ACTEX Learning

Study Manual for Exam PA

12th Edition

Ambrose Lo, PhD, FSA, CERA

An SOA Exam

ACTEX Learning

Study Manual for Exam PA

12th Edition

Ambrose Lo, PhD, FSA, CERA

Actuarial & Financial Risk Resource Materials
Since 1972

Printed in the United States of America.

Who We Are

A Benefit Corporation Experienced at Teaching Actuaries!

Experienced

More than 50 years experience helping students prepare and pass actuarial exams! We are an eLearning technology and education company leveraging experts in the field to constantly update our learning content in a format that works for you.

Trustworthy

ACTEX Learning is a leading US based provider of study materials for actuarial exams. Our authors and content contributors are renowned academics and Actuaries that are proud to have their names on the cover of our manuals and textbooks!

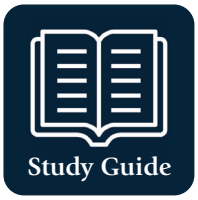
Mission Focused

We are a Benefit Corporation focusing on the mission of accessible high quality actuarial education. We're dedicated to empowering actuarial students by offering test prep materials that are not just effective and efficient but also tailored to suit every type of student.

Become an ACTEX Champion

Join our Global ACTEX Champion Program and bring the benefits to your Actuarial Club. To learn more about the program, scan the QR code on the right. If you have any questions or would like to speak to a Champion Coordinator, please do not hesitate to reach out to us at champions@actexlearning.com.

ACTEX Has the Solutions to Help You with Exam Prep



Our study guides offer the most recommended actuarial prep program. Search our interactive manuals for different topics and toggle easily between concepts and study materials.

Available for P, FM, FAM, ALTAM, ASTAM, PA. ATPA, MAS-I, MAS-II, CAS 5, CAS 6 US & CAN, CAS 7, CAS 8, CAS 9

Want to know you're prepared for your exam? Practice efficiently with our robust database of questions and solutions and predict your success through GOAL's innovative scoring system. GOAL also features dedicated instructor support so you can get help where you need it and pass the exam with confidence!

Available for P, FM, FAM, ALTAM, ASTAM, MAS-I, MAS-II, CAS 5, CAS 6 US & CAN

Master key topics and formulas with our flashcards, which allow you to filter by topic. To help you direct your focus, each card is rated to indicate its importance on the exam.

Available for P, FM, FAM, ALTAM, ASTAM, PA. ATPA, MAS-I, MAS-II, CAS 5, CAS 6 US & CAN, CAS 7, CAS 8, CAS 9

Studies have shown video learning can lead to better retention. We offer hours of video instruction to aid you in your studies. They're a great way to deepen your learning on challenging topics and gain a variety of perspectives from our expert instructors.

Available for P, FM, FAM, ALTAM, ASTAM, PA. ATPA, MAS-I, MAS-II, CAS 5, CAS 6 US & CAN, CAS 7, CAS 8, CAS 9

ACTEX offers convenient online courses approved for CAS VEE credits. All courses are available on-demand. Students complete the courses at their own pace and take the final exams on the date of their choosing.

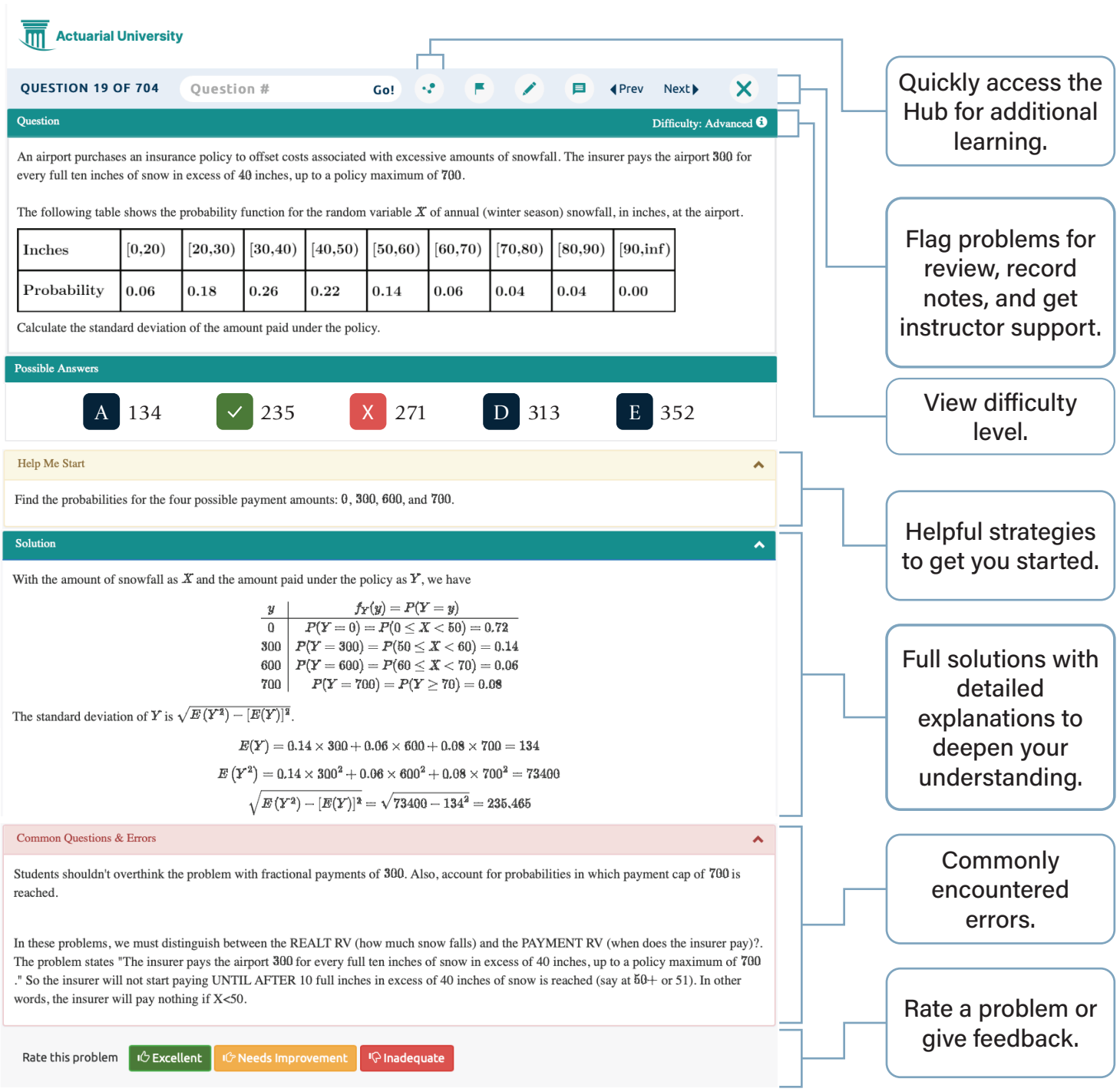
Available for Accounting & Finance, Mathematical Statistics, and Economics

Study Materials are Available for the Following:

SOA: P, FM, FAM, ALTAM, ASTAM, SRM, PA, ATPA, CFE, GI, GH, ILA, RET
CAS: MAS-I, MAS-II, CAS 5, CAS 6C, CAS 6US, CAS 7, CAS 8, CAS 9

GOAL Improves Your Studies

How you can prepare for your exam confidently with GOAL custom Practice Sessions, Quizzes, & Simulated Exams:



Actuarial University

QUESTION 19 OF 704 | Question # | Go! | Prev | Next

Question — Difficulty: Advanced

An airport purchases an insurance policy to offset costs associated with excessive amounts of snowfall. The insurer pays the airport 300 for every full ten inches of snow in excess of 40 inches, up to a policy maximum of 700.

The following table shows the probability function for the random variable X of annual (winter season) snowfall, in inches, at the airport.

Inches	[0,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,inf)
Probability	0.06	0.18	0.26	0.22	0.14	0.06	0.04	0.04	0.00

Calculate the standard deviation of the amount paid under the policy.

Possible Answers

A 134 ✓ 235 X 271 D 313 E 352

Help Me Start

Find the probabilities for the four possible payment amounts: 0, 300, 600, and 700.

Solution

With the amount of snowfall as X and the amount paid under the policy as Y, we have

y	$f_Y(y) = P(Y = y)$
0	$P(Y = 0) = P(0 \le X < 50) = 0.72$
300	$P(Y = 300) = P(50 \le X < 60) = 0.14$
600	$P(Y = 600) = P(60 \le X < 70) = 0.06$
700	$P(Y = 700) = P(Y \ge 70) = 0.08$

The standard deviation of Y is $\sqrt{E(Y^2) - [E(Y)]^2}$.

$$E(Y) = 0.14 \times 300 + 0.06 \times 600 + 0.08 \times 700 = 134$$

$$E(Y^2) = 0.14 \times 300^2 + 0.06 \times 600^2 + 0.08 \times 700^2 = 73400$$

$$\sqrt{E(Y^2) - [E(Y)]^2} = \sqrt{73400 - 134^2} = 235.465$$

Common Questions & Errors

Students shouldn't overthink the problem with fractional payments of 300. Also, account for probabilities in which payment cap of 700 is reached.

In these problems, we must distinguish between the REALT RV (how much snow falls) and the PAYMENT RV (when does the insurer pay)?. The problem states "The insurer pays the airport 300 for every full ten inches of snow in excess of 40 inches, up to a policy maximum of 700 ." So the insurer will not start paying UNTIL AFTER 10 full inches in excess of 40 inches of snow is reached (say at 50+ or 51). In other words, the insurer will pay nothing if X<50.

Rate this problem | Excellent | Needs Improvement | Inadequate

Quickly access the Hub for additional learning.

Flag problems for review, record notes, and get instructor support.

View difficulty level.

Helpful strategies to get you started.

Full solutions with detailed explanations to deepen your understanding.

Commonly encountered errors.

Rate a problem or give feedback.

Thank You for Choosing ACTEX Learning!

We're committed to helping you succeed on your actuarial journey.

For the latest study guides, textbooks, free Formula Sheets, and more resources for SOA, CAS, IFoA, and IAI exams, visit:

ACTEX Learning https://actexlearning.com/

Your destination for comprehensive actuarial exam preparation and professional development.

Looking for additional study material or other actuarial books?

THE ACTUARIAL BOOKSTORE https://www.actuarialbookstore.com/

The #1 online source for actuarial books and study guides.

Scan to Learn More

Don’t Miss Out on All of Your Materials!

Your study manual includes access to the following additional tools and resources to help you prepare for your exam:

STUDY PLANNER – enter the date of your exam and it will plan your studies day by day, topic by topic, resource by resource. This will keep you on track.

GOAL – exam-style problems with detailed solutions. Over 22,000 questions, practice, adaptive quizzes, simulated exams. GOAL Score analyzes your strengths and weaknesses by category, topic, and level of difficulty. Achieve a GOAL Score of 70 and you are ready for exam day! Stuck on a question? Instructors are there for you. Set yourself up for success, Practice with GOAL!

FLASHCARDS – great memorization and quizzing tool. Test your knowledge, filter by importance and master key definitions and formulas.

FORMULA SHEET – easy access to all the formulas in one place for quick reference and printable.

TOPIC SEARCH – find definitions of terms and tools for that topic across multiple resources.

DISCORD – take part in discussion forums with fellow students and receive guidance from professors.

MORE FREE RESOURCES - click on the Actuarial University logo to clear your Exam Dashboard and scroll down to Helpful Links to easily print your Formula Sheet and find more free resources.

