Final Exam Practice Questions, with Answers

EPID 785R: Applied Foundations of Study Design and Regression Modeling, Fall 2026

Ashley I Naimi

What this is. Twenty-five practice questions, one per page, in the form of the pop quizzes: one topic, one small table or one fitted model, two or three short parts. They cover the topics in the final exam study guide; the table below gives each question's week and exam topic numbers. Work them closed book and by hand. Where a question prints an equation or a value of exp, log or a square root, use it; where it prints none, the definitions are part of what the exam expects you to know.

With answers. This copy has the answers at the foot of each page; try each question before reading them.

Conventions. The exposure is X (1 = exposed), the outcome is Y, covariates are C or Z. "Risk" means cumulative incidence over the stated window. A 95% interval is estimate $\pm 1.96\times$ SE throughout. The minutes are a rough guide.

#	Week	Exam topics	Min
1	1	1, 3	9
2	1	2	8
3	2	18	7
4	3	4	7
5	3	5	6
6	3	4	7
7	4	6	8
8	5	7	8
9	5	7	6
10	5	8	7
11	6	19, 11	6
12	7	11, 7	5
13	7	11	6
14	8	9	9
15	8	9, 10	6
16	9	10	7
17	10	12	5
18	10	8, 12	5
19	11	14	5
20	12	13, 19	7
21	12	13	6
22	11	12	6
23	13	15	10
24	13	16	8
25	13	17	7

About 170 minutes of work in all.

Question 1

Week 1. Exam topics 1 and 3. About 9 minutes.

An insurance-claims study emulates a target trial of starting drug D after a new diagnosis of condition E, with all-cause mortality within one year as the outcome. Every patient enters at diagnosis (day 0). The exposure was defined from the pharmacy record as *on drug D at the six-month visit (day 180)*: patients with an activeprescription for D on day 180 were analyzed as the treated arm and all other patients as the untreated arm. Follow-up for mortality ran from day 0 to day 365 in both arms.

The table shows what was observed.

Arm, as analyzed	Patients	Deaths, days 0 to 180	Deaths, days 181 to 365	Deaths by one year
On D at day 180	500	0	50	50
Not on D at day 180	500	50	63	113

(a) Which component of the protocol was assessed with information from after time zero? Name the rule it breaks, and state the direction of the bias in the as-analyzed risk difference, with one sentence on why the first 180 days are the problem.

(b) Compute the one-year risk difference as analyzed. Then move time zero to day 180: using only the patients alive on day 180, compute the risk difference for death between day 181 and day 365.

(c) In one sentence each, say what the landmark analysis in (b) changed about the population and about the question.

Answers. (a) The treatment strategy (classification into arms) component of a target trial emulation protocol was assessed from the day-180 pharmacy record, after time zero at diagnosis. Classification and the start of follow-up do not coincide, and the "alignment rule" is broken. A patient had to survive 180 days to be "on D", so the treated arm holds 180 immortal days per person and none of the early deaths; the as-analyzed RD is biased toward benefit (too negative).

(b) RD = risk(treated) − risk(untreated). As analyzed: $50/500 - 113/500 = 0.100 - 0.226 = -0.126$. Time zero at day 180 (950 alive): $50/500 - 63/450 = 0.100 - 0.140 = -0.040$.

(c) With the landmark analysis, the population changed to the 950 survivors to day 180, not the 1,000 diagnosed. The question being answered by this analysis is: six-month mortality after day 180 by treatment status on day 180, not one-year mortality from diagnosis under "start D".

Question 2

Week 1. Exam topic 2. About 8 minutes.

A trial randomized 800 patients after cardiac surgery to a 12-week supervised exercise program ($X = 1$) or usual care ($X = 0$), 400 per arm. $A = 1$ means the patient completed the program (at least 75% of sessions); some usual-care patients joined an equivalent community program on their own and also count as $A = 1$. The outcome Y is hospital readmission within six months. The table gives the number of patients and the number readmitted in each cell.

Assigned X	Completed A	Patients	Readmitted	Risk
1	1	300	30	0.10
1	0	100	30	0.30
0	1	100	10	0.10
0	0	300	60	0.20

Given. ITT $= P(Y{=}1 \mid X{=}1) - P(Y{=}1 \mid X{=}0)$. Naive as-treated $= P(Y{=}1 \mid A{=}1) - P(Y{=}1 \mid A{=}0)$. Naive per-protocol $= P(Y{=}1 \mid X{=}1, A{=}1) - P(Y{=}1 \mid X{=}0, A{=}0)$.

(a) Compute the three contrasts.

(b) Which contrast does randomization identify, and why? Say in one sentence what comparing the outcome risk in groups defined by A rather than by X reintroduces, naming the arrows in the trial's DAG that do it.

Answers. (a) ITT $= 60/400 - 70/400 = 0.150 - 0.175 = -0.025$. As-treated $= 40/400 - 90/400 = 0.100 - 0.225 = -0.125$. Per-protocol $= 30/300 - 60/300 = 0.100 - 0.200 = -0.100$.

(b) The ITT: X was randomized, so the two arms are exchangeable (with positivity and consistency by design). A was not: grouping by A reintroduces confounding through $C \to A$ and $C \to Y$ (patients who complete differ in prognosis), so the other two are observational contrasts inside a trial.

Question 3

Week 2. Exam topic 18. About 7 minutes.

A cohort study of 20,000 adults aged 60 and over measured alcohol intake once, by questionnaire at enrollment, and coded the exposure as *current drinker* (any alcohol in the past 30 days) versus *non-drinker* (no alcohol in the past 30 days). Participants were followed for 15 years for incident dementia. The abstract concludes: "Current drinking was associated with a 20% lower rate of dementia (hazard ratio 0.80, 95% CI 0.70 to 0.91). These findings support moderate alcohol consumption as a protective behavior in older adults."

(a) State the intervention implied by this exposure coding as precisely as the coding allows, and list what it leaves unspecified.

(b) Describe who the referent group ("non-drinker") can contain in this population. In which direction would that composition move the hazard ratio?

(c) Say what a single baseline measurement can and cannot tell you about timing (prevalent versus newly started drinking; sustained versus one-time drinking), and name one thing the study should have collected or done at enrollment to repair the comparison.

Answers. **(a)** "Drink some alcohol in a 30-day window" versus "none in that window". Unspecified: the dose (one drink or forty), the duration (one month says nothing about the years before or after), when the intervention starts and at what age, the pattern and beverage, and what the non-drinker does instead.

(b) Lifelong abstainers plus former drinkers, some of whom may have stopped because of illness, medication, frailty or early cognitive change (sick quitters). The referent can therefore carry more dementia risk than "never started", which pushes the hazard ratio below 1, toward protection, whatever drinking does.

(c) One measurement gives prevalent status at one moment: it cannot separate a new drinker from a lifelong one, or sustained drinking from one month, so neither "start drinking" nor "keep drinking" has a time zero. Collect drinking history (never, former, current; years; dose) and repeat the measure during follow-up, or restrict the referent to never-drinkers.

