Conditional expectation and the law of total variance

Ten problems on conditional expectation: E[Y | X = x] from a joint table, E[Y | X] as a random variable with its own mean and variance, the law of total variance on the same table, the mean and variance of a two-component Gaussian mixture, a random sum of purchases with a Poisson count, E[Y | X] from the joint density x + y, E[Y | X] as the best mean-squared-error predictor with a residual uncorrelated with every function of X, the conditional mean and variance of a bivariate Gaussian, Rao–Blackwell variance reduction for a Monte Carlo estimate, and the noise that dropout adds split from the spread of the data, with worked solutions and the mistakes of averaging the conditional variances only, evaluating E[Y | X] at the mean of X, inverting a regression line, taking E[N]σ² as the variance of a random sum and averaging conditional means without their weights.

BEFORE YOU START

Conditional expectation answers the question "what is the average of Y once I know X?". For each value x it is a number, the mean of Y in the slice $X = x$; taken over all values of X at once it is a random variable, a function of X. That second view is what makes it useful. Averaging it gives back $E[Y]$, its spread and the spread left inside each slice add up to $\mathrm{Var}(Y)$, and it is the best possible prediction of Y from X under squared error, which is why a regression network trained on mean-squared error is learning $E[Y \mid X]$. The same two laws compute the moments of mixtures and random sums, give the regression line of a Gaussian, and explain why replacing a noisy sample by its conditional mean, as Rao–Blackwellization and many gradient estimators do, can only lower the variance. These ten problems work each of them by hand.

- For discrete variables, $E[Y \mid X = x] = \sum_y y \Pr(Y = y \mid X = x)$, with $\Pr(Y = y \mid X = x) = \Pr(X = x, Y = y) / \Pr(X = x)$. For densities, $E[Y \mid X = x] = \int y\, f(x, y)\, dy \,/\, f_X(x)$.
- Writing $g(x) = E[Y \mid X = x]$, the conditional expectation $E[Y \mid X]$ is the random variable $g(X)$: it is not a number until a value of X is plugged in.
- **Tower law.** $E\big[E[Y \mid X]\big] = E[Y]$, and more generally $E[h(X)\,Y] = E\big[h(X)\,E[Y \mid X]\big]$ for any function h.
- **Law of total variance.** $\mathrm{Var}(Y) = E\big[\mathrm{Var}(Y \mid X)\big] + \mathrm{Var}\left(E[Y \mid X]\right)$: the average spread within the slices plus the spread of the slice means.
- **Best predictor.** Among all functions g, $E[(Y - g(X))^2]$ is smallest for $g(X) = E[Y \mid X]$, and the residual $Y - E[Y \mid X]$ has mean 0 given X, so it is uncorrelated with every $h(X)$.

- **Random sums.** If $S = X_1 + \cdots + X_N$ with the X_i independent of N and of each other, each with mean μ and variance σ^2, then $E[S] = E[N]\mu$ and $\mathrm{Var}(S) = E[N]\sigma^2 + \mathrm{Var}(N)\mu^2$.
- **Bivariate Gaussian.** $E[Y \mid X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ and $\mathrm{Var}(Y \mid X) = \sigma_Y^2(1 - \rho^2)$, the same for every x.

Builds on: Discrete random variables: expectation, variance and the binomial, geometric and Poisson, Continuous random variables: densities, CDFs, expectation and the uniform, exponential and normal, Variance, covariance and correlation

Problems

1. The joint distribution of X and Y is difficulty ·

	$Y=1$	$Y=2$	$Y=3$
$X=0$	0.2	0.1	0.1
$X=1$	0.1	0.2	0.3

Find $E[Y \mid X = 0]$, $E[Y \mid X = 1]$ and $E[Y]$, and check that the conditional means average to $E[Y]$.