To access all these resources be sure to activate the keycode that was sent via email when you placed your order.

GOAL Improves Your Studies

GOAL is a way to practice what you've learned in class or from independent study using textbooks and study manuals.

The GOAL (Guided Online Actuarial Learning) platform offers a database of 22,000 exam-style problems with detailed solutions, 3 learning modes (Practice, Quiz, Simulated Exams) and 3 levels of difficulty (Core, Advanced and Mastery). You control your learning mode, difficulty level and topics.

GOAL is currently available for the following exams:

P | FM | FAM | ALTAM | ASTAM | SRM

MAS-I | MAS-II | CAS 5C | CAS 6 CA | CAS 6 US

Use GOAL Score to Gauge Your Exam Readiness

Measure how prepared you are to pass your exam with a tool that suits any study approach. A GOAL Score of 70 or above indicates readiness.

Your score is broken into categories, allowing you to study efficiently by concentrating on problem areas. GOAL Score quantifies your exam readiness by measuring both your performance and the consistency of your performance. Your GOAL Score also analyzes your strengths and weaknesses by category, topic, and level of difficulty.

Scan to Learn More

Contents

Preface

⚠ NOTE TO STUDENTS ⚠

Please read this preface carefully, even if it looks long. It contains **VERY IMPORTANT** information that will help you navigate this manual smoothly and study for Exam PA effectively.

Why this Study Manual?

"The PA modules are so difficult to follow."

"The PA modules make things unnecessarily complicated and are riddled with errors."

"I feel that the PA modules don't cover enough ground for me to handle the exam well. I have to supplement my learning with external resources."

"I hate having to alternate among the PA modules, the R Markdown files, the required textbooks, and online readings."

"There is a lack of useful study resources for Exam PA in the market."

These are some of the most common comments PA exam candidates who studied for the exam solely using the Society of Actuaries (SOA)'s e-learning modules have voiced on Internet forums, e.g., the old Actuarial Outpost, Reddit, Discord. These "complaints" and the importance of passing this exam to earn the Associateship of the Society of Actuaries (ASA) designation in today's exam curriculum have motivated me to develop a completely new Exam PA study manual with the goal of streamlining, synthesizing, and augmenting the materials in the PA e-learning modules in a coherent and exam-oriented format. With this manual, you will have in your possession a reliable learning resource that hosts all of the useful materials in a single place and shows you how to prepare for this exam effectively and efficiently. There is no longer a need to alternate among the e-learning modules, the suggested textbooks in the syllabus, R markdown files, and additional online readings. Starting from the very basics and adopting a case study approach, we will learn fundamental concepts in predictive analytics, make some fancy and informative graphs in R (a powerful programming language), implement predictive models step by step in concrete settings, understand what the output in R means, and write your responses to the liking of PA exam graders. No prior knowledge in R or the SRM exam material is assumed.

P.1 About Exam PA

Exam Administrations

Exam PA (Predictive Analytics) is a 3.5-hour computer-based exam offered for the first time in December 2018 by the SOA. There are two sittings each year, one in April and one in October,[1] and each testing window lasts for four days. In April 2025, the exam will be delivered via computer-based testing (CBT) 💻 in a Prometric exam center on **April 15-18**. The registration deadline is March 11, 2025. You can check out the exam's official homepage for more information:

https://www.soa.org/education/exam-req/edu-exam-pa-detail/.

After you register for the exam online (and pay the exorbitant $1,170 exam fee! **$$$**) at

https://www.soa.org/education/exam-req/registration/edu-registration/,

you will receive an email confirmation letter ✉ from the SOA containing your candidate ID, which will allow you to schedule an appointment at Prometric (https://www.prometric.com/soa). You will also receive access to the SOA's PA e-learning modules until the end of the month in which the exam is administered (April 30 for the April sitting, October 31 for the October sitting). According to the exam homepage and syllabus, these modules

> "provide support designed to enhance candidates' knowledge from the SRM Exam learning objectives and readings"

and

> "guidance regarding knowledge and approaches that will be expected in the exam."

There are a total of 5 modules, plus an additional module that provides an introduction to R. (Well, as the previous page says, the modules are not easy to read, and with this study manual, it is not really necessary to go over the modules. 😊)

What is Exam PA Like and How to Study for It?

Typically one of the last exams students take before attaining their ASA designation, Exam PA is the first of its kind in the history of actuarial exams that heavily integrates predictive modeling, written communication, and R programming in a fully proctored setting, and this new exam style calls for a completely different approach to assessment as well as learning.

Exam format. In Exam PA, you will be asked to perform a data-driven analysis of a business problem,[2] using a combination of general tools for constructing and evaluating predictive models (e.g., training/test set split, cross-validation), and specific types of models and techniques (e.g., generalized linear models, decision trees, principal components analysis, clustering). Such an analysis does not lend itself to the multiple-choice format of many other preliminary exams you have taken, which can only elicit a simple response. Instead, Exam PA is a computer-based

[1] From 2018 to 2022, Exam PA was held in June and December.

[2] The business problem is not necessarily (and usually not) actuarial or financial in focus. Even if it is actuarially related, there is no expectation that candidates have specific product or practice-area knowledge.

written-answer exam consisting of a set of well-defined and independent tasks (usually 9 to 10 tasks), most of which are further broken down into one or more subtasks that carry 1 to 3 points each and require reasonably short answers. (You need not write long essays or reports!) The whole exam carries a total of **70 points**, with the points for each task and subtask shown at the beginning of the (sub)task *in italics*. As the exam lasts for 3.5 hours, or 210 minutes, on average you should spend $210/70 = 3$ minutes per exam point. A 10-point task, for example, should translate into approximately 30 minutes of work. If you have worked on that task for 50 minutes, then you know that it is time to move on.

Wondering where to put your written answers? The exam paper, available in Microsoft Word format, includes designated spaces labeled "**ANSWER:**" for you to type your written responses to different exam subtasks. At the end of the exam, you will upload the entire Word file for grading. You will be assessed on both the technical accuracy of your answers as well as the clarity of your thought process. Unlike other ASA exams, questions in Exam PA tend to be more open-ended and often there is not a unique best answer, as is true of predictive modeling in practice. To score high, you are expected to justify your answers carefully and adequately, based on the business problem and your prior knowledge of predictive analytics. For a written-answer exam like PA, it is important to note that:

> Credit is awarded depending on how *good* or *bad* your answers are, not (only) whether they are *right* or *wrong*.

Typical exam questions. As I said, PA consists of written-answer (rather than multiple-choice) questions. What are these questions like? To give you a first taste of the exam, here are some representative tasks taken from the latest released exams.

- **Type 1: Conceptual questions**

  Each exam has quite a number of subtasks that require you to *describe* or *explain* the predictive analytic concepts covered in the syllabus. Here are some good examples:

  ▷ October 2024 exam, Task 6 (a):

  > (*2 points*) Describe the differences between using weights and offsets in an ordinary least squares model.

  ▷ October 2024 exam, Task 10 (a):

  > (*2 points*) Describe the key assumptions of the generalized linear model (GLM).

  ▷ April 2024 exam, Task 1 (b):

  > (*2 points*) Describe the steps to calculate the within cluster sum of squares using latitude and longitude [two of the predictors].

  ▷ April 2024 exam, Task 7 (c):

  > (*2 points*) Compare and contrast using ridge vs. LASSO regression to address the overfitting concern.

▷ October 2023 exam, Task 2 (a):

(*4 points*) Describe two similarities and two differences between K-means clustering and hierarchical clustering.

▷ October 2023 exam, Task 5 (c):

(*2 points*) Describe a bivariate visualization that can be applied to understand the relationship between a numeric variable and a categorical variable.

▷ April 2023 exam, Task 5 (a):

(*3 points*) Compare and contrast single decision tree and tree-based ensemble models.

▷ April 2023 exam, Task 8 (c):

(*2 points*) Describe the process of searching for the optimal value of the hyperparameter lambda in a lasso regression.

These descriptive subtasks are good ways for the SOA to test your conceptual understanding of predictive analytics. You can secure these easy exam points simply by studying this manual (in particular, Chapter 1 and the conceptual foundations sections in Chapters 3 to 6) carefully and practicing explaining different concepts. These subtasks also mean that there are definitions and descriptions you have to memorize in advance as part of your exam preparation. There are definitely things to study!

- **Type 2: Analytical questions**

  In the majority of the exam tasks, you will examine some externally generated graphs and output, and provide explanations (e.g., why the model behaves in the way shown), interpretations (e.g., what does the output mean or imply?), or recommendations (e.g., which model is the best, in what sense?). Here are some examples:

  ▷ October 2024 exam, Task 3 (b):

  (*2 points*) Recommend how to change the cutoff value of the model to achieve the desired objective. Explain the directional impact of the change.

  ▷ October 2024 exam, Task 7 (a):

  (*2 points*) Interpret the graphic [a correlation heatmap] and identify which independent variables exhibit collinearity.

  ▷ April 2024 exam, Task 6 (a)-(b):

  (a) (*1 point*) Interpret the intercept and the coefficient for **fare**.

  (b) (*1 point*) Recommend and justify which model is better based upon the output above.

  ▷ October 2023 exam, Task 2 (b):

  (*3 points*) Explain the tradeoff between selecting a value of $K = 2$ and $K = 4$. Recommend a value for K and justify your recommendation.

▷ October 2023 exam, Task 4 (b):

(*2 points*) Interpret the Complexity Parameter table. Recommend and justify a CP value to use for the model.

▷ April 2023 exam, Task 1 (b):

(*2 points*) Explain, using the graph above, why the **Daytype** variable is statistically significant while the **DayofWeek** variable is not.

▷ April 2023 exam, Task 5 (c):

(*2 points*) Determine if this tree shows an interaction between month and year. If there is an interaction, describe it. If not, explain why there is no interaction.

Compared to tasks of type 1 above, which mainly test the ability to recall, these analytical tasks are more demanding (and interesting!), because you are required to formulate your answers based on the given output coupled with your prior knowledge of predictive analytics. It is not enough to memorize; you will have to reason and apply.

- **Type 3: Simple calculation questions** 🖩

  There are also some subtasks where you are asked to use the given output to calculate certain model quantities by hand and do some simple analysis. Examples include:

  ▷ October 2024 exam, Task 1 (b):

  (*2 points*) Calculate the impact of a 10 square foot increase on total energy usage for each community based on the model above:

  a. O'Hare

  b. Roseland

  c. Austin

  ▷ October 2024 exam, Task 3 (a):

  (*4 points*) Calculate the following values from the confusion matrix, and interpret the significance of the results.

  a. Accuracy

  b. Sensitivity

  c. Precision

  ▷ April 2024 exam, Task 5 (a):

  (*3 points*) Determine the information gain of this split using the entropy measure.

▷ April 2024 exam, Task 11 (a):

(*1 point*) Calculate how many observations your total sample will contain. Assume you can find 10 observations for each pair of regions.

▷ October 2023 exam, Task 4 (a):

(*4 points*) Calculate the change in the **Absolute Error**, using the testing data row, between the first decision tree model and building a bagged model using both decision trees. State which of these two approaches yields a better result for this observation. Show all work.

▷ October 2023 exam, Task 7 (b):

(*3 points*) Calculate the model's predicted 7-year loan repayment rate for each scenario below and show your work:

▷ October 2023 exam, Task 10 (c):

(*3 points*) Calculate the RMSE and MAE for the test data above using the tree model. Show your work.