Question 4

Week 3. Exam topic 4. About 7 minutes.

A closed cohort of 2,000 people was followed for two years with no loss to follow-up. The table gives, by exposure, the number of people at baseline, the number who became cases, and the person-time they accumulated.

Exposure	People at baseline	Cases	Person-years
Exposed	1,000	200	1,800
Unexposed	1,000	100	1,900

Three investigators then drew three control series from this same cohort:

- Series I: 10% of the people who were still non-cases at the end of follow-up: 80 exposed, 90 unexposed.
- Series II: 10% of the baseline cohort, drawn at enrollment regardless of what happened to them later: 100 exposed, 100 unexposed.
- Series III: one control per 10 person-years of follow-up, drawn from those at risk as the cases occurred: 180 exposed, 190 unexposed.

(a) From the full cohort, compute the risk ratio, the odds ratio, and the rate ratio.

(b) For each of the different control series, compute the case-control odds ratio. Match each result to one of the three cohort quantities in (a) and name the sampling scheme that produced it.

Answers. (a) RR = (200/1000)/(100/1000) = 2.00. OR = (200/800)/(100/900) = 0.250/0.111 = 2.25. Rate ratio = (200/1800)/(100/1900) = 0.111/0.0526 = 2.11.

(b)

Series I: $(200 \times 90)/(100 \times 80) = 2.25 =$ OR; cumulative (survivor) sampling, the controls stand for the non-cases.

Series II: $(200 \times 100)/(100 \times 100) = 2.00 =$ RR; case-base sampling, the controls stand for the baseline cohort (the risk denominators).

Series III: $(200 \times 190)/(100 \times 180) = 2.11 =$ rate ratio; incidence-density sampling, the controls stand for person-time.

Question 5

Week 3. Exam topic 5. About 6 minutes.

A cohort of $N = 5{,}000$ people was followed for one year. To save on measuring the exposure, the investigators used a case-cohort design: a random subcohort of 10% of the cohort was drawn at baseline, and the exposure was measured on everyone in the subcohort and on every case in the full cohort. The table gives what was measured.

	Exposed	Unexposed
Subcohort members	100	400
Cases among the subcohort members	6	8
Cases in the full cohort	60	80

(a) Rebuild the weighted 2×2 table (weighted people and cases, by exposure) and run the design check: what should the weighted people sum to, and do they?

(b) Compute the two weighted risks (in the exposed and unexposed), the risk ratio, and the risk difference.

Answers. **(a)** The subcohort is a 10% sample, so each subcohort non-case stands for 10 people (weight $1/0.10 = 10$); every case is already in the data and counts once (weight 1), inside or outside the subcohort. Exposed: $(100 - 6) \times 10 + 60 = 940 + 60 = 1{,}000$ weighted people, 60 cases. Unexposed: $(400 - 8) \times 10 + 80 = 3{,}920 + 80 = 4{,}000$ weighted people, 80 cases. Check: the weighted people should rebuild the cohort, $N = 5{,}000$, and $1{,}000 + 4{,}000 = 5{,}000$.

(b) Risks $60/1{,}000 = 0.060$ (exposed) and $80/4{,}000 = 0.020$ (unexposed). RR $= 0.060/0.020 = 3.0$. RD $= 0.060 - 0.020 = 0.040$.

Question 6

Week 3. Exam topic 4. About 7 minutes.

In a source population of 1,000,000 people followed for one year, 4,000 developed the outcome. A case-control study enrolled every case (sampling fraction $\pi_1 = 1$) and a 1% random sample of the non-cases ($\pi_0 = 0.01$). Logistic regression of case status on a binary exposure X in the case-control data gave

$$\operatorname{logit} P(\text{case} \mid X) = -1.0 + 0.7\,X.$$

Given. $\beta_0(\text{population}) = \beta_0(\text{sample}) - \log(\pi_1/\pi_0)$. $P = \exp(\eta)/\{1 + \exp(\eta)\}$. $\log(100) = 4.6$; $\exp(0.7) = 2.0$.

(a) Compute the intercept the same model would have in the source population.

(b) Does exp(0.7) from the fitted model need a correction? Why or why not?

Answers. (a) $\beta_0(\text{population}) = -1.0 - \log(1/0.01) = -1.0 - 4.6 = -5.6$.

(b) No. Write the source population as a 2×2 and sample it: every case is kept with probability π_1 and every non-case with probability π_0, whatever the exposure. Each cell of the sample is the population cell times its row's fraction:

	Exposed	Unexposed
Cases	$a \to \pi_1 a$	$c \to \pi_1 c$
Non-cases	$b \to \pi_0 b$	$d \to \pi_0 d$

$$\text{sample OR} = \frac{\pi_1 a \times \pi_0 d}{\pi_0 b \times \pi_1 c} = \frac{\pi_1 \pi_0}{\pi_1 \pi_0} \times \frac{ad}{bc} = \frac{ad}{bc} = \text{population OR}.$$

Each fraction appears once in the numerator and once in the denominator, so with $\pi_1 = 1$ and $\pi_0 = 0.01$, $(a \times 0.01d)/(0.01b \times c) = ad/(bc) = 2.0$. They do not cancel in the case odds: among the unexposed, sample odds $= \pi_1 c/(\pi_0 d) = (\pi_1/\pi_0) \times c/d$, so the log odds, the intercept, shifts by $\log(\pi_1/\pi_0) = 4.6$, which is (a). The cancellation needs sampling that depends on the outcome only; if the fractions differed by exposure column, the ratio in front of $ad/(bc)$ would not be 1.

Question 7

Week 4. Exam topic 6. About 8 minutes.

A hospital fit a logistic regression of 30-day readmission on enrollment in a new discharge-planning program (X), age, sex and diabetes, in 4,000 discharged patients. The program was offered preferentially to patients with diabetes, and the prevalence of readmission in the data was roughly 30%. The coefficients for the model's fit were:

Term	Estimate	SE	p
Intercept	-0.80	0.07	< 0.001
Program (X)	-0.40	0.12	0.001
Age (per 10 years, centered at 65)	0.15	0.05	0.003
Male	0.10	0.11	0.36
Diabetes	0.05	0.10	0.62

The program's director then used this fit to estimate the program's effect on readmission in their hospital.

(a) A colleague proposes dropping diabetes "because $p = 0.62$" and refitting. Describe both the (i) causal and (ii) statistical consequences that might arise as a result of dropping diabetes and refitting the model to estimate the program's effect.

(b) The draft report calls every coefficient in the table an "independent risk factor for readmission," and reports the coefficients and standard errors as "effects." Name the fallacy entailed by this practice, and describe the specific problems with doing so.

