

Variance Estimation and Importance Sampling

Week 2 problem class Questions and worked solutions

Junqi Tang

Based on the lecture notes developed by Prof. Jinglai Li

What we will practise

1. Distinguish a population variance from an estimator of it.
2. Explain the correction from n to $n - 1$.
3. Calculate two variance estimates from the same data.
4. Estimate one expectation using two sampling distributions.
5. Check importance weights, support and numerical answers.

Two questions to keep asking

What is the fixed target? What is the random quantity being averaged?

Question 1(a): is this estimator unbiased?

Let $X_1, \dots, X_n$ be iid with

$$\mathbb{E}[X_i] = \mu, \quad 0 < \text{Var}(X_i) = \sigma^2 < \infty, \quad n \geq 2.$$

Consider the following *estimator of the population variance*:

$$\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Show that $\tilde{S}^2$ is biased and calculate its bias.

Try first 3 minutes

Expand the square and use $\sum_i X_i = n\bar{X}$.

A useful identity for any finite-variance variable

For $m = \mathbb{E}[Y]$,

$$\begin{aligned}\text{Var}(Y) &= \mathbb{E}[(Y - m)^2] \\ &= \mathbb{E}[Y^2] - 2m\mathbb{E}[Y] + m^2 \\ &= \mathbb{E}[Y^2] - m^2.\end{aligned}$$

Hence

$$\boxed{\mathbb{E}[Y^2] = \text{Var}(Y) + (\mathbb{E}[Y])^2.}$$

No Gaussian assumption

This is an algebraic identity. It applies to any random variable with finite second moment.

Solution 1(a): simplify the estimator

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2.\end{aligned}$$

Divide by n :

$$\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2.$$

We now need $\mathbb{E}[X_i^2]$ and $\mathbb{E}[\bar{X}^2]$.

Solution 1(a): take expectations

Identical distributions and independence give

$$\mathbb{E}[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}.$$

Therefore

$$\mathbb{E}[X_i^2] = \sigma^2 + \mu^2, \quad \mathbb{E}[\bar{X}^2] = \frac{\sigma^2}{n} + \mu^2.$$

Substitute into the expanded estimator:

$$\begin{aligned} \mathbb{E}[\tilde{S}^2] &= \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2 \right) \\ &= \boxed{\frac{n-1}{n} \sigma^2}. \end{aligned}$$

What does downward bias mean?

$$\text{Bias}(\tilde{S}^2) = \mathbb{E}[\tilde{S}^2] - \sigma^2 = \boxed{-\frac{\sigma^2}{n}}.$$

- ▶ The estimator underestimates variance *on average over repeated samples*.
- ▶ An individual estimate can still exceed the true variance.
- ▶ The target σ^2 is fixed; $\tilde{S}^2$ is random before observing data.

Why the bias appears

The sample mean is fitted to the same data. Squared deviations around this fitted centre are smaller than around the true mean.

Question 1(b): correct the bias

Define

$$S^2 = \frac{n}{n-1} \tilde{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Show that S^2 is unbiased for σ^2 .

Pause 1 minute

Use the expectation found in part (a). No new expansion is needed.

Solution 1(b): the correction cancels the bias

$$\mathbb{E}[S^2] = \frac{n}{n-1} \mathbb{E}[\tilde{S}^2] = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \boxed{\sigma^2}.$$

Another useful identity is

$$\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Taking expectations gives $n\sigma^2 - \sigma^2 = (n-1)\sigma^2$.

Be precise

S^2 is unbiased for the variance. Its square root S is generally *not* unbiased for the standard deviation.

Question 1(c): calculate both estimates

For the data

$\{10, 15, 30, 45, 57, 52, 63, 72, 81, 93, 102, 105\}$,

calculate $\tilde{S}^2$ and S^2 .

1. Find n and $\bar{x}$.
2. Calculate the squared-deviation sum $\sum_i (x_i - \bar{x})^2$.
3. Divide by n and by $n - 1$, respectively.

Try first 3 minutes

A convenient check is $\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2$.