▷ April 2023 exam, Task 8 (f):

(2 points) You are provided with the confusion matrix produced by the lasso model with a positive response cutoff threshold of 0.5.

Calculate sensitivity and specificity. Show all work.

As you can see, the shift of exam focus from working out multiple-choice problems efficiently to crafting computer-aided written responses makes Exam PA a completely different (and hopefully more enjoyable and practical!) learning experience compared with all other ASA exams you have taken. To study for this exam effectively: ⚠

It is very important to spend time *understanding* the subject, at least at a conceptual level, and learning how to *communicate* your thoughts precisely and concisely. Unlike other ASA exams, you can't expect to do well just by drilling mechanical practice problems again and again mindlessly. Instead, make an effort to *understand*, *describe*, and *explain* things. You will find that translating your thoughts into words is harder than you imagined.

(Having taught in a CAE university for about 10 years and graded hundreds of mock exams submitted by past PA students, I can say written communication is an area in which actuarial science students leave much to be desired. 😵)

New Exam Format Effective from April 2023

Ever since Exam PA was introduced in December 2018, its format and style have undergone significant changes (see Chapter 7 for more details). The latest revamp took place in the April 2023 sitting, effective from which the exam time has been reduced remarkably from 5 hours to 3.5 hours. Perhaps the more striking change is:

> Starting with the April 2023 administration, R and RStudio (a convenient platform to implement R) will not be available on the exam.

How will this "big" change affect the exam and our preparation? My answer, which is confirmed by all of the released exams following the new format (i.e., the October 2024, April 2024, October 2023, and April 2023 exams), is:

> **You will learn the material and prepare for the exam in essentially the same way, perhaps paying less attention to R code syntax. </>**

Even when R and RStudio were available on the exam from December 2018 to October 2022, Exam PA was never designed as a coding exam. Candidates did have to know *some* R, but only to the extent that they understood what the code (contained in a separate R markdown file generously provided by the SOA) was doing and knew how to make minor adjustments if necessary. The focus of the exam has always been on *understanding*, *interpretation*, and *communication*, reflected by the abundance of past exam questions belonging to Types 1 and 2 above. With R and RStudio no longer available, the only major difference is that the code and output relevant to the exam tasks will be provided directly in the exam paper; you need not take the trouble to run the code in RStudio or see the R output. The emphasis on conceptual understanding and interpretation is very likely to remain (or will even be greater).

⚠ There is one change I do expect to see in the new exam format:

> There may be more tasks of Type 3, where you have to do some simple manual calculations based on the R code and R output given.

With all the useful code and output given in the exam paper, you may be asked to explain what a certain number in the output means or how it is calculated, or to use the output to do some simple arithmetic calculations. This is a good way for the SOA to test your deeper understanding. You can't just rely on R to do everything!

Historical Pass Rates %

The following table[3] shows the number of sitting candidates, number of passing candidates, and pass rates for Exam PA since it was offered in December 2018:

(For written-answer exams, including PA, the SOA does not announce pass *marks*, i.e., the actual score you have to get to pass the exam, nor does it release the grading rubric. Yes, the grading is very much a black-box process. 📦)

Sitting	# Candidates	# Passing Candidates	Pass Rate
October 2024	2240	1467	65.5%
April 2024	2188	1413	64.6%
October 2023	2315	1468	63.4%
April 2023	2193	1552	70.8% (highest ever!)
October 2022	1554	1005	64.7%
April 2022	1171	773	66.0%
December 2021	1922	1321	68.7%
June 2021	1691	1055	62.4%
December 2020	1954	1228	62.8%
June 2020	1389	812	58.5%
December 2019	2048	1098	53.6% (I took this exam! 😅)
June 2019	1282	642	50.1%
December 2018	1042	524	50.3%

The pass rates, which are close to **65%**,[4] are higher than those of other ASA-level exams, which are typically 40-50%. Meanwhile, about 35% of the candidates failed every time, even among those who have reached this far in the ASA journey. I have heard of candidates who have failed PA twice or thrice (☹), so the exam is neither a beast nor a breeze! Worst of all, the exam is offered only twice a year, so in the unfortunate event that you do not pass, you will need to wait another six months, which adds a lot to your travel time to ASA.

[3]Some figures in the table differ slightly from those on *Actuarial Lookup* (http://www.actuarial-lookup.com/exams/pa). On occasion, a handful of missing exams were found after pass lists were posted. Some of them were passing papers and the exam statistics on https://www.soa.org/education/exam-results/ were updated at a later stage.

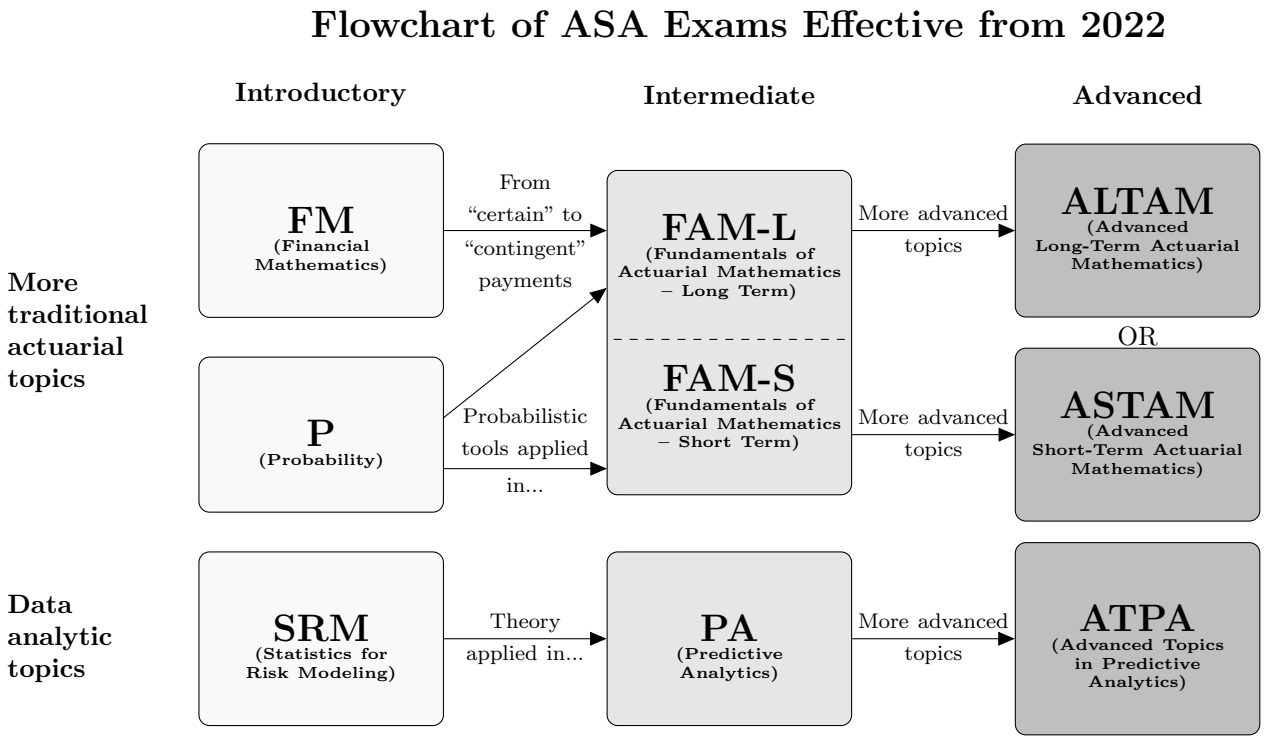
[4]They became noticeably higher after COVID-19 broke out possibly due to the temporary refund policy, which allowed students to withdraw 14 days before the exam started. This policy is likely to end soon.

Predictive Analytics Trio ♻: SRM, PA, and ATPA

Since 2018, the SOA has redesigned the ASA curriculum to reflect more contemporary and powerful predictive analytic methods that have proved useful in actuarial practice. In the current curriculum, there are a total of 3 exams (or assessments) with a heavy focus on predictive analytics: (Together, they form the recently introduced *Data Science for Actuaries* Micro-credential; see https://www.soa.org/programs/soa-ready/micro-credentials/.)

- SRM (Statistics for Risk Modeling)
- PA (where we are!)
- ATPA (Advanced Topics in Predictive Analytics)

The flowchart below shows how these 3 exams (and other ASA exams for your information) are related. While there is no set order in which the exams should be taken, students typically attempt exams from left to right, or from introductory, intermediate, to advanced. In the case of the predictive analytics trio, that means taking SRM, PA, and ATPA, in this order.



SRM vs. PA. From December 2018 to June 2021, Exam SRM was a formal prerequisite for Exam PA. Although this prerequisite is no longer in place, knowledge of the SRM materials is still assumed. As the PA exam syllabus says,

> "Exam PA assumes knowledge of probability, mathematical statistics, and selected analytical techniques as covered in Exam P (Probability), VEE Mathematical Statistics, and Exam SRM (Statistics for Risk Modeling),"

so it is reasonable to prepare for PA at the same time as or shortly after taking SRM, e.g., taking SRM in early September and PA in mid-October, or SRM in early January and PA in mid-April.

In essence, Exams SRM and PA share the same theme of working with *models*, but test it differently. As a precursor, Exam SRM is a traditional multiple-choice exam that serves to provide you with the foundational knowledge behind the modeling process. The emphasis is on the underlying theory, including the uses, motivations, mechanics, pros and cons, do's and don'ts of, and similarities and differences between different predictive analytic techniques. As a natural continuation, Exam PA will have you apply the theory you learned in Exam SRM to a business problem and see first hand how things play out. Although SRM is an important stepping stone to PA and the two exams have a rather big overlap, I would still recommend spending **at least 2 months** studying for PA intensively, even if you have taken SRM. Here are the reasons:

- *(Different skills tested)* Even though you will not apply mathematical formulas or do calculations by hand as often as in SRM, you will need time to gain hands-on experience with fitting and interpreting predictive models in R, and need practice on communicating your thoughts in writing. The written-answer format of Exam PA means that the SOA can test the material of SRM in greater breadth and depth, and assess your higher-level thinking, e.g., can you describe a certain concept or explain why something is true? You have to know how things work, at least at a conceptual level, and organize your thoughts in words.
- *(Scope)* There are some additional concepts (e.g., exploratory data analysis, elastic nets, performance metrics for classifiers, the elbow method for K-means clustering) and practical considerations that are tested in PA, but not seen in SRM. You do have to study!

PA vs. ATPA. Introduced in January 2022, the ATPA Assessment is a 96-hour take-home computer-based assessment (rather than a proctored exam) that tests additional data and modeling concepts on the basis of those in Exams SRM and PA, and consists of more involved and open-ended tasks than those in PA. As a result, ATPA is preferably taken after passing SRM and PA.

Although ATPA is a take-home assessment and 4 days seem a lot of time, you would be wise not to underestimate the amount of time and effort necessary to master the topics that can be tested, and the workload and pressure that the assessment can create. Unlike PA, which only requires some basic knowledge of R programming, proficiency with R is critical to success in ATPA. During the 96-hour window, you will spend most of your time dealing with various data issues, constructing and evaluating more advanced predictive models than those covered in PA, and finally turning your results into a written report. Make sure that you have set aside enough free time in your schedule for the next 4 days before you start the assessment. In my experience, you may need more than a day just to clean the data and get it in good shape in R before building any models. You will be busy doing coding and writing!

Note that I have written a separate study manual for ATPA. To learn more, please check out:

https://www.actexlearning.com/exams/atpa/exam-atpa-study-manual.