Answers. **(a)** (i) Causal: the program was offered preferentially to patients with diabetes, and diabetes raises readmission risk, so diabetes is very likely a common cause of X and Y, a confounder. Conditional exchangeability for the program's effect then requires it in the adjustment set; drop it and the program coefficient absorbs the diabetes-readmission association that enrollment carries, whatever its p-value. The $p = 0.62$ tests the conditional diabetes-readmission association in this sample given the other terms; it says nothing about whether diabetes causes enrollment or readmission, or about how far the program coefficient moves without it. (ii) Statistical: dropping a nuisance term can lower the variance of the program coefficient, since diabetes and the program are strongly associated by design and one fewer parameter is estimated. For the effect estimate there is therefore a bias-variance tradeoff: a small confounding bias might be worth a smaller variance (a lower mean squared error), especially with few events per parameter. But $p = 0.62$ is not a valid way to make that trade: the confounding bias depends on the diabetes-exposure association as much as on the outcome coefficient, and the p-value measures neither the bias removed nor the variance saved. The decision belongs to the causal structure (a DAG) and, if a trade is wanted, to a criterion aimed at the exposure coefficient itself (its change and its standard error with and without the term, or shrinkage of the nuisance term), not to a significance test on a nuisance coefficient.

(b) The Table 2 fallacy. The non-exposure rows are the nuisance function, estimated to adjust the program coefficient, not to be read as effects: each is a conditional association given the program and the other terms, with no adjustment set of its own. Age and sex were never adjusted for their own confounders. The diabetes coefficient is conditional on the program, which was targeted at diabetics and may carry part of the effect of diabetes, so 0.05 is a conditional association, not the effect of diabetes, and conditioning on the program may open other paths. One adjustment set identifies one effect; "independent risk factor" claims a causal effect with everything else held fixed, which the fit does not establish, and the standard errors then state the precision of quantities that are not effects.

Question 8

Week 5. Exam topic 7. About 8 minutes.

A Poisson regression of the number of hospitalizations Y over each person's follow-up time T (in years) on an exposure X (0/1) and age gave

$$\log E(Y \mid X, \text{age}, T) = -4.0 + 0.70\,X + 0.03\,(\text{age} - 50) + \log T,$$

where the coefficient of $\log T$ was fixed at 1 rather than estimated.

Given. $\exp(0.70) = 2.0$; $\exp(-3.0) = 0.050$. Expected count = rate $\times T$.

(a) Label the five parts of this model: link, distribution, target function, nuisance function, and offset.

(b) What does $\exp(0.70)$ report? Provide a full interpretation, including the units of the contrast.

(c) For an exposed 60-year-old followed for two years, compute the expected hospitalization rate (per person-year) and the expected number of hospitalizations. At what age does the age term contribute nothing?

(d) In one sentence, what would $\exp(0.70)$ report if $\log T$ were left out of the model entirely?

Answers. (a) Link: log. Distribution: Poisson (variance equal to the mean). Target function: the exposure term $0.70\,X$, the log rate ratio the question is about. Nuisance function: the intercept -4.0 and the age term $0.03(\text{age}-50)$, there to adjust, not to be read. Offset: $\log T$, a covariate whose coefficient is fixed at 1 rather than estimated; moving it to the left turns the model for the expected count into a model for the rate, $\log\{E(Y)/T\}$.

(b) $\exp(0.70) = 2.0$ is an age-adjusted rate ratio: among people of the same age, the exposed have 2.0 times the hospitalization rate of the unexposed, two hospitalizations per person-year for every one in the unexposed. Each rate is in hospitalizations per person-year, so the contrast is a ratio of two rates and has no units; the model makes it the same at every age and over all of follow-up.

(c) Linear predictor $-4.0 + 0.70 + 0.03 \times 10 = -3.0$; rate $= \exp(-3.0) = 0.050$ per person-year; expected count over two years $= 0.050 \times 2 = 0.10$ hospitalizations. The age term is zero at age 50, the centering value.

(d) A ratio of expected counts over whatever follow-up each group accumulated, so the rate ratio multiplied by the ratio of the groups' follow-up times: equal to 2.0 only if exposed and unexposed were followed equally long, and inflated or deflated otherwise.

Question 9

Week 5. Exam topic 7. About 6 minutes. (Pairs with Question 12.)

A logistic regression of an outcome Y on an exposure X (0/1), age, sex ($F = 1$ for women) and an exposure-by-sex product term gave

$$\text{logit}\, P(Y{=}1) = -1.5 + 0.8\, X - 0.04\, (\text{age} - 60) + 0.5\, F - 0.3\, XF.$$

(a) Compute the predicted probability of the outcome for an exposed woman aged 70.

(b) Calculate the odds ratio for both men and women using the coefficients from the model.

(c) Say in one sentence what the product-term coefficient of -0.3 measures. In other words, provide an interpretation of exp(-0.3) from this model.

Answers. (a) Linear predictor for $X = 1$, age 70, $F = 1$: $\eta = -1.5 + 0.8 - 0.04 \times 10 + 0.5 - 0.3 = -0.9$. $P = \exp(-0.9)/\{1 + \exp(-0.9)\} = 0.41/1.41 = 0.29$.

(b) Men ($F = 0$): the exposure log odds ratio is 0.8, so OR $= \exp(0.8) = 2.23$. Women ($F = 1$): the log odds ratio is $0.8 - 0.3 = 0.5$, so OR $= \exp(0.5) = 1.65$. Both hold at every age, since the model has no age-by-exposure term.

(c) -0.3 is the difference between the women's and the men's exposure log odds ratios, so $\exp(-0.3) = 0.74$ is the ratio of the two odds ratios, $1.65/2.23 = 0.74$: the exposure odds ratio among women is 0.74 times the exposure odds ratio among men, a departure from a constant odds ratio across sex (multiplicative interaction on the odds scale). The same number is the ratio of the sex odds ratios between exposed and unexposed.

Question 10

Week 5. Exam topic 8. About 7 minutes.

The table gives a binary outcome by binary exposure X and binary covariate C status, with 100 people in every cell.

C	X	People	Cases	Risk
0	1	100	20	0.20
0	0	100	10	0.10
1	1	100	60	0.60
1	0	100	30	0.30

Given. Three additive models with three parameters each: identity link, $P = \alpha_0 + \alpha_1 X + \alpha_2 C$; log link, $\log P = \gamma_0 + \gamma_1 X + \gamma_2 C$; logit link, $\operatorname{logit} P = \delta_0 + \delta_1 X + \delta_2 C$. $\log(0.1) = -2.30$; $\log 2 = 0.69$; $\log 3 = 1.10$; $\exp(-0.51) = 0.60$.

(a) Within each stratum of C, compute the RD, the RR and the OR. Which measure is the same in both strata?

(b) Which of the three additive models reproduces all four cell risks exactly? Write its three coefficients using the printed logs, and show that its fitted risk for $X = 1, C = 1$ equals the table.