2. For the table of Problem 1, write $E[Y \mid X]$ as a function of X. Find its mean and variance, and use the tower law to get $E[XY]$ and $\mathrm{Cov}(X, Y)$. difficulty ·

3. For the table of Problem 1, find $\mathrm{Var}(Y \mid X = 0)$ and $\mathrm{Var}(Y \mid X = 1)$, and check the law of total variance against $\mathrm{Var}(Y)$ computed directly. difficulty ··

4. A model draws each output from a two-component mixture: with probability 0.3 from $N(0, 1)$, and with probability 0.7 from $N(4, 2^2)$. Find the mean and variance of the output. difficulty ··

5. The number of customers in an hour is Poisson with mean 4, and each customer spends an amount with mean 20 dollars and standard deviation 5 dollars, independently of the others and of how many come. Find the mean and standard deviation of the hour's revenue S. difficulty ··

6. X and Y have joint density $f(x,y) = x + y$ on the unit square $0 \le x, y \le 1$. Find f_X, the conditional mean $E[Y \mid X = x]$ and its values at $x = 0$ and $x = 1$, and check the tower law. difficulty ··

7. For the table of Problem 1, compare three predictors of Y from X by mean-squared error: the constant $E[Y]$, the conditional mean $E[Y \mid X]$, and $g(X) = 2 + 0.5X$. Show that the residual $R = Y - E[Y \mid X]$ is uncorrelated with X. difficulty ···

8. Heights X (cm) and weights Y (kg) are bivariate Gaussian with $\mu_X = 170$, $\sigma_X = 10$, $\mu_Y = 70$, $\sigma_Y = 12$ and correlation $\rho = 0.5$. Find the mean and standard deviation of the weight of a person 185 cm tall, and check the law of total variance. difficulty ··

9. To estimate $E[Y]$ where $X \sim \text{Uniform}(0,1)$ and $Y \mid X \sim N(X, 1)$, one can average $n = 100$ samples of Y, or average $E[Y \mid X_i] = X_i$ over 100 samples of X. Find the variance of one term of each estimator, their standard errors, and how many samples of Y match the second estimator's accuracy. difficulty ···

10. A unit computes $s = \sum_i m_i w_i x_i / 0.8$ with weights $w = (1, -2, 3)$ and independent dropout masks $m_i \sim \text{Bernoulli}(0.8)$. The input is $x_a = (2, 1, 1)$ or $x_b = (1, 1, 0)$ with probability $\frac{1}{2}$ each. Find $E[s \mid x]$ and $\text{Var}(s \mid x)$ for each input, then $E[s]$ and $\text{Var}(s)$, and the share of $\text{Var}(s)$ that dropout causes. difficulty ···

Answers

1. $E[Y \mid X=0] = 1.75$, $E[Y \mid X=1] = \frac{7}{3} \approx 2.333$, $E[Y] = 0.4 \cdot 1.75 + 0.6 \cdot \frac{7}{3} = 2.1$
2. $E[Y \mid X] = \frac{7}{4} + \frac{7}{12}X$, with mean 2.1 and variance $\frac{49}{600}$; $E[XY] = 1.4$ and $\operatorname{Cov}(X,Y) = 1.4 - 0.6 \cdot 2.1 = 0.14$
3. $\operatorname{Var}(Y \mid X=0) = \frac{11}{16}$, $\operatorname{Var}(Y \mid X=1) = \frac{5}{9}$, and $0.6083 + 0.0817 = 0.69 = \operatorname{Var}(Y)$
4. $E[Y] = 2.8$ and $\operatorname{Var}(Y) = 3.1 + 3.36 = 6.46$, a standard deviation of about 2.54
5. $E[S] = 80$ dollars and $\operatorname{Var}(S) = 100 + 1600 = 1700$, so the standard deviation is about 41.23 dollars
6. $f_X(x) = x + \frac{1}{2}$, $E[Y \mid X=x] = \frac{3x+2}{6x+3}$, and $E\big[E[Y \mid X]\big] = \int_0^1 \left(\frac{x}{2} + \frac{1}{3}\right)dx = \frac{7}{12} = E[Y]$
7. The errors are 0.69, $\frac{73}{120} \approx 0.608$ and 0.65; $E[RX] = E\big[X\,E[R \mid X]\big] = 0$, and $E[R] = 0$, so $\operatorname{Cov}(R,X) = 0$
8. Mean 79 kg, standard deviation $\sqrt{108} \approx 10.39$ kg
9. Variances $\frac{13}{12}$ against $\frac{1}{12}$, standard errors 0.104 against 0.029; matching it takes $13 \cdot 100 = 1300$ samples of Y
10. $E[s] = 1$ and $\operatorname{Var}(s) = 2.75 + 4 = 6.75$; dropout causes $\frac{2.75}{6.75} = \frac{11}{27} \approx 41\%$ of it

Worked solutions

Problem 1

> The joint distribution of X and Y is
>
> | | $Y=1$ | $Y=2$ | $Y=3$ |
> |---|---|---|---|
> | $X=0$ | 0.2 | 0.1 | 0.1 |
> | $X=1$ | 0.1 | 0.2 | 0.3 |
>
> Find $E[Y \mid X=0]$, $E[Y \mid X=1]$ and $E[Y]$, and check that the conditional means average to $E[Y]$.