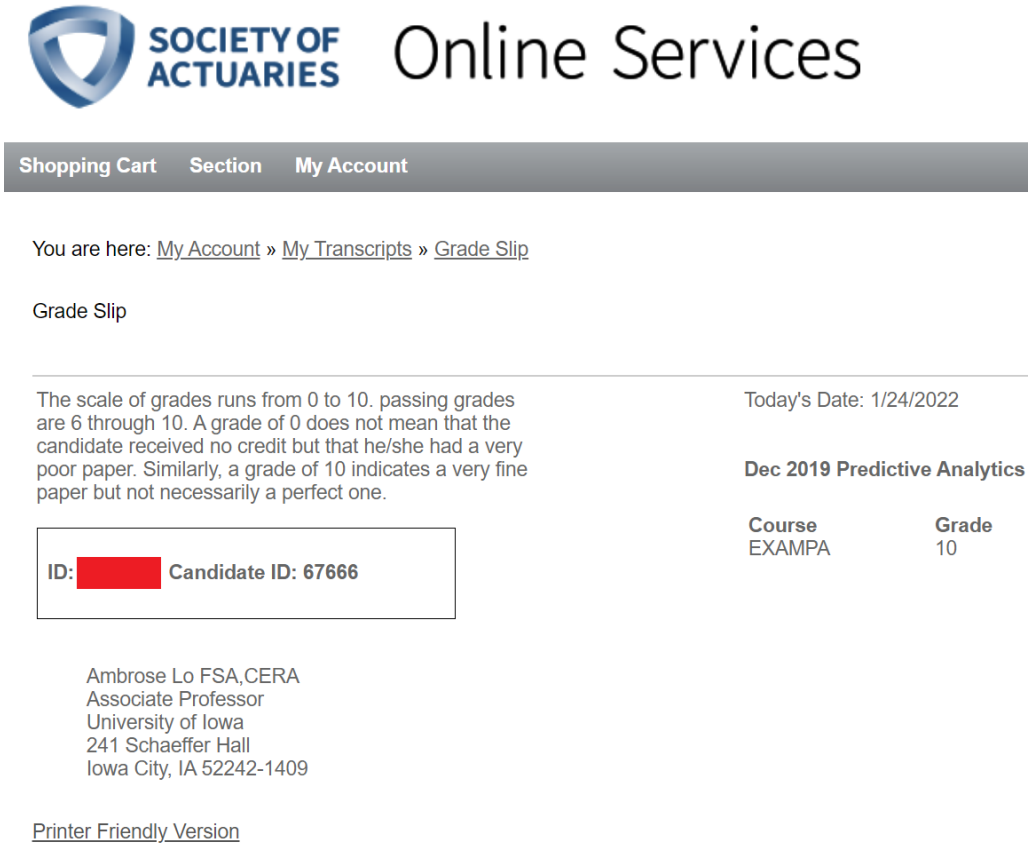
P.2 About this Study Manual

What is Special about This Study Manual?

Having been an actuarial student and teacher, I fully understand that you have an acutely limited amount of study time and that Exam PA, as a written-answer exam with a new format effective from April 2023, is not easy to prepare for. With this in mind, the overriding objective[5] of this study manual is to help you develop a conceptual understanding of and hands-on experience with the materials of Exam PA as effectively and efficiently as possible, so that you will pass the exam on your first try easily, go on to ATPA confidently, and get your ASA ASAP. Here are some unique features of this manual to make this possible.

Feature 1: The Coach DID Play!

Usually coaches don't play 😉, but as a study manual author, I took the initiative to write the **December 2019 Exam PA** and the **February-April 2023 ATPA Assessment** to experience first-hand what the real exams were like, despite having been an FSA since 2013 (and technically free from exams thereafter!). I made this decision in the belief that *teaching* for an exam and *taking* an exam are rather different activities, and braving the exam myself is the best way to ensure that this manual is indeed effective for exam preparation. If the manual is useful, then at the minimum the author himself can do well, right? I am thrilled that...


SOCIETY OF ACTUARIES Online Services

Shopping Cart Section My Account

You are here: My Account » My Transcripts » Grade Slip

Grade Slip

The scale of grades runs from 0 to 10. passing grades are 6 through 10. A grade of 0 does not mean that the candidate received no credit but that he/she had a very poor paper. Similarly, a grade of 10 indicates a very fine paper but not necessarily a perfect one.

Today's Date: 1/24/2022

Dec 2019 Predictive Analytics

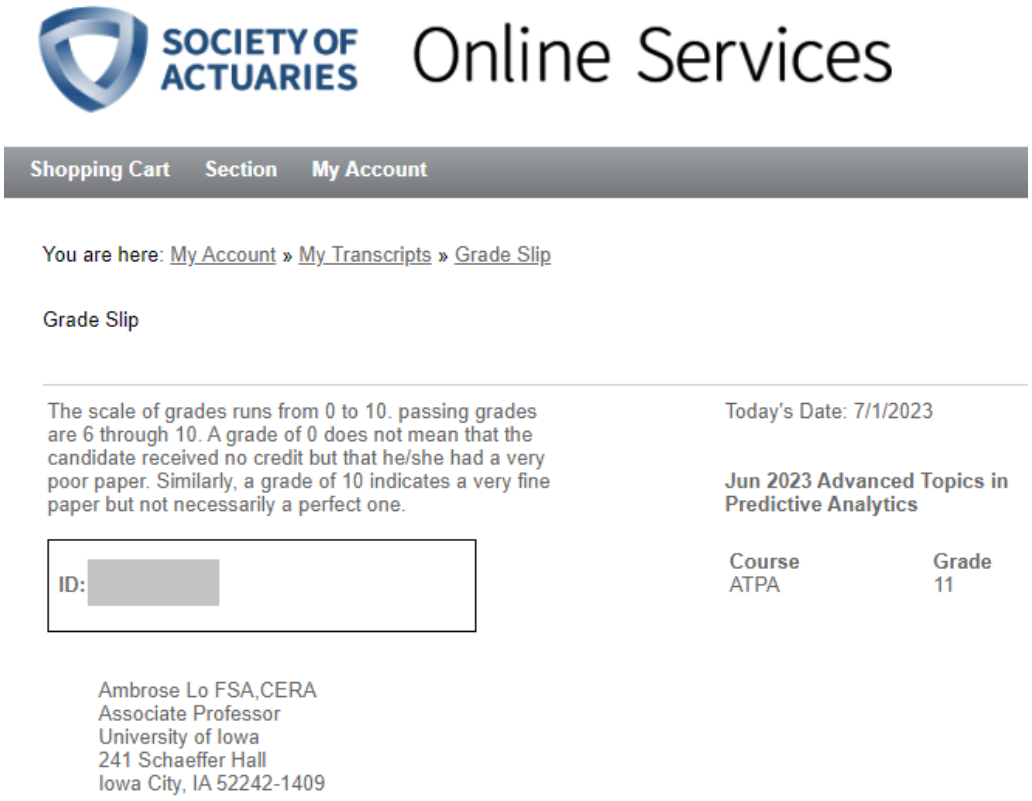
Course	Grade
EXAMPA	10

ID: [redacted] Candidate ID: 67666

Ambrose Lo FSA,CERA
Associate Professor
University of Iowa
241 Schaeffer Hall
Iowa City, IA 52242-1409

Printer Friendly Version

[5] A secondary but still important objective is to let you have some fun along the way. 😜


SOCIETY OF ACTUARIES Online Services

Shopping Cart Section My Account

You are here: My Account » My Transcripts » Grade Slip

Grade Slip

The scale of grades runs from 0 to 10. passing grades are 6 through 10. A grade of 0 does not mean that the candidate received no credit but that he/she had a very poor paper. Similarly, a grade of 10 indicates a very fine paper but not necessarily a perfect one.

Today's Date: 7/1/2023

Jun 2023 Advanced Topics in Predictive Analytics

Course	Grade
ATPA	11

ID:

Ambrose Lo FSA,CERA
Associate Professor
University of Iowa
241 Schaeffer Hall
Iowa City, IA 52242-1409

If you use this PA study manual, you can rest assured that it is written from an exam taker's perspective by a professional instructor who has experienced the "pain" of PA-ATPA candidates and truly understands their needs. Drawing upon his "real battle experience" and firm grasp of the exam topics, the author will go to great lengths to help you prepare for this challenging exam in the best possible way. You are in capable hands. 👍

Feature 2: Three-part Structure

To maximize your learning effectiveness and efficiency, I have divided this study manual into three parts:

- **Part I: Introduction to Predictive Analytics**

  To set the stage for the whole manual, this introductory part, consisting of Chapters 1 and 2, provides a general and coherent introduction to predictive analytics and presents a broad overview of the whole model building process. The discussion is deliberately designed to be generic so that it applies to virtually all predictive modeling problems. At the completion of this part, you will be equipped with the fundamental concepts that permeate predictive analytics and set foot 🚶 in some specific types of predictive models.

- **Part II: Theory of and Case Studies in Predictive Analytics**

  Armed with the foundation planted in Part I, you will learn the theory of specific predictive analytic techniques, both classical and modern (well, relatively modern!), illustrated by a series of case studies in the second part (Chapters 3 to 6), also the linchpin, of this manual. Each chapter in this part follows the same arrangement:

  ▷ *Theory:* Each chapter begins with a conceptual foundations section describing the mechanics of various predictive analytic techniques, including linear models (Chapter 3), generalized linear models (Chapter 4), decision trees (Chapter 5), and principal components and cluster analyses (Chapter 6). The explanations in these sections are thorough, but *exam-focused* and *learning-oriented*. Instead of showing unnecessary technicalities that add little value to your preparation for PA (predictive analytics can be a very mathematical subject!), I strive to follow the SOA's PA modules very closely and cover enough (but just enough) ground for you to understand predictive analytics at the level commensurate with Exam PA. You will also see a nice blend of verbal, pictorial, and mathematical expositions and a good balance between heuristic and formal presentations.

  ▷ *Practice:* After learning the ins and outs, pros and cons, and do's and don'ts of these techniques, we will turn to their practical implementations and gain some hands-on experience through a number of task-based case studies using R. Do read these case studies carefully as they illustrate a wide range of skills necessary for tackling various types of tasks in Exam PA, ranging from data pre-processing, data exploration, model construction, model evaluation, and model selection.

- **Part III: Final Preparation**

  Last but not least, the third part concludes this manual with the following resources:

  ▷ *Chapter 7:* This chapter includes my commentary on the SOA's past PA exams, which are good indicators of the SOA's expectations of PA candidates and what you will see in future exams.

  ▷ *Chapter 8:* This chapter presents two original full-length practice exams updated for the new exam format and designed to mimic the real PA exam in terms of style and difficulty, with detailed illustrative solutions provided.

  ▷ *Appendix A:* This appendix categorizes all relevant exam tasks since June 2019 (i.e., all past exams following a task-based format) by topic. A cursory glance at this appendix can reveal the themes that consistently emerge in past exams, and you may take advantage of it to identify relevant exam questions on a certain topic and make your learning more focused.

  ▷ A downloadable and printable cheat sheet, available as a separate file, provides a "helicopter" view of the entire PA exam, and is useful for both regular review and last-minute exam preparation. The cheat sheet can be accessed from

  https://www.actexlearning.com/formula-and-review-sheets.

After completing Part III, you will be fully ready to take (and pass!) the April 2025 PA exam.