Answers. **(a)** $C = 0$: RD $= 0.20 - 0.10 = 0.10$; RR $= 0.20/0.10 = 2.0$; OR $= (0.2/0.8)/(0.1/0.9) = 0.25/0.111 = 2.25$. $C = 1$: RD $= 0.60 - 0.30 = 0.30$; RR $= 0.60/0.30 = 2.0$; OR $= (0.6/0.4)/(0.3/0.7) = 1.5/0.429 = 3.5$. The RR is the same in both strata; the RD and the OR are not.

(b) The log-link model: the risk doubles with X in both strata and triples with C in both exposure groups ($0.30/0.10 = 0.60/0.20 = 3$), which is exactly what additivity on the log scale says. $\gamma_0 = \log(0.10) = -2.30$, $\gamma_1 = \log 2 = 0.69$, $\gamma_2 = \log 3 = 1.10$. Fitted risk for $X = 1, C = 1$: $\exp(-2.30 + 0.69 + 1.10) = \exp(-0.51) = 0.60$, the table's value (and $\exp(-2.30) = 0.10$ for the referent cell). The identity and logit models cannot do it with three parameters, because the RD (0.10 versus 0.30) and the OR (2.25 versus 3.5) differ between strata.

Question 11

Week 6. Exam topics 19 and 11. About 6 minutes.

In a cohort of 200 people with a binary outcome (100 exposed with risk 0.20; 100 unexposed with risk 0.10), a log-binomial model of the risk ratio failed to converge, so the analyst fit a Poisson model with a log link to the same 0/1 outcome and reported $\exp(\beta_1) = 2.0$ with a model-based standard error.

Given. Variance functions: Gaussian, constant; binomial, $\mu(1-\mu)$; Poisson, μ; gamma, μ^2. Approximate variance of the log of a group's risk, for a group of n people with risk μ: binomial, $(1-\mu)/(n\mu)$; Poisson, $1/(n\mu)$. $\sqrt{1.25} = 1.12$.

(a) For the exposed group ($n = 100$, $\mu = 0.20$), compute the variance of the log risk under the binomial family and under the Poisson family, and the ratio of the two standard errors (Poisson to binomial).

(b) In which direction is the Poisson model-based standard error wrong, for any risk strictly between 0 and 1? Name what repairs it without changing the coefficient.

Answers. (a) Binomial: $(1 - 0.20)/(100 \times 0.20) = 0.80/20 = 0.040$. Poisson: $1/(100 \times 0.20) = 1/20 = 0.050$. Ratio of standard errors, Poisson to binomial: $\sqrt{0.050/0.040} = \sqrt{1.25} = 1.12$, so the Poisson standard error is 12% too large for this group.

(b) Too large, always: for any risk strictly between 0 and 1, $\mu > \mu(1-\mu)$, so the Poisson variance function overstates the variance of a 0/1 outcome and the model-based standard error is conservative (intervals too wide, p-values too large); the ratio is $\sqrt{1/(1-\mu)}$, which grows with the risk. The repair is a robust (sandwich) standard error, which estimates the variance from the observed residuals rather than from the family's variance function and leaves the coefficient, $\exp(\beta_1) = 2.0$, exactly where it was.

Question 12

Week 7. Exam topics 11 and 7. About 5 minutes. (Pairs with Question 9.)

Return to the logistic model of Question 9, $\text{logit}\, P(Y{=}1) = -1.5 + 0.8\, X - 0.04\,(\text{age} - 60) + 0.5\, F - 0.3\, XF$. The part of its estimated variance-covariance matrix for the coefficients of X, F and XF is:

	X	F	XF
X	0.040	-0.010	-0.035
F	-0.010	0.050	-0.030
XF	-0.035	-0.030	0.090

Given. $\text{Var}(A + B) = \text{Var}(A) + \text{Var}(B) + 2\,\text{Cov}(A, B)$. $\text{Var}(A + wB) = \text{Var}(A) + w^2\,\text{Var}(B) + 2w\,\text{Cov}(A, B)$. $\sqrt{0.060} = 0.245$; $\sqrt{0.130} = 0.361$; $\exp(0.02) = 1.02$; $\exp(0.98) = 2.66$.

(a) Compute the variance and the standard error of the exposure log odds ratio among women, $\beta_X + \beta_{XF}$.

(b) Compute its 95% Wald confidence interval, and the interval for the odds ratio.

Answers. **(a)** $\text{Var}(\beta_X + \beta_{XF}) = \text{Var}(\beta_X) + \text{Var}(\beta_{XF}) + 2\,\text{Cov}(\beta_X, \beta_{XF}) = 0.040 + 0.090 + 2(-0.035) = 0.130 - 0.070 = 0.060$. SE $= \sqrt{0.060} = 0.245$.

(b) The log odds ratio among women is $0.8 - 0.3 = 0.5$. Wald interval: $0.5 \pm 1.96 \times 0.245 = 0.5 \pm 0.48$, so $(0.02,\ 0.98)$ on the log scale. Odds ratio $\exp(0.5) = 1.65$, interval $(\exp(0.02),\ \exp(0.98)) = (1.02,\ 2.66)$.

Question 13

Week 7. Exam topic 11. About 6 minutes.

A study of 60 primary-care clinics (20 patients each, 1,200 patients) compared systolic blood pressure after one year between the 30 clinics that adopted a new hypertension protocol and the 30 that did not; adoption was a clinic-level decision. A linear model of blood pressure on the protocol indicator and patient covariates gave a coefficient of −3.0 mmHg. Four standard errors were computed for it.

Standard error	Value
Model-based (from the fitted family's variance function)	1.00
HC3 robust (sandwich, patients treated as independent)	1.10
Cluster-robust (sandwich, clinics as clusters)	1.50
SD of 1,000 cluster-bootstrap replicates	1.55

Given. 95% interval = estimate $\pm 1.96\times$ SE.

(a) In one sentence each, say what the four standard errors are built from. Which one should be reported here, and what counts as the sample size for it?

(b) Compute the 95% confidence interval from the standard error you chose.

Answers. (a) Model-based: from the family's variance function, one residual variance assumed constant and independent across all 1,200 patients. HC3 robust: from each patient's own squared residual (the sandwich), which allows unequal residual variances but still treats the 1,200 patients as independent. Cluster-robust: from the sums of residual cross-products within each clinic (the sandwich with 60 clusters), which allows any correlation among the 20 patients of a clinic. Cluster bootstrap: from the spread of the coefficient over 1,000 resamples that draw whole clinics with replacement. Report the cluster-robust standard error (the cluster bootstrap agrees with it): adoption was a clinic-level decision, so the 20 patients in a clinic share the exposure and whatever else the clinic does, and their residuals are correlated, which is why the first two standard errors are too small. The sample size that counts is the 60 clinics (30 against 30), not the 1,200 patients.

(b) $-3.0 \pm 1.96 \times 1.50 = -3.0 \pm 2.94$, so $(-5.94,\ -0.06)$ mmHg.

Question 14

Week 8. Exam topic 9. About 9 minutes.