$\Pr(X=0)=0.4,\ \Pr(X=1)=0.6$	Row sums of the table.
$E[Y \mid X=0] = \dfrac{1\cdot 0.2 + 2\cdot 0.1 + 3\cdot 0.1}{0.4} = \dfrac{0.7}{0.4} = 1.75$	Divide the row by its total to get the conditional distribution, then take its mean.
$E[Y \mid X=1] = \dfrac{1\cdot 0.1 + 2\cdot 0.2 + 3\cdot 0.3}{0.6} = \dfrac{1.4}{0.6} = \dfrac{7}{3}$	The same in the second row.
$E[Y] = 0.7 + 1.4 = 2.1$	The column totals are 0.3, 0.3, 0.4, and $1\cdot 0.3 + 2\cdot 0.3 + 3\cdot 0.4 = 2.1$; the two numerators of steps 2 and 3 add to the same thing.
$E[Y \mid X=0] = 1.75$, $E[Y \mid X=1] = \frac{7}{3} \approx 2.333$, $E[Y] = 0.4\cdot 1.75 + 0.6\cdot \frac{7}{3} = 2.1$	The tower law: each conditional mean is weighted by how likely its slice is. The plain average of 1.75 and $\frac{7}{3}$ is $\frac{49}{24} \approx 2.042$ (Mistake 5).

Problem 2

> For the table of Problem 1, write $E[Y \mid X]$ as a function of X. Find its mean and variance, and use the tower law to get $E[XY]$ and $\operatorname{Cov}(X,Y)$.

$E[Y \mid X] = \frac{7}{4} + \frac{7}{12}X$	X takes only the values 0 and 1, so a line through $(0, 1.75)$ and $(1, \frac{7}{3})$ matches both: $\frac{7}{3} - \frac{7}{4} = \frac{7}{12}$.

$E\big[E[Y \mid X]\big] = \frac{7}{4} + \frac{7}{12} \cdot 0.6 = 2.1$	$E[X] = 0.6$; this is $E[Y]$ again.
$\operatorname{Var}\big(E[Y \mid X]\big) = \left(\frac{7}{12}\right)^2 \operatorname{Var}(X) = \frac{49}{144} \cdot 0.24 = \frac{49}{600} \approx 0.0817$	$\operatorname{Var}(X) = 0.6 \cdot 0.4$ for a $0/1$ variable; the constant $\frac{7}{4}$ does not change the variance.
$E[XY] = E\big[X\,E[Y \mid X]\big] = 0.6 \cdot \frac{7}{3} = 1.4$	$X\,E[Y \mid X]$ is 0 when $X = 0$ and $\frac{7}{3}$ when $X = 1$. Directly: $1 \cdot (0.1 \cdot 1 + 0.2 \cdot 2 + 0.3 \cdot 3) = 1.4$.
$E[Y \mid X] = \frac{7}{4} + \frac{7}{12}X$, **with mean** 2.1 **and variance** $\frac{49}{600}$; $E[XY] = 1.4$ **and** $\operatorname{Cov}(X, Y) = 1.4 - 0.6 \cdot 2.1 = 0.14$	The slope is $\operatorname{Cov}(X, Y) / \operatorname{Var}(X) = 0.14/0.24 = \frac{7}{12}$, the regression slope, as it must be when the conditional mean is a straight line.

Problem 3

> For the table of Problem 1, find $\operatorname{Var}(Y \mid X = 0)$ and $\operatorname{Var}(Y \mid X = 1)$, and check the law of total variance against $\operatorname{Var}(Y)$ computed directly.