Other Features

This manual throughout is also characterized by the following features that make your learning as smooth as possible:

- Each chapter in Parts I and II starts by explicitly stating which learning objectives and outcomes of the PA exam syllabus we are going to cover, to assure you that we are on track and hitting the right target.
- Objects in R are shown in `typewriter` font and code chunks with output in gray boxes for aesthetic reasons. (PA exam questions may show R objects in `typewriter` font or in **bold**.)

```
...LOTS OF R CODE HERE...
...LOTS OF R CODE HERE...
...LOTS OF R CODE HERE...
```

  Formulas, functions, and commands that are of great importance are boxed to make them stand out.

- Important exam items and common mistakes committed by students are highlighted by pinkish red boxes that look like:

> **EXAM NOTE**
>
> Be sure to pay special attention to boxes like this!

- The main text of this manual is interspersed with more than 110 exercises, all with complete solutions, to assess your understanding regularly. Some of these exercises are based on recent SOA and CAS exams, but many are original. (If you have used the *ACTEX Study Manual for Exam SRM*, you may have seen some of these past exam questions in some form, but I have rewritten many of them in the language and style of Exam PA. There is also no harm in giving them a second look!) These examples are instrumental in illustrating a number of conceptual and analytical items that can be tested in Exam PA.
- Each chapter in Parts I and II concludes with a number of review questions (more than 110 conceptual review questions and 40 integrated review questions, to be released, in total) designed to help you look back on the important concepts and techniques covered in that chapter.

What is New in the 12th edition of the Manual?

Albeit relatively established, this manual is periodically "re-trained" taking the latest PA exams and students' feedback into account to improve its "predictive power." The 12th edition of the manual has seen substantial updates in terms of clarity, substance, and exam focus, but here are the most significant improvements:

- The most noticeable change is that sections which are dedicated predominantly to R programming and occupy the first two chapters of previous editions of the manual are now relegated to a newly created appendix, Appendix R, and the manual begins directly with a general discussion on predictive analytics. In response to the diminishing role played by R in the current exam format, this arrangement should make the development of the whole manual more logical and coherent, and, more importantly, make students' learning more effective and enjoyable.
- In view of the rearrangement above, some of the instructional videos that accompany Chapters 1-3 will be reworked, also to improve their content and clarity, and released on a rolling basis.
- Another highlight is that 40-50 new review questions of an integrated nature will be added to each chapter in Parts I and II and released together with the commentary on the October 2024 PA exam (Section 7.1) after the SOA posts the model solution online.
- Exam note, exercise, and task boxes are colored to make reading more pleasant. We need more colors to enrich our study! 😀
- Chapter 1: This chapter starts with a study guide that provides learning strategies for making the most of this manual. You can also find the distinction between nominal and ordinal categorical variables in Section 1.1.
- Section 1.3: Some real-life analogues of underfitting and overfitting are added on page 43, mostly for amusement.
- Chapter 2: Section 2.1 features clearer explanations of the pros and cons of summary statistics and graphical displays, and Section 2.2 now ends with a comparison among the three types of split bar charts and considerations that make a graph clearer.
- Section 6.1: The old Figure 6.1.1 is broken down into two figures (Figures 6.1.1 and 6.1.2) to explain the geometric meaning of PCs more effectively. The description of what a biplot is in Task 3 of Subsection 6.1.3 is also clarified.
- Subsection 6.2.2: Four diagrams are added to provide a schematic illustration of the inter-cluster dissimilarity between two clusters for each type of linkage, and a second solution to Example 6.2.1 based on a distance matrix is presented.
- A fair number of new in-text exercises and end-of-chapter conceptual review questions are added: 1.1.1, 2.2.1 (rewritten), 2.2.2; 1.2, 1.12, 2.1-2.12, 5.8, 6.3, 6.9, 6.11, 6.21, 6.24
- Appendix A: Questions of the October 2024 exam have been added and categorized.

Supplementary Files

This study manual comes with a number of supplementary files, e.g., R Markdown files with completely reproducible R code, datasets, and files to be released, that can all be downloaded or accessed from *Actuarial University*. All users of the manual (whether it is the printed or digital version) will receive by email a keycode that provides electronic access to all supplementary files shortly after their order is placed. If you can't retrieve that email (be sure to check your junk/spam folders), please reach out to `support@actexlearning.com` for assistance.

It is a good idea (but not absolutely essential, given the new exam format) to run the R Markdown files as you work through this manual, making sure that your output agrees with what is shown here. This is especially important if you have ordered a printed copy of this study manual—run the code to see the beautiful colors! ☺

⚠ NOTE ⚠

Commentary on the October 2024 PA exam (the SOA's model solution not yet available at the time of writing) and the 40-50 integrated review problems will be available on *Actuarial University* shortly after the SOA posts the exam with solutions online.

Two Add-ons

If you have purchased this manual and are interested in upgrading your product to include any of the following add-ons, please email Customer Service at `support@actexlearning.com`.

(1) Instructional videos. Instructional videos (https://www.actexlearning.com/exams/pa/exam-pa-study-manual) accompanying the core of this manual (Parts I and II, or Chapters 1 to 6) are available for purchase as an add-on. In these videos, I (Ambrose) will walk you through the fundamental concepts in predictive analytics and the construction of predictive models in R step by step, with a strong emphasis on key test items in Exam PA. With the aid of visuals, these videos aim to make the materials in the manual as accessible as possible and will add substantial value to your learning.

When it comes to learning strategies, some students find it useful to watch the videos to get the "big picture," then read the manual to learn the details. Alternatively, you may first read the manual, then watch the videos to consolidate your understanding. Both modes of learning are perfectly fine and which one is better depends entirely on your preferences.

⚠ NOTE ⚠

The instructional videos support and add value to, but do NOT fully replace this study manual, which should be the backbone of your study program.

(2) Graded mock exam. In addition to the two practice exams in Chapter 8 of this manual, we offer a separate mock exam (https://www.actexlearning.com/exams/pa/exam-pa-mock-exam), with completely different questions, and an optional 1:1 live feedback session.

A common "complaint" against Exam PA is that its written and somewhat open-ended exam format makes it difficult for students to self-evaluate their work even after reading the SOA's model solutions to past exams, e.g., if you write this, how many points can you expect to get? How to improve your answers? This is precisely why we create this mock exam with grading service, which provides a valuable opportunity for you to assess your overall understanding of the PA exam syllabus and, more importantly, have your work graded from a **critical eye**, and receive **personalized feedback (not generated by AI in any way!)**. You will work on the mock exam under simulated exam conditions and submit your solutions to us. Having taken PA in the past and now teaching for PA, we (Ambrose and his team) will then grade your work from start to finish, with a score out of 70, and offer specific feedback that will help you enhance the quality of your write-up and improve your performance on the real exam.

For the April 2025 sitting, the graded mock exam (currently available for pre-order) is expected to be released on *Actuarial University* in early February and the last day of submission is March 31, 2025 (Monday). Within 2 weeks of your submission, you will receive the following by email:

(1) Your graded mock exam with personalized (and possibly critical!) feedback

(2) Detailed illustrative solutions to the mock exam along with grading rubric

Contact Us

If you encounter problems with your learning, we always stand ready to help.

- For **technical issues** (e.g., not able to access, download, or print supplementary files from *Actuarial University*, extending your digital license, upgrading your product, exercising the Pass Guarantee), please email ACTEX Learning's Customer Service at

  `support@actexlearning.com`

  The list of FAQs on https://www.actuarialuniversity.com/help/faq may also be useful.

- For questions related to **specific contents** of this manual and Exam PA, including potential errors (typographical or otherwise), please feel free to raise them in the PA community on ACTEX's Discord channel, which provides a convenient platform for you to network with other PA students (and me!), and I will strive to respond to your questions ASAP. Please note:

  ▷ A list of errata (if any) will be maintained and posted in the PA community. I would greatly appreciate it if you could bring any potential errors, no matter how minor, to my attention so that they can be fixed in a future edition of the manual.

▷ Instead of saying

"You mention (somewhere) in your manual that...,"

it would be great to quote the specific page(s) of the manual your questions are about. This will provide a concrete context and make our discussion much more fruitful.

▷ *(Less important in the new exam format)* If you experience issues with R, e.g., your code can't run and you keep seeing weird error messages, please provide the version of R (not RStudio!) you are using and a screenshot of the error messages.

Acknowledgments

I am grateful to Mr. Tony Pistilli for proofreading an early version of this study manual and many past students for taking the time to send me comments and suggestions, which have improved the quality of the manual in no small measure. All errors that remain are solely mine.

About the Author

Ambrose Lo, PhD, FSA, CERA, is the author of several study manuals for professional actuarial examinations and an Adjunct Associate Professor at the Department of Statistics and Actuarial Science, the University of Hong Kong (HKU). He earned his BSc in Actuarial Science (first class honors) and PhD in Actuarial Science from HKU in 2010 and 2014, respectively, and attained his Fellowship of the Society of Actuaries (FSA) in 2013. He joined the Department of Statistics and Actuarial Science, the University of Iowa (UI) as Assistant Professor of Actuarial Science in August 2014, and was promoted to Associate Professor with tenure in July 2019. His research interests lie in dependence structures, quantitative risk management as well as optimal (re)insurance. His research papers have been published in top-tier actuarial journals, such as *ASTIN Bulletin: The Journal of the International Actuarial Association*, *Insurance: Mathematics and Economics*, and *Scandinavian Actuarial Journal*. He left the UI and returned to Hong Kong in July 2023.

Besides dedicating himself to actuarial research, Ambrose attaches equal (if not more!) importance to teaching and education, through which he nurtures the next generation of actuaries and serves the actuarial profession. He has taught courses on a wide range of actuarial science topics, such as financial derivatives, mathematics of finance, life contingencies, and statistics for risk modeling. He is also the (co)author of the *ACTEX Study Manuals for Exams ATPA, MAS-I, MAS-II, PA, and SRM*, a *Study Manual for Exam FAM*, and the textbook *Derivative Pricing: A Problem-Based Primer* (2018) published by Chapman & Hall/CRC Press. Although helping students pass actuarial exams is an important goal of his teaching, inculcating students with a thorough understanding of the subject and logical reasoning is always his top priority. In recognition of his outstanding teaching, Ambrose has received a number of awards and honors ever since he was a graduate student, including the 2012 Excellent Teaching Assistant Award from the Faculty of Science, HKU, public recognition in the *Daily Iowan* as a faculty member "making a positive difference in students' lives during their time at UI" for nine years in a row (2016 to 2024), and the 2019-2020 Collegiate Teaching Award from the UI College of Liberal Arts and Sciences.

Part I

Introduction to Predictive Analytics

Chapter 2

Data Exploration and Visualization

*****FROM THE PA EXAM SYLLABUS*****

2. Topic: Data Exploration and Visualization (20-30%)

Learning Objectives
The Candidate will be able to work with various data types, understand principles of data design, and construct a variety of common visualizations for exploring data.

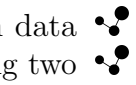
Learning Outcomes
The Candidate will be able to:

d) Apply the key principles of constructing graphs.

e) Apply univariate data exploration techniques.

f) Apply bivariate data exploration techniques.

Chapter overview: As discussed in Stage 3 in Section 1.2, an integral part of any predictive analytic exercise is to explore the characteristics of the variables in the data, both on their own and in relation to one another, by means of such tools as summary statistics and graphical displays. Synthesizing the material in the first four chapters of the book *Data Visualization: A Practical Introduction* (which is listed in the PA exam syllabus), this chapter draws upon R's functions in the base installation and the more specialized `ggplot2` package to perform data exploration, more formally called ***exploratory data analysis*** (EDA), with the following two important goals:



Q2.1

(1) *Data validation*

From a technical point of view, EDA allows us to perform commonsense checks on the data and identify outright nonsensical data values (e.g., a negative value for age, which is impossible), which are potential data errors that may lead to unreasonable model results, and outliers that merit further consideration. After inappropriate and anomalous data values have been removed or processed, the data becomes ready for analysis.

