.
D. It equals the original sample mean of the outputs.

(h) No association with the control [1]

If $\mathrm{Cov}(Y, C) = 0$ and $\mathrm{Var}(C) > 0$, which fixed coefficient minimizes $\mathrm{Var}\{Y + c(C - \mu_C)\}$?

A. $c = 1$.
B. $c = 0$.
C. $c = -1$.
D. Every value gives the same variance.

Question 3: Precision and Interpretation [8]

Two runs estimate the same $\theta = E[g(U)]$, where $U \sim \text{Uniform}(0,1)$. Each evaluates g exactly **1,200 times**, and each evaluation has the same cost. The plain run averages independent outputs; the antithetic run averages independent pair averages A_k.

Method	Independent contributions	Estimate	Sample SD of contributions
Plain	$N = 1200$ outputs	1.721	0.600
Antithetic	$K = 600$ pair averages	1.718	0.120

Treat the displayed SDs as exact for these calculations. Retain full precision until the final answer.

(a) MCSE [2]

Report the estimated MCSE for **plain** and **antithetic** Monte Carlo, in that order. Each value is worth **1 point**.

(b) Relative precision [2]

Compute the ratio of estimated **estimator variances**, $\widehat{\text{Var}}(\hat{\theta}_{\text{plain}})/\widehat{\text{Var}}(\hat{\theta}_{\text{anti}})$ [**1**]. In one sentence, identify the more precise estimator and interpret this variance ratio at the stated computational budget (do not call it an MCSE ratio) [**1**].

(c) A Monte Carlo interval [2]

Using the antithetic result, give the lower and upper endpoints of the approximate 95% interval $\hat{\theta} \pm 1.96\widehat{\text{MCSE}}$ (**1 point each**).

(d) A separate control-variate calculation [2]

In a separate simulation, $E[C] = 2$, $\text{Var}(C) = 0.25$, and $\text{Cov}(Y, C) = 0.50$. Use the fixed optimal coefficient in

$$Z = Y + c(C - 2), \qquad c^* = -\frac{\text{Cov}(Y, C)}{\text{Var}(C)}.$$

The run gives $\overline{Y} = 3.4$ and $\overline{C} = 2.1$. Report c^* and the adjusted estimate $\overline{Z}$ (**1 point each**).

Question 4: Guided Preview: Cross-Validation [6]

New material. Read Chapter 6: prediction on new observations and K-fold CV.

Reading guide. Split the development data into K folds. For each candidate and each fold, fit using the other folds and predict the held-out observations. Average their squared errors to obtain the CV MSE. Use the same folds for all candidates. Select the candidate with the smallest CV MSE, refit it on all development observations, and assess it once on a separate test set that was not used for selection. Data-dependent preprocessing and feature selection belong inside each training fold.

(a) Data roles [2]

There are 100 independent development observations in five equal folds. For each fit, how many observations are used for **training** and for **validation**, respectively (**1 point each**)?

(b) Choose a candidate [2]

Three polynomial regression candidates use the same five folds and squared-error loss. Training MSE is calculated after fitting each candidate to all 100 development observations; CV MSE uses held-out predictions from the five separate training fits:

Degree	Training MSE	CV MSE
1	9.0	10.0
2	4.0	6.0
6	1.0	12.0

Which degree should be selected by CV [**1**]? In one sentence, explain why the smallest training MSE does not determine this choice [**1**].

(c) Valid assessment [2]

Enter `T` or `F` for each statement (**1 point each**).

1. Selecting features using all development outcomes before forming the folds gives a valid CV assessment of that feature-selection procedure.
2. After choosing the degree by CV, it is appropriate to refit on all development data and evaluate once on an untouched test set.

Additional Exam 2 Practice (Not Graded)

These ten questions complement Questions 1-4 and Assignment 3. They practice core concepts in new situations; the longer explanations are for learning and do not prescribe the length of an exam answer. They are **not an additional submission requirement** and do not change the 30-point assignment total. No programming is required; use short explanations and calculations where requested.

- P1-P5: random-variable generation and checking (Chapter 4).
- P6-P8: Monte Carlo precision and variance reduction (Chapter 5).
- P9-P10: validation design and prediction loss (Chapter 6, with reading guides).

Vectors are bold and are column vectors. Gamma parameters are shape and rate unless stated otherwise. Exclude the optional and supplementary material in the notes.

P1. Inverse transformation

A lifetime X has CDF

$$F(x) = \begin{cases} 0, & x \le 0, \\ x^2/4, & 0 < x < 2, \\ 1, & x \ge 2. \end{cases}$$

For $U \sim \text{Uniform}(0, 1)$, select the correct generator. Check your choice by substituting it into F.

A. $X = 2U^2$.
B. $X = \sqrt{U}/2$.
C. $X = 2\sqrt{U}$.
D. $X = 4U$.

P2. Construct a distribution from normal draws

Let $Z_1, \ldots, Z_4$ be independent standard normal variables and define

$$T = \frac{Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2}{6}.$$

Which distribution does T have? Here $\text{Gamma}(a, b)$ uses shape a and **rate** b.

A. $\text{Gamma}(2, 3)$.
B. $\text{Gamma}(4, 6)$.
C. $\text{Gamma}(2, 1/3)$.
D. χ^2_4.

P3. Check the joint distribution

Let $\mathbf{Z} = (Z_1, Z_2)^\top \sim N_2(\mathbf{0}, I_2)$ and define

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} \mathbf{Z}, \qquad \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{5} \end{pmatrix} \mathbf{Z}.$$

Which statement is correct? Treat the vectors as columns.