$E[Y^2 \mid X = 0] = \dfrac{0.2 + 0.4 + 0.9}{0.4} = 3.75,$ so $\operatorname{Var}(Y \mid X = 0) = 3.75 - 1.75^2 = \frac{11}{16}$	$1^2 \cdot 0.2 + 2^2 \cdot 0.1 + 3^2 \cdot 0.1 = 1.5$, divided by 0.4.
$E[Y^2 \mid X = 1] = \dfrac{0.1 + 0.8 + 2.7}{0.6} = 6,$ so $\operatorname{Var}(Y \mid X = 1) = 6 - \frac{49}{9} = \frac{5}{9}$	The second row: $3.6/0.6$.
$E\big[\operatorname{Var}(Y \mid X)\big] = 0.4 \cdot \frac{11}{16} + 0.6 \cdot \frac{5}{9} = 0.275 + \frac{1}{3} = \frac{73}{120} \approx 0.6083$	The within-slice spread, weighted by the slice probabilities.
$E\big[\operatorname{Var}(Y \mid X)\big] + \operatorname{Var}\big(E[Y \mid X]\big) = \frac{73}{120} + \frac{49}{600} = \frac{414}{600} = 0.69$	The between-slice spread from Problem 2.
$\operatorname{Var}(Y \mid X = 0) = \frac{11}{16}$, $\operatorname{Var}(Y \mid X = 1) = \frac{5}{9}$, **and** $0.6083 + 0.0817 = 0.69 = \operatorname{Var}(Y)$	Directly, $E[Y^2] = 1.5 + 3.6 = 5.1$ and $5.1 - 2.1^2 = 0.69$. Stopping at step 3 gives 0.6083, which is too small (Mistake 1).

Problem 4

A model draws each output from a two-component mixture: with probability 0.3 from $N(0, 1)$, and with probability 0.7 from $N(4, 2^2)$. Find the mean and variance of the output.

Let Z be the component, so $E[Y \mid Z] \in \{0, 4\}$ and $\mathrm{Var}(Y \mid Z) \in \{1, 4\}$	Condition on the hidden label; given it, Y is a single Gaussian.
$E[Y] = 0.3 \cdot 0 + 0.7 \cdot 4 = 2.8$	The tower law over Z.
$E\big[\,\mathrm{Var}(Y \mid Z)\big] = 0.3 \cdot 1 + 0.7 \cdot 4 = 3.1$	The average within-component variance. Note $2^2 = 4$: the second component has standard deviation 2.
$\mathrm{Var}\left(E[Y \mid Z]\right) = 0.3 \cdot 0.7 \cdot (4 - 0)^2 = 3.36$	A two-valued variable taking values a and b with probabilities p and $1 - p$ has variance $p(1 - p)(b - a)^2$.
$E[Y] = 2.8$ **and** $\mathrm{Var}(Y) = 3.1 + 3.36 = 6.46$, **a standard deviation of about** 2.54	Check with second moments: $E[Y^2] = 0.3 \cdot 1 + 0.7 \cdot (4 + 16) = 14.3$, and $14.3 - 2.8^2 = 6.46$. The mixture is wider than either component because its centers are far apart.

Problem 5

The number of customers in an hour is Poisson with mean 4, and each customer spends an amount with mean 20 dollars and standard deviation 5 dollars, independently of the others and of how many come. Find the mean and standard deviation of the hour's revenue S.

$E[S \mid N] = 20N$ and $\mathrm{Var}(S \mid N) = 25N$	Given $N = n$, S is a sum of n independent amounts: means add and variances add.
$E[S] = E[20N] = 20 \cdot 4 = 80$	The tower law.
$E\big[\,\mathrm{Var}(S \mid N)\big] = 25\,E[N] = 100$	The spread of the amounts, for the average number of customers.

$\operatorname{Var}\big(E[S \mid N]\big) = 20^2 \operatorname{Var}(N) = 400 \cdot 4 = 1600$	A Poisson count has variance equal to its mean; this term is the uncertainty in how many customers come.
$E[S] = 80$ **dollars and** $\operatorname{Var}(S) = 100 + 1600 = 1700$, **so the standard deviation is about** 41.23 **dollars**	Most of the risk is in the count, not the amounts. Treating the count as fixed at 4 gives a standard deviation of 10 dollars (Mistake 4).