(2) *Generating insights for modeling*

From a predictive analytic point of view, EDA is even more valuable. It helps us understand the characteristics of and relationships between the variables in the data. Such an understanding has substantial modeling implications, e.g., to identify potentially useful predictors of the target variable, to generate new features that may improve the prediction performance and interpretability of the predictive models we will construct, to decide on an appropriate type of predictive model that suits our needs.

Statistics + graphs. Typically, EDA is accomplished by a *combination* of two kinds of tools:

(1) *Descriptive statistics* (a.k.a. summary statistics) that quickly summarize different distributional properties of the variable(s) of interest by a set of numbers

Examples: Mean, variance, mode, correlation, table of frequency counts

(2) *Graphical displays* (a.k.a. graphical representations, visual displays) that turn data into a visual format, which in turn promotes understanding and insights

Examples: Histograms, boxplots, bar charts, and their variants

Q2.2 These two types of tools complement each other and have their relative pros and cons.

	Summary Statistics	Graphical Displays
Pros	• Precise and objective (a number is a number!) • Easily comparable across variables, e.g., based on the mean, we can say one variable X tends to be higher than another variable Y.	• Users can gain a quick visual impression of the distribution(s) of the variable(s) of interest holistically. • Can reveal information not easily captured by summary statistics, e.g., the presence of multiple modes, non-linear relationships
Cons	• Can only capture a certain aspect of a variable's distribution, not the full picture • Some statistics (e.g., mean) can be easily distorted by outliers.	• Information provided is not as precise as summary statistics (exact values/levels of the original observations are often lost). • Less comparable across variables, e.g., not always easy or possible to say one variable tends to be smaller than another variable based on a plot. • For complex data, graphs can be hard to read and interpret.

⚠ EXAM NOTE ⚠

EDA is tested in EVERY PA exam. The following tasks from the three most recent PA exams will convince you of how important EDA is.

October 2024 PA Exam: Task 9

Your assistant graphs the number of days that the maximum temperature exceeds different levels in each month against monthly residential Kilowatt hours for April through October.

(b) (*2 points*) Briefly summarize the relationship between monthly residential KWH usage and days in a month exceeding a specific temperature based on the graph above.

April 2024 PA Exam: Task 4

Your manager is interested in the relationship between the market share held by the airline with the highest market share between two cities (**large_ms**) and the average fare between those cities (**fare**). Your assistant produces the graphs below.

(a) (*3 points*) Describe one pro and one con of each visualization in explaining the relationship between **large_ms** and **fare**.

(b) (*2 points*) Recommend a visualization for your assistant to create to understand the modeling implications of the graphic above.

October 2023 PA Exam: Task 5

Your client is interested in obtaining a deeper understanding of how tuition prices and the size of the universities are reflected in the dataset.

(a) (*3 points*) Suggest two numerical variables from the Data Dictionary for this analysis and describe two univariate techniques that can be used to explore them.

(b) (*2 points*) Suggest a categorical variable from the Data Dictionary for this analysis and describe a univariate technique to explore the variable.

(c) (*2 points*) Describe a bivariate visualization that can be applied to understand the relationship between a numeric variable and a categorical variable.

(d) (*2 points*) Interpret the plot above.

Case study: Personal injury insurance dataset. To illustrate data exploration and visualization techniques, in this chapter we will look at a personal injury insurance dataset.[1] This dataset contains the information of 22,036 settled personal injury insurance claims. These claims were reported during the period from July 1989 to the end of 1999, with claims settled with zero payment excluded. The variables in the dataset are described in Table 2.1.

Variable	Description
`amt`	settled claim amount (a continuous numeric variable)
`inj`	injury code, with seven levels: 1 (no injury), 2, 3, 4, 5, 6 (fatal), 9 (not recorded)
`legrep`	legal representation (a binary variable; 0 = no, 1 = yes)
`op_time`	operational time (a standardized amount of time elapsed between the time when the injury was reported and the time when the claim was settled)

Table 2.1: Data dictionary for the personal injury (`persinj`) insurance claims dataset.

In Section 4.2, we will build a generalized linear model to predict the size of personal injury insurance claims using other variables in the dataset. For now, we will content ourselves with data exploration and visualization. The insights we gain here will go a long way towards constructing a good predictive model.

To get started, run CHUNK 1 to load an external CSV file containing the dataset into R as a data frame called `persinj` (meaning "personal injury") using the `read.csv()` function, which takes the name of the CSV file supplied as a character string as an argument.

```
# CHUNK 1
persinj <- read.csv("persinj.csv")
```

ONE MORE REMINDER!

Please read page xxviii of the preface of this manual about how to access the Rmd files as well as datasets that go with this manual. There is absolutely NO need to copy the R code in the manual to RStudio or type the code manually line by line!

2.1 Univariate Data Exploration

Let's begin with *univariate* data exploration—exploration that sheds light on the distribution of one and only one variable at a time. The specific summary statistics and graphical tools to use

[1] This dataset is a pre-processed version of the `ausautoBI8999` data in the `CASdatasets` package, which in turn accompanies the textbook *Generalized Linear Models for Insurance Data* (2008), by de Jong and Heller.

will depend on whether the variables you are analyzing are numeric or categorical (recall the definitions of numeric and categorical variables given on page 11). Both types of variable are part of the dataset in a typical PA/ATPA exam and in real life.

2.1.1 Numeric Variables

Descriptive statistics. Summary statistics are mainly used to reveal two aspects of the distribution of a **numeric variable**:

- *Central tendency:* The central tendency of a numeric variable, be it continuous or discrete, is often quantified by its **mean** and **median**. These two metrics capture, in a loose sense, the typical "size," or center, of the variable and can be readily produced in R by applying the `summary()` function to the variable of interest.
- *Dispersion:* Common measures of dispersion include **variance**, **standard deviation**, and **inter-quartile range** (defined as the difference between the 75% quantile and the 25% quantile), all of which measure in a way how spread out the values of the numeric variable are over its range. In R, we can use the `sd()` function to determine standard deviations.

In the `persinj` data, there are two numeric variables, `amt` and `op_time`. In CHUNK 2, we focus on the `amt` variable for the purposes of illustration and generate its "summary" statistics using the `summary()` function.

```
# CHUNK 2
summary(persinj$amt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10    6297   13854   38367   35123 4485797

sd(persinj$amt)

## [1] 90981.11
```

When the `summary()` function is applied to a numeric variable, a six-number "summary" is produced. We can see that the mean of claim amount (38,367) is way higher than its median (13,854) and the 75th percentile is much further away from the median than the 25th percentile, indicating that the distribution of claim amount is highly skewed to the right. The right skew suggests that the values to the right of the mean of claim amount tend to be further away from the mean than those to the left, so there is a heavy tail that extends to the far right, and some claim amounts are extraordinarily large. In fact, the largest claim amount, 4,485,797, is almost an astronomical figure compared with the mean or median.

In the following exercise, you will calculate the summary statistics for the two groups of injuries classified by legal representation, which is a binary categorical variable. Doing so helps us understand the effect of legal representation on claim amount.

Exercise 2.1.1. **(Calculating the summary statistics for two groups of observations)** Write R code to determine the summary statistics for claim amount separately for injuries with legal representation and those without legal representation. Comment on the central tendency and dispersion of claim amount for these two groups of injuries.

Solution. To extract the two groups of injuries, we can use the method of logical subsetting (see Section R.2) to split the dataset into two subsets:

(1) A subset called `persinj.0` corresponding to injuries without legal representation (`legrep = 0`)

(2) A subset called `persinj.1` corresponding to injuries with legal representation (`legrep = 1`)

Then we look at the summary statistics of the claim amount variable within the two subsets. This is done in CHUNK 3.

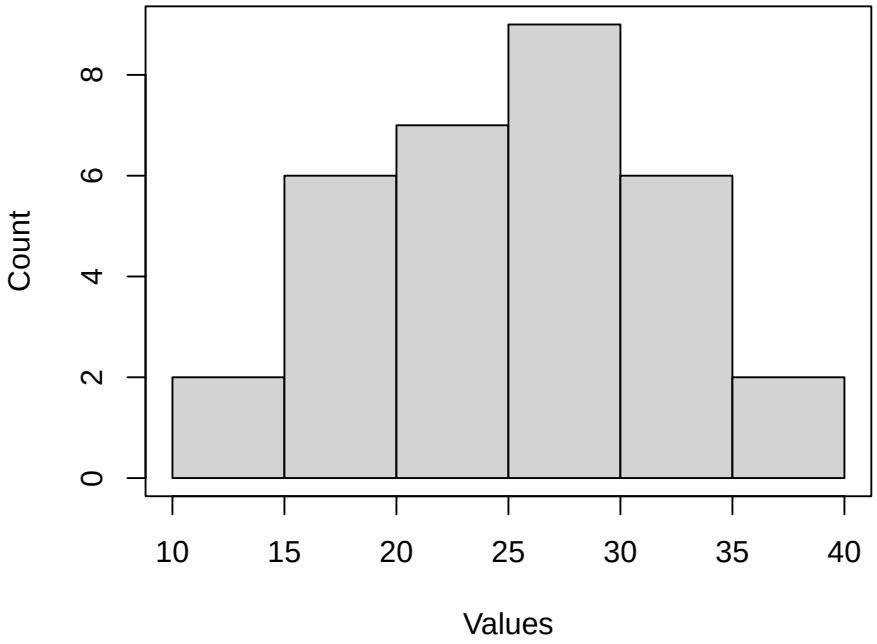
```
# CHUNK 3
persinj.0 <- persinj[persinj$legrep == 0, ]
persinj.1 <- persinj[persinj$legrep == 1, ]
summary(persinj.0$amt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10    4061   11164   32398   29641 2798362

summary(persinj.1$amt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      20    7305   15309   41775   38761 4485797

sd(persinj.0$amt)

## [1] 77820.33

sd(persinj.1$amt)

## [1] 97541.38
```

The comparison is apparent: Claims with legal representation not only are larger on average (perhaps with legal advice the claimant is able to fight for larger settled claims), but also are more spread out. □

Histograms. Now we turn to visual representations. To visualize the distribution of a numeric variable, a convenient graphical tool is a ***histogram***, which divides the observations of the Q2.3 variable into several equally spaced bins (a.k.a. buckets) and provides a visual summary of the

number of observations in each bin. Looking at a histogram, we can learn about the overall shape of the distribution of the variable and where most observations lie.

Here is a prototypical histogram, where the distribution ranges between 10 and 40, and is roughly symmetric around 25.



In the rest of this manual, we will be following the PA modules and mostly using functions in the `ggplot2` package to make graphs. In CHUNK 4, we install and load the package. (An R note: Make sure to install a package the first time you use it!)

```
# CHUNK 4
# Uncomment the next line the first time you use ggplot2
# install.packages("ggplot2")
library(ggplot2)
```

With the last command, we can use all the functions in the `ggplot2` package until the end of the current R session.