Z is a cause of the outcome but is unrelated to the exposure: half of each exposure group has $Z = 1$. There are 100 people in every cell.

Z	X	People	Cases	Risk
0	1	100	50	0.50
0	0	100	20	0.20
1	1	100	80	0.80
1	0	100	50	0.50

(a) Compute the RD, the RR and the OR within each stratum of Z, and in the whole table ignoring Z.

(b) Show the arithmetic that recovers the overall RR from the four stratum risks

(c) Is there a comparable arithmetic (i.e., weighted sum) that can recover the overall OR from the stratum specific ORs? Why or why not?

Answers. (a) $Z = 0$: RD $= 0.50 - 0.20 = 0.30$; RR $= 0.50/0.20 = 2.5$; OR $= (0.5/0.5)/(0.2/0.8) = 1/0.25 = 4.0$. $Z = 1$: RD $= 0.80 - 0.50 = 0.30$; RR $= 0.80/0.50 = 1.6$; OR $= (0.8/0.2)/(0.5/0.5) = 4/1 = 4.0$. Whole table: exposed $130/200 = 0.65$, unexposed $70/200 = 0.35$; RD $= 0.30$; RR $= 0.65/0.35 = 1.86$; OR $= (0.65/0.35)/(0.35/0.65) = 1.857/0.538 = 3.45$.

(b) Because Z is split 50:50 in both exposure groups, each overall risk is the average of its two stratum risks: exposed $0.5 \times 0.50 + 0.5 \times 0.80 = 0.65$, unexposed $0.5 \times 0.20 + 0.5 \times 0.50 = 0.35$, and RR $= 0.65/0.35 = 1.86$. The same number is a weighted average of the stratum RRs, with weights proportional to each stratum's share of the unexposed risk: $w_0 = 0.10/0.35 = 0.286$, $w_1 = 0.25/0.35 = 0.714$, and $0.286 \times 2.5 + 0.714 \times 1.6 = 1.86$. The RR is collapsible: the overall value sits between 1.6 and 2.5.

(c) No. The overall OR, 3.45, lies outside the range of the stratum ORs, which are both 4.0, and no weighted average with non-negative weights that sum to one can land outside the range of the numbers it averages. The reason is that marginalizing over Z averages risks (as in (b)), and the odds of an average risk is not the average of the odds: odds are a curved function of risk, so the OR is not a weighted average of stratum ORs, even with a constant conditional OR and no confounding. This is non-collapsibility, a property of the measure, not a bias; the RD, a linear function of the risks, is strictly collapsible (0.30 everywhere).

Question 15

Week 8. Exam topics 9 and 10. About 6 minutes.

A logistic regression of a binary outcome on the exposure X and a binary covariate Z gave

$$\operatorname{logit} P(Y{=}1 \mid X, Z) = -1.39 + 1.39\,X + 1.39\,Z.$$

In the study population 30% of people have $Z = 1$, and Z is the only confounder.

Given. $P = \exp(\eta)/\{1 + \exp(\eta)\}$. $\exp(-1.39) = 0.25$; $\exp(0) = 1$; $\exp(1.39) = 4.0$. Standardized risk under $X = x$: $\sum_z P(Y{=}1 \mid X{=}x, Z{=}z)\,P(Z{=}z)$.

(a) Compute the four fitted risks, then the standardized risk if everyone were exposed and if no one were exposed.

(b) Compute the standardized (marginal) RD, RR and OR, and compare the OR with exp(1.39).

Answers. **(a)** Linear predictors and fitted risks: $X = 0, Z = 0$: $\eta = -1.39$, $P = 0.25/1.25 = 0.20$. $X = 1, Z = 0$: $\eta = 0$, $P = 1/2 = 0.50$. $X = 0, Z = 1$: $\eta = 0$, $P = 0.50$. $X = 1, Z = 1$: $\eta = 1.39$, $P = 4/5 = 0.80$. Standardized over Z (weights 0.7 and 0.3): everyone exposed, $0.7 \times 0.50 + 0.3 \times 0.80 = 0.35 + 0.24 = 0.59$; no one exposed, $0.7 \times 0.20 + 0.3 \times 0.50 = 0.14 + 0.15 = 0.29$.

(b) RD $= 0.59 - 0.29 = 0.30$. RR $= 0.59/0.29 = 2.03$. OR $= (0.59/0.41)/(0.29/0.71) = 1.439/0.408 = 3.52$, against the conditional OR $\exp(1.39) = 4.0$: the marginal OR is closer to the null although Z is fully adjusted and every person's own odds ratio is 4.0, which is non-collapsibility of the OR.

Question 16

Week 9. Exam topic 10. About 7 minutes.

A cohort of 200 people, with a binary confounder C and a binary outcome.

C	X	People	Cases	Risk
0	1	20	4	0.20
0	0	80	8	0.10
1	1	60	30	0.50
1	0	40	16	0.40

Given. For a person with $X = x$ and $C = c$: unstabilized weight $w = 1/P(X{=}x \mid C{=}c)$; stabilized weight $sw = P(X{=}x)/P(X{=}x \mid C{=}c)$. Weighted risk in arm x = (sum of weight × cases) ÷ (sum of weight × people) over the cells with $X = x$.

(a) Compute $P(X{=}1 \mid C{=}0)$, $P(X{=}1 \mid C{=}1)$, $P(X{=}1)$, and the unstabilized and stabilized weights for each of the four cells.

(b) Compute the total weighted population under each set of weights and, under the stabilized weights, the proportion exposed within each level of C. State what each of these three numbers checks.

Answers. (a) $P(X{=}1 \mid C{=}0) = 20/100 = 0.20$; $P(X{=}1 \mid C{=}1) = 60/100 = 0.60$; $P(X{=}1) = 80/200 = 0.40$. Unstabilized weights: $C = 0, X = 1$: $1/0.20 = 5$; $C = 0, X = 0$: $1/0.80 = 1.25$; $C = 1, X = 1$: $1/0.60 = 1.67$; $C = 1, X = 0$: $1/0.40 = 2.5$. Stabilized weights (numerator 0.40 for the exposed, 0.60 for the unexposed): $0.40/0.20 = 2$; $0.60/0.80 = 0.75$; $0.40/0.60 = 0.67$; $0.60/0.40 = 1.5$.

(b) Unstabilized total: $20 \times 5 + 80 \times 1.25 + 60 \times 1.67 + 40 \times 2.5 = 100 + 100 + 100 + 100 = 400 = 2N$: each exposure arm is inflated to a copy of the whole cohort, so the pseudo-population holds everyone exposed and everyone unexposed. Stabilized total: $20 \times 2 + 80 \times 0.75 + 60 \times 0.67 + 40 \times 1.5 = 40 + 60 + 40 + 60 = 200 = N$: stabilization keeps the cohort's size. Exposed share under the stabilized weights: $C = 0$: $40/(40 + 60) = 0.40$; $C = 1$: $40/(40 + 60) = 0.40$, both equal to $P(X{=}1) = 0.40$: in the pseudo-population C no longer predicts X (confounding by C is gone) and the marginal exposure prevalence is preserved.