Problem 6

> X and Y have joint density $f(x, y) = x + y$ on the unit square $0 \le x, y \le 1$. Find f_X, the conditional mean $E[Y \mid X = x]$ and its values at $x = 0$ and $x = 1$, and check the tower law.

$f_X(x) = \int_0^1 (x + y)\, dy = x + \frac{1}{2}$	Integrate out y.
$\int_0^1 y(x + y)\, dy = \frac{x}{2} + \frac{1}{3}$	The numerator of the conditional mean.
$E[Y \mid X = x] = \dfrac{x/2 + 1/3}{x + 1/2} = \dfrac{3x + 2}{6x + 3}$	Multiply top and bottom by 6.
$E[Y \mid X = 0] = \frac{2}{3}$ and $E[Y \mid X = 1] = \frac{5}{9} \approx 0.556$	At larger x the density $x + y$ is flatter in y, so the conditional mean moves toward $\frac{1}{2}$.
$f_X(x) = x + \frac{1}{2}$, $E[Y \mid X = x] = \frac{3x+2}{6x+3}$, **and** $E\big[E[Y \mid X]\big] = \int_0^1 \left(\frac{x}{2} + \frac{1}{3}\right) dx = \frac{7}{12} = E[Y]$	In the tower law the $f_X(x)$ in the denominator cancels the weight $f_X(x)\, dx$. Directly, $E[Y] = \int_0^1 y(y + \frac{1}{2})\, dy = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$. Plugging $E[X] = \frac{7}{12}$ into the curve gives $\frac{15}{26}$ instead (Mistake 2).

Problem 7

> For the table of Problem 1, compare three predictors of Y from X by mean-squared error: the constant $E[Y]$, the conditional mean $E[Y \mid X]$, and $g(X) = 2 + 0.5X$. Show that the residual $R = Y - E[Y \mid X]$ is uncorrelated with X.

$E[(Y - 2.1)^2] = \operatorname{Var}(Y) = 0.69$	Problem 3.

$E\big[(Y - E[Y \mid X])^2\big] = E\big[\operatorname{Var}(Y \mid X)\big] = \frac{73}{120} \approx 0.6083$	Inside each slice the squared error around the slice mean is the slice variance.
$E[(Y - g(X))^2] = 0.4 \cdot 0.75 + 0.6 \cdot \frac{7}{12} = 0.65$	Given $X = 0$: $(0.2 \cdot 1 + 0.1 \cdot 0 + 0.1 \cdot 1)/0.4 = 0.75$. Given $X = 1$: $(0.1 \cdot 2.25 + 0.2 \cdot 0.25 + 0.3 \cdot 0.25)/0.6 = \frac{7}{12}$.
$0.65 = \frac{73}{120} + E\big[(E[Y \mid X] - g(X))^2\big] = \frac{73}{120} + \frac{1}{24}$	$E[Y \mid X] - g(X)$ is -0.25 at $X = 0$ and $-\frac{1}{6}$ at $X = 1$: $0.4 \cdot 0.0625 + 0.6 \cdot \frac{1}{36} = \frac{1}{24}$. The cross term vanishes because $E[R \mid X] = 0$.
The errors are 0.69, $\frac{73}{120} \approx 0.608$ **and** 0.65; $E[RX] = E\big[X\,E[R \mid X]\big] = 0$, **and** $E[R] = 0$, **so** $\operatorname{Cov}(R, X) = 0$	Every other predictor pays the extra $E[(E[Y \mid X] - g(X))^2]$. The share of $\operatorname{Var}(Y)$ that X explains is $\frac{49/600}{0.69} = \frac{49}{414} \approx 0.118$.

Problem 8

> Heights X (cm) and weights Y (kg) are bivariate Gaussian with $\mu_X = 170$, $\sigma_X = 10$, $\mu_Y = 70$, $\sigma_Y = 12$ and correlation $\rho = 0.5$. Find the mean and standard deviation of the weight of a person 185 cm tall, and check the law of total variance.