⚠ NOTE ⚠

In the rest of this chapter, you will see quite a lot of R code based on functions in the `ggplot2` package and some explanatory comments, for the benefit of those who are interested in reproducing the graphs and learning `ggplot2` for their own work. The technicalities of the (somewhat complicated) code syntax are covered in more detail in Section R.5, which you may look at if you are interested. For learning and exam prep purposes, try to focus on **why we use a certain plot for a certain purpose**, and **what you can make of a given plot**, which are important exam items in PA.

In `ggplot2`, a histogram is constructed by the `geom_histogram()` function, which requires the use of the `x` aesthetic to capture the numeric variable of interest. (The notion of aesthetic mappings is covered in Subsection R.5.1.) In CHUNK 5, we make four histograms colored in blue for the claim amount variable (Figure 2.1.1) with different choices of the `bins` parameter, which controls the number of bins in a histogram. Note that:

- To produce better visual effects, we have restricted the range of the horizontal axis to be between 0 and 100,000.[2] (Try to see how the histograms look if the command `xlim(0, 100000)` is lifted.)
- The `fill` argument is for coloring the *interior* of shapes. Had we used the `color` argument and typed `color = "blue"` inside the `geom_histogram()` function, only the border lines of the vertical bars would be in blue and the interior would remain gray. (Try it!)
- One thing that differentiates a ggplot from a plot created from the base R installation is that a ggplot can be saved as an object in R and manipulated further. Using the `grid.arrange()` function in the `gridExtra` package, we can place several ggplots (four histograms, `p1`, `p2`, `p3`, and `p4`, in this case) in a single figure for ease of comparison.

By default, `geom_histogram()` chooses the number of bins based on a rule of thumb (which is 30 here). The value of `bins` should be small enough to highlight interesting features of the distribution, but large enough to visualize the distribution accurately. When plotting histograms, it is a good idea to experiment with a few values of `bins` as different values of the parameter can reveal quite different patterns. In Figure 2.1.1, 30 (default) and 40 are both fine values for `bins` while 10 makes the histogram too crude and 80 may have described the data too finely. Regardless of the value of `bins`, the histograms corroborate the earlier findings in CHUNK 2 that the distribution of claim amount is heavily lopsided to the right with a pronounced tail.

```
# CHUNK 5
library(ggplot2)
p1 <- ggplot(persinj, aes(x = amt)) +
  geom_histogram(bins = 10, fill = "blue") +
  xlim(0, 100000) +
  ggtitle("Bins = 10")
p2 <- ggplot(persinj, aes(x = amt)) +
  geom_histogram(fill = "blue") +
  xlim(0, 100000) +
  ggtitle("Bins = 30 (default value)")
p3 <- ggplot(persinj, aes(x = amt)) +
  geom_histogram(bins = 40, fill = "blue") +
  xlim(0, 100000) +
  ggtitle("Bins = 40")
p4 <- ggplot(persinj, aes(x = amt)) +
```

[2] In fact, if you use the `xlim` and `ylim` functions directly without using the `coord_cartesian()` function, you are in effect looking at the *conditional* distribution of the data given that the observations lie within the ranges specified.

```
  geom_histogram(bins = 80, fill = "blue") +
  xlim(0, 100000) +
  ggtitle("Bins = 80")
```

```
# CHUNK 5 (Cont.)
# uncomment the next line the first time you use the gridExtra package
# install.packages("gridExtra")
library(gridExtra)
grid.arrange(p1, p2, p3, p4, ncol = 2)
```

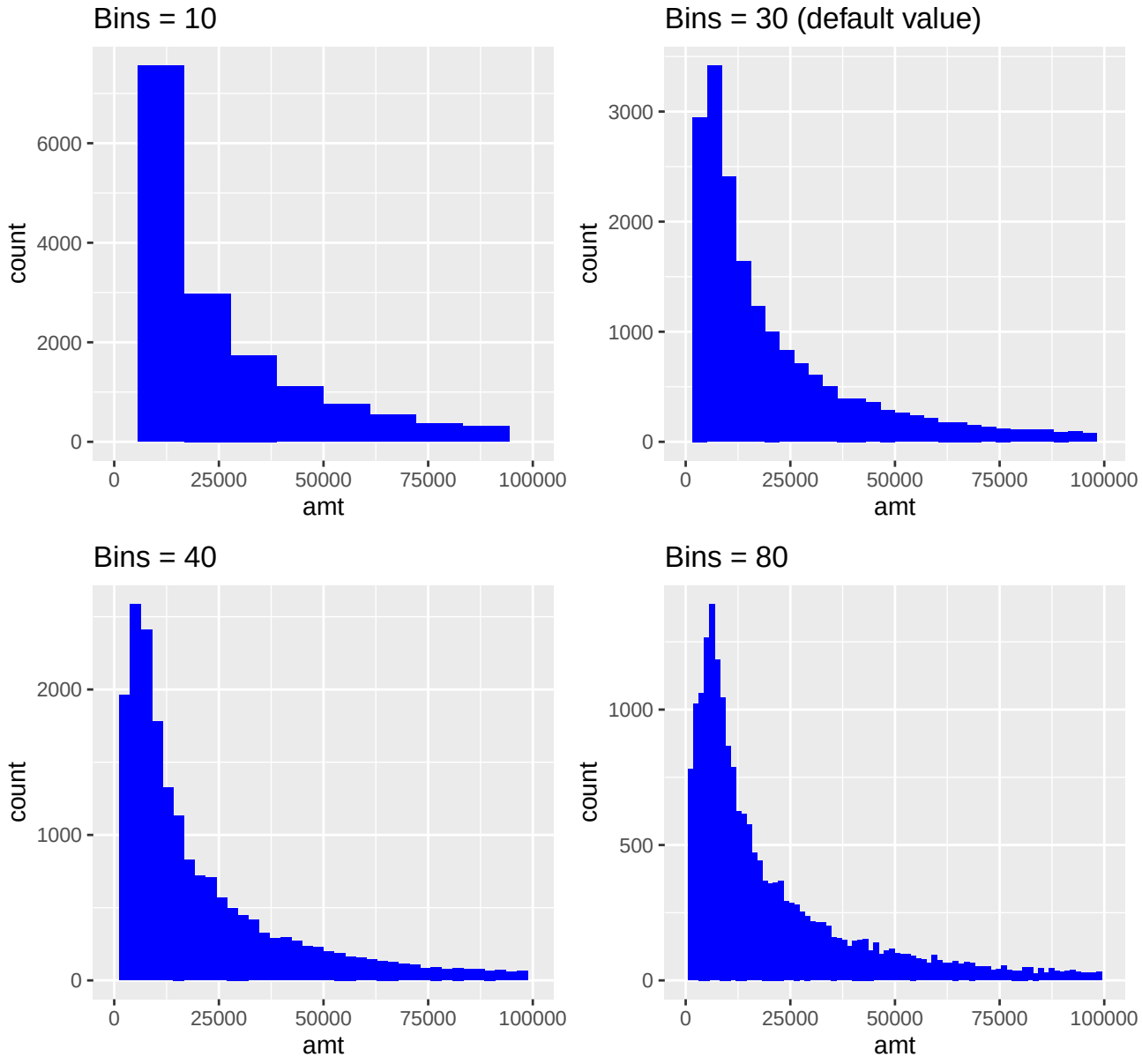


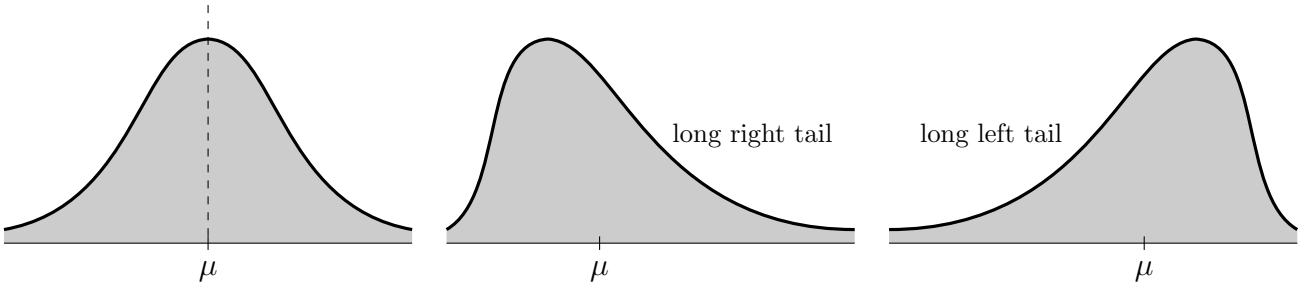
Figure 2.1.1: Four histograms of `amt` in the `persinj` dataset with different values of `bins`.

An aside: Problems with skewed data and possible solutions. The statistics in CHUNK 2 and the histograms in Figure 2.1.1 extending to the far right indicate that the claim amount variable has a prominent right skew. You may recall from Exam P/FAM/(A)STAM that **skewness** is a measure of the asymmetry of a distribution. For comparison purposes, it may help to distinguish the three cases below:

Case 1. For a symmetric distribution, i.e., a distribution whose histogram is symmetric about the mean μ, values above μ are, on average, of the same distance from μ as those below; see Figure 2.1.2 (a).

Case 2. A *right-skewed* distribution is one for which values above the mean tend to be further away from the mean than those below, leading to a histogram with a "long tail" that extends to the right; see Figure 2.1.2 (b).

Case 3. A *left-skewed* distribution is the opposite: Values below the mean tend to be further away from the mean than those above, creating a histogram with a long left tail; see Figure 2.1.2 (c).



(a) A symmetric distribution. (b) A right-skewed distribution. (c) A left-skewed distribution.

Figure 2.1.2: Graphical illustrations of the density functions (or histograms) of symmetric, right-skewed, and left-skewed distributions.

In real modeling work, it is *right*-skewed distributions (to different degrees) that arise most frequently, especially in insurance and financial applications. Many variables such as income and loss amount are, by design, non-negative, so there is a limit to how long the left tail can be, but the right tail may be extraordinarily thick—the largest values, called the **outliers**, can be astronomically large compared to the mean.

There is no universal quantitative definition of outliers. We generally think of outliers as anomalous data points that substantially differ from the overall pattern of the data and appear to be strange. (For categorical variables, observations in sparse factor levels can be considered outliers.) They typically arise in two ways:

- *Errors:* Outliers can arise due to errors in the data collection process, such as data entry errors. They are observations whose values are so ridiculous that they can be safely dismissed as erroneous, such as a negative age and a current customer who was born in 1800. If an outlier is erroneous, then it is reasonable to correct the error or remove it entirely.

- *Natural:* Outliers can also arise naturally. Not necessarily errors, they are simply observations whose values are far away from the rest of the data, but are possible in theory. Examples include a policyholder currently aged 120 and a legendary actuary earning $3 million a year. (Well, some actuaries are insanely rich!)

The presence of a right-skewed variable (or outliers) can be problematic for two reasons: Q2.4

- *(Model fitting)* A number of predictive models (e.g., linear models, regression trees, to be covered in Chapters 3 and 5, respectively) are fitted by minimizing the sum of the squared discrepancies between the observed values and fitted values of the target variable:

$$\min \sum_{i=1}^{n_{\text{tr}}} (y_i - \hat{y}_i)^2.$$

  If the target variable is right-skewed, then the outliers will contribute substantially to the sum due to the squaring and have a disproportionate effect (or "leverage") on the whole model. For example, a difference of 1,000 gets squared to 1,000,000, which is way larger. The trained model will unwittingly pay more attention to fitting the outliers at the expense of other "well-behaved" observations. This may be undesirable unless the right tail of the target variable is where our main concern lies.