Question 17

Week 10. Exam topic 12. About 5 minutes.

A linear spline (piecewise linear function) with knots at 2 and 5 was fit for the mean of a continuous outcome Y as a function of a continuous exposure x (0 to 8 units):

$$E(Y \mid x) = 10 + 3\,x - 4\,(x-2)_+ + 2\,(x-5)_+,$$

where $(x-\kappa)_+$ equals $x-\kappa$ when $x > \kappa$ and 0 otherwise.

(a) Compute the predicted mean at $x = 1$ and at $x = 6$, and the contrast between them.

(b) Say why a spline fit is generally reported as predictions and contrasts of those predictions at chosen values of x, rather than as its coefficients.

Answers. (a) At $x = 1$ both hinge terms are zero: $10 + 3(1) = 13$. At $x = 6$: $10 + 3(6) - 4(6-2) + 2(6-5) = 10 + 18 - 16 + 2 = 14$. Contrast $14 - 13 = 1$.

(b) The coefficients belong to basis columns, not to the exposure: 3 is the slope below the first knot only, and -4 and 2 are changes in slope at the knots, so no coefficient is "the effect of x", and their values depend on the basis chosen (knot positions, hinge parameterization) although the fitted curve does not. A prediction at a value of x, or a contrast between two values, is the same under any equivalent basis, is in the outcome's units, and answers the question asked; its standard error combines several coefficients through their covariances, which a single printed coefficient and standard error cannot give.

Question 18

Week 10. Exam topics 8 and 12. About 5 minutes.

Three logistic models of the same outcome on a binary exposure X and one continuous confounder C (60 distinct values) were fit to the same people. They differ only in how C enters: linearly; as a spline with 4 degrees of freedom; as one indicator (i.e., dummy variable) per distinct value of C.

Model for C	Parameters p	Deviance	Exposure coefficient (SE)
Linear	3	950.0	0.60 (0.20)
Spline, 4 df	6	940.0	0.45 (0.21)
One indicator per value	61	900.0	0.44 (0.24)

(a) Rank each of the three fits along the "parametric-to-nonparametric" spectrum, and say which end carries less bias and which less variance.

(b) The question is causal: the effect of X on the outcome. Choose a form for C as the model from which you would report the effect of X. Justify your choice in one or two sentences.

Answers. (a) Linear in C (3 parameters) is the parametric end; the 4-df spline (6 parameters) is semiparametric, in the middle; one indicator per value (61 parameters, saturated in C) is the nonparametric end. The nonparametric end carries less bias, since it can reproduce any shape of the C-outcome relation and so leaves no residual confounding from a wrong form, but more variance, since 59 nuisance parameters are estimated from the same data (the exposure SE grows from 0.20 to 0.24). The parametric end has less variance but more bias when its form is wrong: the exposure coefficient moves from 0.60 to 0.45 once C is allowed to bend, so the linear term left confounding in the estimate. The deviance falls monotonically (950, 940, 900) because fit always improves with flexibility; it measures fit, not the accuracy of the target.

(b) The spline. It moves the exposure coefficient almost all the way to the saturated answer (0.45 against 0.44) while spending 3 extra parameters instead of 58, so it removes the residual confounding the linear term left (the shift from 0.60) at a small cost in variance (SE 0.21 against 0.24). The linear form is ruled out by the shift itself, and the saturated form adds variance without changing the answer. (AIC agrees: 956, 952, 1022.)

Question 19

Week 11. Exam topic 14. About 5 minutes.

Four candidate models for a nuisance function $g(C)$ in $E(Y \mid X, C) = \beta X + g(C)$ were fit to 2,000 people. The table gives each candidate's in-sample mean squared error (MSE) and its 5-fold cross-validated MSE.

Candidate for $g(C)$	In-sample MSE	5-fold CV MSE
Linear in C	25.0	25.4
Spline, 3 df	24.0	24.3
Spline, 8 df	23.0	24.9
One indicator per value of C	20.0	28.0

A super learner fit on the same folds put weight 0.70 on the 3-df spline and 0.30 on the linear model (0 on the other two). For one new person, the 3-df spline predicts 4.0 and the linear model predicts 5.0.

(a) Which candidate does in-sample loss pick, and which does cross-validation pick? Which is more valid?

(b) Compute the non-discrete super learner's prediction for the new person, and show how you calculate it. Name the algorithm that the "discrete super learner" would pick among the four candidates.

Answers. (a) In-sample loss picks the indicator-per-value model (MSE 20.0), and would pick the most flexible candidate whatever the table said, because fitting the same data more closely will almost always lower the in-sample error. Cross-validation picks the 3-df spline (CV MSE 24.3): the held-out error falls from the linear model to the 3-df spline and then rises for the 8-df spline (24.9) and the indicator model (28.0), the signature of overfitting, and the gap between the two columns grows with flexibility (0.4, 0.3, 1.9, 8.0). The CV choice is the valid one: it estimates the error on data the fit did not see, which is the quantity a nuisance model is chosen for, whereas in-sample error is optimistic by an amount that grows with the number of parameters.

(b) Super learner prediction $= 0.70 \times 4.0 + 0.30 \times 5.0 = 2.8 + 1.5 = 4.3$; the weights sum to one, so the prediction is a weighted average of the two candidates' predictions (the other two carry weight 0). The discrete super learner picks the single candidate with the smallest cross-validated error: the 3-df spline.

Question 20

Week 12. Exam topics 13 and 19. About 7 minutes.

	Cases	Non-cases	Total
Exposed	7	0	7
Unexposed	20	20	40

Logistic regression of case status on exposure returns an exposure coefficient of 19.6 with a standard error of 2,900.

(a) Say why the odds ratio for this table is not a number. How, exactly, does the maximum-likelihood fitting loop from the logistic model produce an estimate of 19.6?

(b) Use Firth's method and add half a person to each cell, and compute the odds ratio in this sparse data table. Then, suppose the exposed group had instead been 80 total people with 70 cases (the unexposed row unchanged). Compute the odds ratio for this new table, and also use Firth's method to compute the "half-person" odds ratio in this new table. Compare the two odds ratios from the two tables (the fit's coefficient of 19.6, an OR of exp(19.6), and Firth's OR in the sparse table; the OR and Firth's OR in the unsparse table). Say in one sentence what happens to the half-person's pull as the size of the cells in the table increase.

(c) Name the two remedies for the empty cell discussed in the course. For each remedy, discuss whether it changes the question being asked, or the estimator being used, or both.

Answers. (a) OR $= (7 \times 20)/(0 \times 20)$: the exposed non-case cell is empty, so the exposed odds of being a case are $7/0$ and the odds ratio is infinite, not a number (exposure separates the data). The log-likelihood keeps rising as the exposure coefficient grows and has no maximum at any finite value, so the iterative fitting loop takes step after step uphill, each step raising the likelihood by less, until the change falls below its convergence tolerance (or the iteration limit) and it reports wherever it stopped: 19.6, with the fitted risk for the exposed essentially 1. The standard error of 2,900 is the flatness of the likelihood at that point and means "not estimable", not "a wide interval".