$E[Y \mid X = x] = 70 + 0.5 \cdot \frac{12}{10}(x - 170) = 70 + 0.6(x - 170)$	The slope is $\rho\sigma_Y/\sigma_X = \operatorname{Cov}(X, Y)/\operatorname{Var}(X)$.
$E[Y \mid X = 185] = 70 + 0.6 \cdot 15 = 79$	1.5 standard deviations above the mean in height, but only 0.75 above in weight.
$\operatorname{Var}(Y \mid X) = 144(1 - 0.25) = 108$	It does not depend on x for a Gaussian.
$\operatorname{Var}\big(E[Y \mid X]\big) = 0.6^2 \cdot 100 = 36$, and $108 + 36 = 144 = \sigma_Y^2$	The regression line explains $\rho^2 = 25\%$ of the variance.

Mean 79 **kg, standard deviation** $\sqrt{108} \approx 10.39$ **kg**	Knowing the height narrows the spread from 12 kg only to 10.39 kg at $\rho = 0.5$. The reverse question, the expected height of a 79 kg person, is not 185 cm (Mistake 3).

Problem 9

To estimate $E[Y]$ where $X \sim \text{Uniform}(0, 1)$ and $Y \mid X \sim N(X, 1)$, one can average $n = 100$ samples of Y, or average $E[Y \mid X_i] = X_i$ over 100 samples of X. Find the variance of one term of each estimator, their standard errors, and how many samples of Y match the second estimator's accuracy.

$E[Y] = E[X] = \frac{1}{2}$	Both estimators are unbiased, by the tower law.
$\text{Var}(Y) = E[1] + \text{Var}(X) = 1 + \frac{1}{12} = \frac{13}{12} \approx 1.083$	Total variance: the noise around X plus the spread of X.
$\text{Var}\left(E[Y \mid X]\right) = \text{Var}(X) = \frac{1}{12} \approx 0.0833$	Conditioning removes the within-slice term entirely.
$\sqrt{\frac{13/12}{100}} \approx 0.1041$ and $\sqrt{\frac{1/12}{100}} \approx 0.0289$	The standard error of an average of n independent terms is $\sqrt{\text{Var}\,/n}$.
Variances $\frac{13}{12}$ **against** $\frac{1}{12}$**, standard errors** 0.104 **against** 0.029**; matching it takes** $13 \cdot 100 = 1300$ **samples of** Y	This is Rao–Blackwellization: replacing a sample by its conditional mean keeps the mean and can only lower the variance, by exactly $E[\text{Var}(Y \mid X)]$. Baselines and marginalized gradient estimators use the same idea.

Problem 10

A unit computes $s = \sum_i m_i w_i x_i / 0.8$ with weights $w = (1, -2, 3)$ and independent dropout masks $m_i \sim \text{Bernoulli}(0.8)$. The input is $x_a = (2, 1, 1)$ or $x_b = (1, 1, 0)$ with probability $\frac{1}{2}$ each. Find $E[s \mid x]$ and $\text{Var}(s \mid x)$ for each input, then $E[s]$ and $\text{Var}(s)$, and the share of $\text{Var}(s)$ that dropout causes.

$E[s \mid x] = w^\top x$	$E[m_i / 0.8] = 1$, so dropout with this rescaling leaves the mean unchanged.

$\mathrm{Var}(s \mid x) = \dfrac{0.2}{0.8}\sum_i w_i^2 x_i^2$	$\mathrm{Var}(m_i/0.8) = 0.8 \cdot 0.2/0.8^2 = 0.25$, and the masks are independent.
x_a: $E = 2 - 2 + 3 = 3$, $\mathrm{Var} = 0.25(4 + 4 + 9) = 4.25$; x_b: $E = 1 - 2 = -1$, $\mathrm{Var} = 0.25(1 + 4) = 1.25$	Plug in each input.
$E\big[\mathrm{Var}(s \mid x)\big] = \frac{1}{2}(4.25 + 1.25) = 2.75$ and $\mathrm{Var}\left(E[s \mid x]\right) = \frac{1}{4}(3 - (-1))^2 = 4$	Equal weights: a two-valued variable with $p = \frac{1}{2}$ has variance $\frac{1}{4}(b - a)^2$.
$E[s] = 1$ **and** $\mathrm{Var}(s) = 2.75 + 4 = 6.75$; **dropout causes** $\frac{2.75}{6.75} = \frac{11}{27} \approx 41\%$ **of it**	The other 59% is the genuine difference between the inputs. At test time dropout is switched off, which removes exactly the first term.