Solution 1(c): numerical values

$$n = 12, \quad \sum_i x_i = 725, \quad \sum_i x_i^2 = 54995, \quad \bar{x} = \frac{725}{12} \approx 60.4167.$$

$$\sum_i (x_i - \bar{x})^2 = 54995 - \frac{725^2}{12} \approx 11192.9167.$$

Estimator	Calculation	Estimate
$\tilde{S}^2$	11192.9167/12	932.7431
S^2	11192.9167/11	1017.5379

The estimates satisfy $S^2 = (12/11)\tilde{S}^2$.

Question 2: one target and two sampling methods

Let $X \sim U[-1, 1]$, $f(x) = e^{-x^2/2}$ and $I = \mathbb{E}[f(X)]$.

Given uniform samples

0.1812, -0.6245, -0.0422, 0.2049, 0.3150

Given standard normal samples

0.1772, -0.4802, 1.2499, 0.7698, -1.3827

1. Estimate I using direct Monte Carlo and the uniform samples.
2. Estimate the same I using importance sampling and the normal samples.

Write the target density explicitly

For $X \sim U[-1, 1]$,

$$p(x) = \frac{1}{2} \mathbf{1}_{\{|x| \leq 1\}}.$$

Thus

$$I = \mathbb{E}_p[f(X)] = \frac{1}{2} \int_{-1}^1 e^{-x^2/2} dx.$$

- ▶ p is the probability density; f is the function being averaged.
- ▶ The expectation is half the unweighted integral over $[-1, 1]$.
- ▶ Sampling from p already accounts for this density factor.

Do not multiply a direct estimate of this expectation by the interval length.

Direct Monte Carlo: evaluate and average

x_i	$f(x_i) = e^{-x_i^2/2}$
0.1812	0.983717
-0.6245	0.822835
-0.0422	0.999110
0.2049	0.979227
0.3150	0.951598

$$\hat{I}_{\text{MC}} = \frac{1}{5} \sum_{i=1}^5 f(x_i) = \boxed{0.947297 \dots \approx 0.9473}.$$

Calculations use unrounded values; displayed entries are rounded.

Importance sampling: change the sampling density

Draw $Z_i \sim N(0, 1)$ with density

$$q(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

The correct output is $H_i = f(Z_i)p(Z_i)/q(Z_i)$, so

$$\hat{l}_{\text{IS}} = \frac{1}{5} \sum_{i=1}^5 H_i.$$

Support check

q is positive wherever p is positive. Samples outside $[-1, 1]$ are allowed: their target density, and hence their weight, is zero.

The weight is *target density divided by proposal density*.

Simplify before calculating

$$w(z) = \frac{p(z)}{q(z)} = \frac{\sqrt{2\pi}}{2} e^{z^2/2} \mathbf{1}_{\{|z| \leq 1\}},$$

$$\begin{aligned} h(z) &= f(z)w(z) \\ &= e^{-z^2/2} \frac{\sqrt{2\pi}}{2} e^{z^2/2} \mathbf{1}_{\{|z| \leq 1\}} \\ &= \boxed{\frac{\sqrt{2\pi}}{2} \mathbf{1}_{\{|z| \leq 1\}}}. \end{aligned}$$

Pause 2 minutes

How many of the five normal samples lie in $[-1, 1]$? Use that count to calculate the estimate.

Importance sampling: the corrected numerical table

z_i	$f(z_i)$	$w(z_i)$	$h(z_i) = f(z_i)w(z_i)$
0.1772	0.984423	1.273146	1.253314
-0.4802	0.891102	1.406476	1.253314
1.2499	0.457891	0	0
0.7698	0.743567	1.685542	1.253314
-1.3827	0.384455	0	0

Three draws contribute the same nonzero output:

$$\hat{l}_{\text{IS}} = \frac{3}{5} \frac{\sqrt{2\pi}}{2} = \boxed{0.751988 \dots \approx 0.7520}.$$

The value 0.5830 in the supplied solution is a numerical error.

Why do we divide by five rather than three?