(b) Sparse table, half a person in each cell: OR $= (7.5 \times 20.5)/(0.5 \times 20.5) = 7.5/0.5 = 15.0$, against the fit's $\exp(19.6) \approx 3 \times 10^8$, infinite in effect. Unsparse table: OR $= (70 \times 20)/(10 \times 20) = 7.0$; half-person OR $= (70.5 \times 20.5)/(10.5 \times 20.5) = 70.5/10.5 = 6.71$. The same half-person that takes the sparse estimate from infinity to 15 moves the unsparse estimate by 4%: its pull fades as the cells fill, because it is a fixed amount of pseudo-information set against a growing amount of real information.

(c) Pooling: merge the sparse cell or category with a neighbour so the empty cell never arises; the estimator is unchanged (maximum likelihood on the new variable) but the question is, since the contrast is now between different groups. Penalizing: Firth's correction (the half-person, generalized through the information matrix) or a pseudo-record prior keeps every coefficient finite without touching the data; the question is unchanged and the estimator is, a penalized likelihood in place of the likelihood. Firth's penalty also shrinks the intercept, so if the question were about absolute risks rather than the odds ratio, the penalty would touch it too.

Question 21

Week 12. Exam topic 13. About 6 minutes.

An analyst wants the effect of an exposure X (0/1) on a continuous outcome Y, adjusting for three confounders C_1, C_2, C_3, all known to affect both X and Y. Two fits to 1,000 people:

Term	OLS estimate (SE)	Lasso at the CV-chosen λ
X	3.0 (0.8)	2.1
C_1	2.0 (0.5)	1.5
C_2	0.4 (0.4)	0.0
C_3	-1.0 (0.5)	-0.7

C_2 is strongly associated with X. The analyst reports the lasso's coefficient for X with a 95% interval of 1.4 to 2.8, obtained by rerunning the lasso on 2000 bootstrap samples and taking the 2.5th and 97.5th percentiles of the 2000 coefficients.

(a) Name the two things the lasso did to the coefficient of X, and explain why the second one leaves confounding in the estimate.

(b) Say what the bootstrap interval ignores, and what that can do to its coverage.

Answers. (a) It shrank the coefficient toward zero, from 3.0 to 2.1, which the penalty does to every coefficient by construction, and it deleted C_2 by setting its coefficient to exactly zero. The deletion leaves confounding in the estimate because the lasso judged C_2 by its small outcome coefficient (0.4), not by its role in the causal structure: C_2 is strongly associated with X, and the confounding a variable carries is the product of its exposure association and its outcome association, so a column that predicts Y weakly can still move the exposure coefficient a lot once it is gone.

(b) The interval ignores that its center is a shrunken, post-selection estimate: every bootstrap replicate shrinks and selects in the same way, so the 2,000 coefficients spread around 2.1, not around 3.0 or around the true effect, and the selection makes the sampling distribution non-standard besides (a point mass at zero for columns that drop in and out). An interval centered on the wrong number covers the true effect less often than 95%, and the shortfall grows with the shrinkage and with the confounding left behind; resampling cannot repair a biased center.

Question 22

Week 11. Exam topic 12. About 6 minutes.

An analyst wants the effect of a binary exposure A on a continuous outcome Y, adjusting for a set of confounders W, with the partially linear model $E(Y \mid A, W) = \psi A + g(W)$, where g is left unspecified. Two nuisance models were fit with a super learner and 5-fold cross-fitting: the outcome model $E(Y \mid W)$ and the exposure model $E(A \mid W)$. The table gives, for four people from one held-out fold, the exposure, the outcome, and the two out-of-fold predictions.

Person	A	$\hat{E}(A \mid W)$	Y	$\hat{E}(Y \mid W)$
1	1	0.5	10.0	8.0
2	0	0.5	7.0	8.0
3	1	0.2	12.4	10.0
4	0	0.2	9.4	10.0

Given. Exposure residual $\tilde{A} = A - \hat{E}(A \mid W)$; outcome residual $\tilde{Y} = Y - \hat{E}(Y \mid W)$; residual-on-residual estimate (least squares through the origin) $\hat{\psi} = \sum_i \tilde{A}_i \tilde{Y}_i / \sum_i \tilde{A}_i^2$.

(a) Compute the two residuals for each person and the estimate $\hat{\psi}$, treating the four people as the whole sample.

(b) The predictions in the table come from models fit on the other four folds. Why? What goes wrong if each person's own data are used to fit them?

(c) In one sentence, what does $\hat{\psi}$ estimate, and what must hold, beyond the identification conditions, for it to be the average treatment effect?

Answers. **(a)** $\tilde{A}$: $1 - 0.5 = 0.5$; $0 - 0.5 = -0.5$; $1 - 0.2 = 0.8$; $0 - 0.2 = -0.2$. $\tilde{Y}$: $10.0 - 8.0 = 2.0$; $7.0 - 8.0 = -1.0$; $12.4 - 10.0 = 2.4$; $9.4 - 10.0 = -0.6$. Products $\tilde{A}\tilde{Y}$: $1.0, 0.5, 1.92, 0.12$, sum 3.54. Squares $\tilde{A}^2$: $0.25, 0.25, 0.64, 0.04$, sum 1.18. $\hat{\psi} = 3.54/1.18 = 3.0$: the exposure raises the outcome by 3.0 units, with $g(W)$ never specified.

(b) Cross-fitting. A flexible fit reproduces the observations it was trained on too well, so in-sample residuals are too small and carry the fit's adaptivity (its overfitting) into the least squares step, which biases $\hat{\psi}$ and makes its standard error dishonest. Predicting each fold from models that never saw it gives residuals with the right size, so the slope and its ordinary (sandwich) standard error from the least squares step are valid.

(c) The constant exposure effect ψ of the partially linear model, the conditional effect of A given W, which is the average treatment effect under exchangeability given W, positivity and consistency provided the effect is the same for everyone (no exposure-confounder interaction) and the exposure model $E(A \mid W)$ is right (if only the outcome model is right the estimate attenuates toward the null), with both nuisance fits accurate enough for their errors to cancel.

Question 23

Week 13. Exam topic 15. About 10 minutes.

Eight patients were followed from kidney transplant. Event A is graft failure; event B is death with a functioning graft; "censored" means the patient left the study alive with a functioning graft.