Where this goes wrong

1. Averaging the conditional variances and stopping

The law of total variance has two terms, and the first one already looks like "the variance, averaged over X".

$\mathrm{Var}(Y \mid X = 0) = \frac{11}{16}$ and $\mathrm{Var}(Y \mid X = 1) = \frac{5}{9}$	Right so far: Problem 3.
“$\mathrm{Var}(Y)$ is the average of the conditional variances.”	That is only the within-slice spread. The slice means 1.75 and $\frac{7}{3}$ also differ, and that spread counts too.
$\mathrm{Var}(Y) = 0.4 \cdot \frac{11}{16} + 0.6 \cdot \frac{5}{9} \approx 0.6083$ ✗	The missing term is $\mathrm{Var}(E[Y \mid X]) = \frac{49}{600} \approx 0.0817$, and the right total is 0.69. The averaged conditional variance is never larger than the true variance.

2. E[Y | X] evaluated at the mean of X

$E[Y \mid X]$ is a function of X, and the tower law averages it over X; plugging in a single value of X looks like a shortcut.

$E[Y \mid X = x] = \dfrac{3x+2}{6x+3}$ and $E[X] = \frac{7}{12}$	Right so far: Problem 6, and $E[X] = E[Y]$ by the symmetry of $x + y$.
"$E[Y] = E[Y \mid X = E[X]]$."	That treats the random variable $E[Y \mid X]$ as if it were a fixed number. Averaging a function and evaluating it at the average agree only for straight lines.
$E[Y] = \dfrac{3 \cdot \frac{7}{12} + 2}{6 \cdot \frac{7}{12} + 3} = \frac{15}{26} \approx 0.577$ ✗	The curve is not a line, so the answer is off: $E[Y] = \int_0^1 \left(\frac{x}{2} + \frac{1}{3}\right) dx = \frac{7}{12} \approx 0.583$. In Problem 2 the shortcut happens to work, because there $E[Y \mid X]$ is linear in X.

3. Inverting the regression line

The line $E[Y \mid X = x] = 70 + 0.6(x - 170)$ predicts weight from height, and solving it for x looks like it predicts height from weight.

$E[Y \mid X = 185] = 79$	Right so far: Problem 8.
"So a person who weighs 79 kg is expected to be 185 cm tall."	$E[X \mid Y]$ is a different regression, with slope $\rho\sigma_X/\sigma_Y$, not the reciprocal $1/0.6$.
$E[X \mid Y = 79] = 185$ ✗	$E[X \mid Y = 79] = 170 + 0.5 \cdot \frac{10}{12}(79 - 70) = 173.75$ cm. Both regressions pull toward the mean; that is regression to the mean, and only at $\lvert\rho\rvert = 1$ are the two lines the same.

4. The variance of a random sum taken as E[N]σ²

Variances of independent terms add, so n terms of variance σ^2 give $n\sigma^2$, and $E[N]$ seems the natural n.

$\mathrm{Var}(S \mid N) = 25N$	Right so far: Problem 5.

"Replace N by its mean: $\mathrm{Var}(S) = 25\,E[N]$."	That is only $E[\mathrm{Var}(S \mid N)]$. The count itself is random, and $\mathrm{Var}(20N) = 1600$ is missing.
$\mathrm{Var}(S) = 25 \cdot 4 = 100$, a standard deviation of 10 dollars ✗	The right variance is 1700, a standard deviation of about 41.23 dollars, more than four times as wide. The 10-dollar answer is correct only if exactly 4 customers are certain to come.

5. Averaging conditional means without their weights

There are two slices and two conditional means, and the midpoint looks like the overall mean.

$E[Y \mid X = 0] = 1.75$ and $E[Y \mid X = 1] = \frac{7}{3}$	Right so far: Problem 1.
"$E[Y]$ is the average of the two conditional means."	The tower law weights each slice by its probability, 0.4 and 0.6, not $\frac{1}{2}$ each.
$E[Y] = \frac{1}{2}(1.75 + \frac{7}{3}) = \frac{49}{24} \approx 2.042$ ✗	The right value is $0.4 \cdot 1.75 + 0.6 \cdot \frac{7}{3} = 2.1$, which the column totals confirm. The unweighted average is right only when the slices are equally likely, as in Problem 10.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 10 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/conditional-expectation-and-the-law-of-total-variance