A. The two vectors have the same joint distribution because their coordinate variances agree.
B. X_1 and X_2 are independent because Z_1 and Z_2 are independent.
C. $\mathrm{Var}(X_2) = 4$, whereas $\mathrm{Var}(W_2) = 5$.
D. Both vectors have coordinate variances 1 and 5, but $\mathrm{Cov}(X_1, X_2) = 1$ and $\mathrm{Cov}(W_1, W_2) = 0$.

Write the two covariance matrices as a check.

P4. Proposal support

A rejection sampler targets $\mathrm{Uniform}(0, 1)$ and proposes $Y \sim \mathrm{Exp}(1)$ on $(0, \infty)$.

Mark **T or F** and give one sentence: "The proposal is unusable because it can generate values greater than 1."

For this question, you may use the valid envelope constant $M = e$: the target density is 1 on $(0, 1)$ and zero elsewhere, while the proposal density is $g(y) = e^{-y}$ for $y > 0$.

P5. Rejection-sampling efficiency

Two rejection samplers use the same **normalized** target density f and proposal density g. Both envelope constants are valid: $M_A = 1.5$ and $M_B = 3$, meaning $f(y) \le Mg(y)$ throughout the target support. Each sampler uses 600 independent proposals, with the same cost per proposal and fresh independent acceptance draws.

1. Calculate each sampler's overall acceptance probability and expected number of accepted draws.
2. Which sampler gives more accepted draws per proposal, and by what factor? Does the larger valid envelope change the target distribution of accepted draws?

Guide. For normalized target and proposal densities, the overall acceptance probability is $1/M$. An expected count need not equal the count in a particular run.

P6. Combine independent simulation batches

Two independent batches contain $n_1 = 100$ and $n_2 = 400$ independent outputs from the same distribution, with mean θ and variance 8. Let their means be $\overline{Y}_1$ and $\overline{Y}_2$. Compare

$$\hat{\theta}_A = \tfrac{1}{2}\overline{Y}_1 + \tfrac{1}{2}\overline{Y}_2, \qquad \hat{\theta}_B = \tfrac{1}{5}\overline{Y}_1 + \tfrac{4}{5}\overline{Y}_2.$$

1. Are both estimators unbiased? Explain using their expectations.
2. Calculate both estimator variances. Which uses the 500 outputs more efficiently, and why?
3. Does the smaller estimator variance guarantee a closer estimate to θ in every pair of realized runs? Explain briefly.

Use $\mathrm{Var}(aV + bW) = a^2\,\mathrm{Var}(V) + b^2\,\mathrm{Var}(W)$ for independent V, W.

P7. Evaluate a proposed pairing

Each paired run averages K independent pair averages $(Y_k + Y_k')/2$. Both members have the correct target marginal distribution and variance 1. Their within-pair covariance differs:

Pairing	$\mathrm{Cov}(Y_k, Y_k')$
A	-0.6
B	0.4

Each method uses $2K$ function evaluations. The plain comparison averages $2K$ independent outputs of variance 1.

1. Calculate $\mathrm{Var}(\hat{\theta}_{\mathrm{paired}})/\mathrm{Var}(\hat{\theta}_{\mathrm{plain}})$ separately for A and B. Which pairing improves precision?
2. A paired-to-plain ratio of 0.4 for estimator **variance** is reported as a paired-to-plain MCSE ratio of 0.4. Is that conversion correct? Give the correct ratio.

Use $\mathrm{Var}\{(Y + Y')/2\} = \{\mathrm{Var}(Y) + \mathrm{Var}(Y') + 2\,\mathrm{Cov}(Y, Y')\}/4$.

P8. Can a larger adjustment be worse?

Suppose $E[Y] = \theta$, $E[C] = \mu_C$ is known, $\mathrm{Var}(Y) = 9$, $\mathrm{Var}(C) = 4$, and $\mathrm{Cov}(Y, C) = 3$. Use

$$Z_c = Y + c(C - \mu_C).$$

Compare the two fixed coefficients $c = -0.5$ and $c = -2$.

1. State the expectation and variance of Z_c for each coefficient.
2. A student argues that making c more negative must keep improving precision because the covariance is positive. Use your results to assess this claim at the same number of independent draws.

You may use $\mathrm{Var}(Z_c) = \mathrm{Var}(Y) + c^2\,\mathrm{Var}(C) + 2c\,\mathrm{Cov}(Y, C)$.

P9. Match validation to the prediction task

Reading guide. A validation split should resemble the future prediction task. For a new group, keep its observations out of training; for forecasting, train on earlier times and validate on later times. Learn preprocessing from the training portion of each split. See Chapter 6: choosing folds and avoiding leakage.

A retailer records monthly sales for several stores. Address these two different goals:

1. To predict the next month's sales at the **existing stores**, should validation randomly mix past and future months, or hold out later months and train on earlier months? Explain briefly.
2. For the separate goal of predicting sales at a **new store during the same calendar period** (not forecasting future months), should rows from the same store be spread across training and validation, or kept in the same fold? Assume stores are the independent sampling units. Explain briefly.
3. Within a CV fit, a predictor is standardized using a mean and SD estimated from **all** development rows before splitting. State where those quantities should instead be estimated and how they should be applied to the validation rows.

P10. Aggregate held-out loss

A candidate model specification is fixed. It is refitted separately on the other two folds each time, giving the following held-out results. Each observation appears in exactly one validation fold:

Fold	Number of validation observations	Sum of squared prediction errors
1	2	2
2	3	9
3	5	4

1. Calculate the overall CV MSE, giving every validation observation equal weight.
2. A student instead averages the three fold MSEs without weights. Calculate that result and explain why it answers a different averaging question.

Guide. Add the squared held-out errors over all observations and divide by their total count. Unequal folds require size weights if you start from fold MSEs.