Patient	Months	Status
1	1	A, graft failure
2	2	censored
3	3	A, graft failure
4	4	B, death
5	5	A, graft failure
6	6	censored
7	7	B, death
8	8	A, graft failure

Given. Kaplan-Meier all-cause survival: $S(t) = \prod_{t_k \le t}(1 - d_k/n_k)$, the product over the distinct event times t_k up to t, where d_k is the number of events of either type at t_k and n_k the number still at risk just before t_k. Aalen-Johansen risk of cause j: $F_j(t) = \sum_{t_k \le t} S(t_k^-)\, d_{jk}/n_k$, where $S(t_k^-)$ is the Kaplan-Meier all-cause survival just before t_k and d_{jk} the number of cause-j events at t_k.

(a) Before any estimator: using counts alone, give the lowest and the highest value the all-cause risk at 5 months could take, depending on what happened to the patient censored before month 5. Then compute $1 - S(5)$ by Kaplan-Meier and check that it lies between them.

(b) Compute the Aalen-Johansen risk of graft failure (A) at 5 months and of death (B) at 5 months, and check that they sum to $1 - S(5)$.

(c) An analyst codes the deaths (B) as censored and computes a Kaplan-Meier "risk of graft failure" at 5 months. Compute it, say by how much it exceeds the Aalen-Johansen value, and say in one sentence what that excess represents.

Answers. **(a)** By month 5 there are four events (months 1, 3, 4, 5) among eight patients and one patient censored at month 2. If that patient never had an event, the risk is $4/8 = 0.50$; if they had one before month 5, it is $5/8 = 0.625$. Kaplan-Meier: $S(5) = (1-1/8)(1-1/6)(1-1/5)(1-1/4) = (7/8)(5/6)(4/5)(3/4) = 0.4375$, so $1 - S(5) = 0.5625$, inside $[0.50, 0.625]$: censoring shrinks the later risk sets (6, 5, 4 instead of 7, 6, 5), which puts the estimate above the lower bound.

(b) Graft failure: jumps at months 1, 3 and 5, $S(t_k^-) \times d_{Ak}/n_k = 1 \times 1/8 + (7/8) \times 1/6 + (7/8)(5/6)(4/5) \times 1/4 = 0.125 + 0.146 + 0.146 = 0.417$. Death: one jump at month 4, $(7/8)(5/6) \times 1/5 = 0.146$. Sum $0.417 + 0.146 = 0.5625 = 1 - S(5)$.

(c) With deaths censored, the graft-failure risk set at month 5 is still 4 but the month-4 factor disappears: $S_A(5) = (7/8)(5/6)(3/4) = 0.547$, so $1 - S_A(5) = 0.453$, which exceeds the Aalen-Johansen 0.417 by 0.036. The excess is graft-failure risk imputed to the patient who died, as if they had remained at risk and gone on to fail at the same rate as the survivors: the risk of a world in which death is switched off, not the risk in this cohort.

Question 24

Week 13. Exam topic 16. About 8 minutes.

A manuscript says: "Because the hazard ratio is a biased estimate of the causal effect in heterogeneous populations, we report the risk difference at five years instead." For the sake of the question, suppose we knew both potential survival times for all 100 patients in the study:

Would survive to 5 years under …	Patients
both treatment and control	50
control only	20
treatment only	5
neither	25

(a) Say what is wrong with the sentence, in estimand-versus-estimator language, in two sentences.

(b) The course gave two properties that a causal estimand should have. State them. Then count from the table: how many patients are in the denominator of the hazard at 5 years under treatment, how many under control, and how many in the denominator of each arm's 5-year risk? Compute the causal risk difference at 5 years.

(c) Say why the hazard ratio's selection problem grows over follow-up in a heterogeneous population even when every hazard is estimated without error, and name the row of the table that causes it.

Answers. (a) Bias belongs to estimators, measured against an estimand; an estimand cannot be biased, and the hazard ratio is a well-defined contrast that can be estimated without bias. What the sentence means is that the hazard ratio is a poor causal estimand, because it compares different sets of people as follow-up proceeds, and that is a defect of the quantity chosen, not an estimation error.

(b) (A) A comparison of counterfactual marginal distributions, the same outcome under two treatments; (B) a contrast made within a common set of units. Hazard denominators at 5 years: under treatment, those alive at 5 under treatment, $50 + 5 = 55$; under control, $50 + 20 = 70$, a different set of people (the 20 "control only" patients are in one and not the other, the 5 "treatment only" the reverse), so property (B) fails. Risk denominators: all 100 patients in both arms. Risks $1 - 55/100 = 0.45$ under treatment and $1 - 70/100 = 0.30$ under control; RD $= 0.15$, a harm at 5 years.

(c) The hazard at t conditions on surviving to t, and treatment itself decides who survives: in a heterogeneous population the treated survivors and the control survivors differ in composition (the frail are removed faster from whichever arm harms them), and the two surviving groups drift further apart as follow-up continues, so later hazard ratios compare increasingly different people and move toward or across the null even with a constant individual effect and perfectly estimated hazards. The row that causes it is "control only" (region H): alive in the control denominator, absent from the treatment denominator.

Question 25

Week 13. Exam topic 17. About 7 minutes.

Two models were fit to the same data on time to a first cardiovascular event, with a binary exposure X. The Cox model gave $\beta = 0.69$ for X, and the baseline survival for the unexposed at 5 years was estimated afterwards as $S_0(5) = 0.80$. The accelerated failure time (AFT) model gave $\gamma = -0.69$ for X, meaning that the exposed reach any given survival probability in half the time of the unexposed; its fitted survival curve for the unexposed gives $S_0(5) = 0.80$ and $S_0(10) = 0.60$.

Given. Cox: $S(t \mid X{=}1) = S_0(t)^{\exp(\beta)}$. AFT: $S(t \mid X{=}1) = S_0\{t \times \exp(-\gamma)\}$. $\exp(0.69) = 2.0$; $\exp(-0.69) = 0.50$. Risk $= 1 - S$.

(a) From the Cox fit, compute the 5-year risk for an exposed person and for an unexposed person.

(b) From the AFT fit, compute the 5-year risk for an exposed person.

(c) Say what the parametric baseline buys and what it costs, and why the Cox fit alone reports hazard ratios and not risks: what has to be estimated afterwards, and why it is a nuisance function.

Answers. (a) Exposed: $S(5 \mid X{=}1) = 0.80^{\exp(0.69)} = 0.80^2 = 0.64$, risk 0.36. Unexposed: risk $1 - 0.80 = 0.20$.

(b) $S(5 \mid X{=}1) = S_0(5 \times \exp(0.69)) = S_0(10) = 0.60$, risk 0.40: the exposed at 5 years sit where the unexposed sit at 10.

(c) The parametric baseline buys a complete survival curve for every covariate pattern at every t, including times beyond the data, a time-scale interpretation (times are halved), and efficiency when the shape is right; it costs a shape assumption (Weibull, log-normal, and so on) that, when wrong, biases every risk the model reports. The Cox fit estimates β without specifying the baseline hazard, so the fit alone yields hazard ratios; a risk at t needs $S_0(t)$, a function of time estimated afterwards from the fitted model and the data. It is a nuisance function because the exposure contrast does not need it, while any absolute risk does.