Ordinary importance sampling averages *all* proposal outputs:

$$\frac{1}{5}(H_1 + H_2 + H_3 + H_4 + H_5).$$

- ▶ A zero output is still an observation from the proposal.
- ▶ Discarding zero outputs and dividing by three changes the procedure.
- ▶ Dividing by the observed weight sum is self-normalised importance sampling, a different estimator.

Check the expectation

For $c = \sqrt{2\pi}/2$ and $Z \sim N(0, 1)$,

$$\mathbb{E}_q[h(Z)] = c \Pr(|Z| \leq 1) = I.$$

Compare the estimates with a benchmark

Let Φ denote the standard normal CDF. Then

$$I = \frac{\sqrt{2\pi}}{2} [\Phi(1) - \Phi(-1)] \approx 0.855624.$$

Method	Estimate	Absolute error
Direct Monte Carlo	0.947297	0.091673
Importance sampling	0.751988	0.103636

One run is not an efficiency comparison

These are two realised errors from five samples each. Compare theoretical variances or many independent runs to assess accuracy.

Check your understanding

Discuss for two minutes:

1. Does the proof of unbiased sample variance require Gaussian data?
2. Is a downward-biased estimate always below the true variance?
3. Why does $z = 1.2499$ have weight zero in Question 2?
4. Does an unbiased importance sampler always improve accuracy?

Answers: No; no; $p(z) = 0$ outside $[-1, 1]$; no.

Main results to take away

$$\mathbb{E}[\tilde{S}^2] = \frac{n-1}{n}\sigma^2, \quad \mathbb{E}[S^2] = \sigma^2.$$

For the data in Question 1(c):

$$\tilde{S}^2 \approx 932.7431, \quad S^2 \approx 1017.5379.$$

For the samples in Question 2:

$$\hat{l}_{\text{MC}} \approx 0.9473, \quad \hat{l}_{\text{IS}} \approx 0.7520.$$

Reliable calculations

State the target and assumptions, simplify the weighted output, and check numerical values before rounding.

Extension: does this proposal reduce variance?

Write $r = \Pr(|Z| \leq 1) = \Phi(1) - \Phi(-1) \approx 0.682689$. Since $H = c\mathbf{1}_{\{|Z| \leq 1\}}$ and $c = \sqrt{2\pi}/2$,

$$\text{Var}_q(H) = c^2 r(1 - r) \approx 0.340273.$$

For direct sampling,

$$\text{Var}_p(f(X)) = \frac{1}{2} \int_{-1}^1 e^{-x^2} dx - I^2 \approx 0.014731.$$

At the same sample size, the IS estimator has about

$$\frac{0.340273}{0.014731} \approx 23.1$$

times the variance. Unbiasedness alone does not imply efficiency.

Extension: why the normal proposal is inefficient here

Direct output

$$e^{-1/2} \leq f(X) \leq 1.$$

Every draw contributes a moderately sized value.

The relevant shape is $f(x)p(x)$, which vanishes outside $[-1, 1]$.

Importance-weighted output

$$H \in \{0, 1.253314 \dots\}.$$

About 31.7% of proposal draws produce zero output.

Ideal shape

$q^*(x) = f(x)p(x)/I$ is a standard normal density truncated to $[-1, 1]$. It gives constant weighted outputs, but its normalising constant involves the unknown target.

Extension: knowing the mean changes the estimator

If the true mean μ is known, then

$$V_{\mu} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

is already unbiased:

$$\mathbb{E}[V_{\mu}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i - \mu)^2] = \sigma^2.$$

The reason for the correction

The $n - 1$ denominator compensates for estimating the centre from the same sample. It is not a universal denominator for every variance calculation.

Corrections to the supplied solution

- ▶ Question 1(a): say “estimator of population variance” and state the moment assumptions.
- ▶ Question 1(c): include both numerical estimates.
- ▶ Question 2: write $Z \sim N(0, 1)$ and specify its density q .
- ▶ Recalculate normal-sample function values and importance weights.
- ▶ Replace the reported IS estimate 0.5830 by 0.751988....

Source: supplied Week 2 problem sheet (headed “Problem Set 1”), Questions 1–2 and solutions. Original sample values are retained. Extension slides add interpretation of the corrected results.