- *(Predictive power)* After all, our objective in predictive analytics is to study the association between the target variable and predictors in the data over a wide range of variable values. If most of the observations of the target variable cluster narrowly in the small-value range and a handful take extraordinarily large values, this will make it difficult to investigate the effect of the predictors on the target variable globally—we simply don't know enough about the target variable in the right tail.

  The same idea applies to a right-skewed predictor. If a predictor exhibits a heavy right skew, we are unable to differentiate the observations of the target variable effectively on the basis of the values of the predictor, most of which are concentrated in the small-value range.

How do we cope with outliers? Here are some possible options:

- *Remove:* If we (somehow!) know that a natural outlier is not likely to have a material effect on the final model, then it is fine to remove it.
- *Keep:* If the outliers make up only an insignificant proportion of the data and are unlikely to create bias, then it is sensible to leave them in the data.
- *Modify:* We can also modify the outliers to make them more reasonable, like censoring the policyholder age of 120 at 100.
- *Using robust model forms:* Instead of minimizing the squared error between the predicted values and the observed values, we could replace the squared error by the *absolute error*:

$$\min \sum_{i=1}^{n_{\text{tr}}} |y_i - \hat{y}_i|,$$

which places much less relative weight on the large errors and reduces the impact of outliers on the fitted model.

(This is related to the concept of a loss function discussed on page 31.)

⚠ EXAM NOTE ⚠

The above methods for handling outliers have been removed from the PA modules effective from the April 2023 exam sitting, but I'm still including them here because they are more or less common sense and **might** be tested.

The transformation methods below are far more important to know. In fact, the log transformation is one of the (if not, THE) most commonly used variable transformations in Exam PA and in applied modeling in general.

An alternative to the solutions above is to apply a monotone concave function to the whole variable in order to shrink the outliers relative to the smaller values and symmetrize the overall distribution, while preserving the ranks of the observed values of the variable. This will dampen the effects of the outliers on the fitted model and tend to improve its overall goodness of fit compared to using the original right-skewed variable. Two commonly used transformations for dealing with right-skewed variables are:

Q2.5

- ***Log transformation:*** This transformation (typically with respect to base e) can be applied as long as the variable of interest is *strictly* positive. None of the variable values should be zero or negative. (Remember that $\log x$ is not well-defined if $x \leq 0$!)
- ***Square root transformation:*** Although not discussed in the PA modules, the square root transformation $\sqrt{x}$ is in a similar spirit to the log transformation, but is applicable even to non-negative variables, some of whose values can be zero.

⚠ Note that transformations can serve as data exploration tools and/or predictive modeling tools separately. That is, we can make graphical displays for the transformed variables to identify and assess the extent of skewness in the EDA stage, and we can, but not necessarily have to, use them to take care of the skewness when constructing predictive models. Some models like GLMs (see Chapter 4) are capable of handling skewness on their own, without transforming the skewed variables.

To see the effects of the log and square root transformations in action, CHUNK 6 produces a histogram for the log of claim amount and the square root of claim amount for the `persinj` dataset. It is evident that the log transformation has effectively removed the right skewness of claim amount and made the resulting distribution much more symmetric. In contrast, the square root of claim amount remains rather right-skewed. In general, the log transformation does a better job of remedying the right skewness of a variable than the square root transformation, but it may overdo things and make the transformed variable left-skewed (fortunately, not the case in Figure 2.1.3).

```
# CHUNK 6
p1 <- ggplot(persinj, aes(x = log(amt))) +
  geom_histogram()
p2 <- ggplot(persinj, aes(x = sqrt(amt))) +
  geom_histogram()
grid.arrange(p1, p2, ncol = 2)
```

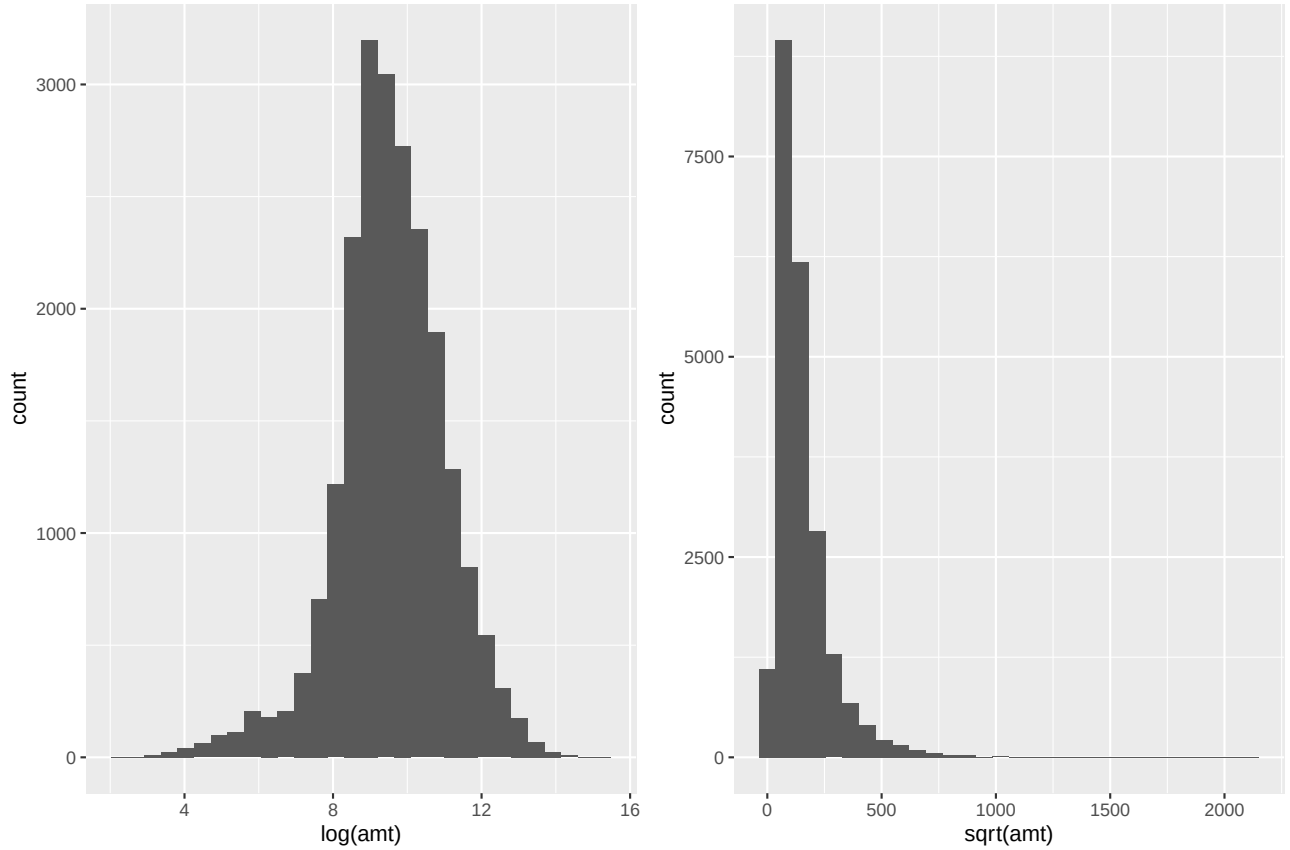


Figure 2.1.3: Histograms of the log of claim amount (left) and square root of claim amount (right) in the `persinj` dataset.

Because of the extreme skewness of claim amount in the `persinj` data, the log transformation is the more appropriate transformation to use and we will adopt it in the rest of this chapter. We will see that the log transformation makes it much easier to discover relationships between variables.

[MINOR] A variation of histograms: Density plots. Before moving on to the next type of graphical displays for numeric variables, let's also mention *density plots*, which are smoothed and scaled versions of histograms displaying "density" rather than counts on the vertical axis. Although there is hardly any mention of density plots in the PA modules and they function in more or less the same way as a histogram, they made an appearance in some old PA exams (e.g., the December 7, 2020 and December 8, 2020 exams), so it is beneficial to get some exposure to these plots.

In CHUNK 7, we use the `geom_density()` function to make the density plots for the log of claim amount and the square root of claim amount (see Figure 2.1.4).

```
# CHUNK 7
p1 <- ggplot(persinj, aes(x = log(amt))) +
  geom_density()
p2 <- ggplot(persinj, aes(x = sqrt(amt))) +
  geom_density()
grid.arrange(p1, p2, ncol = 2)
```

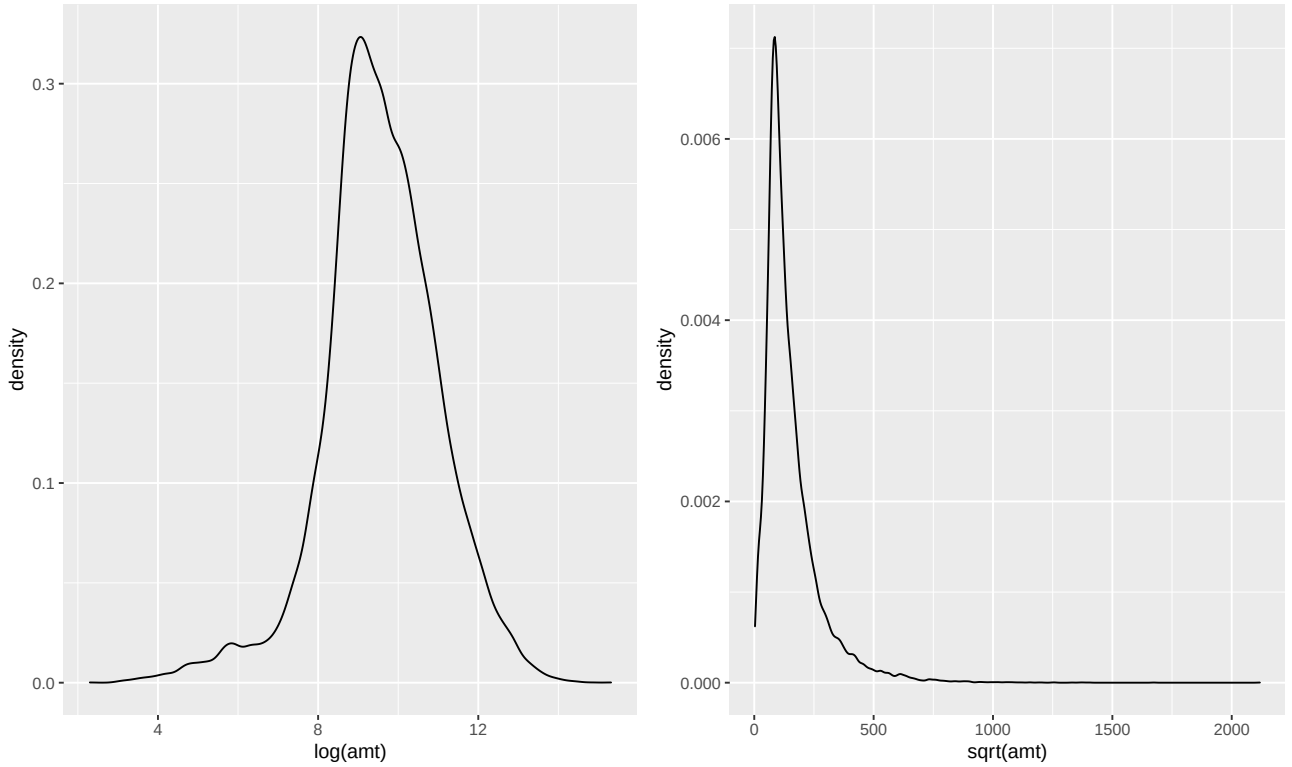


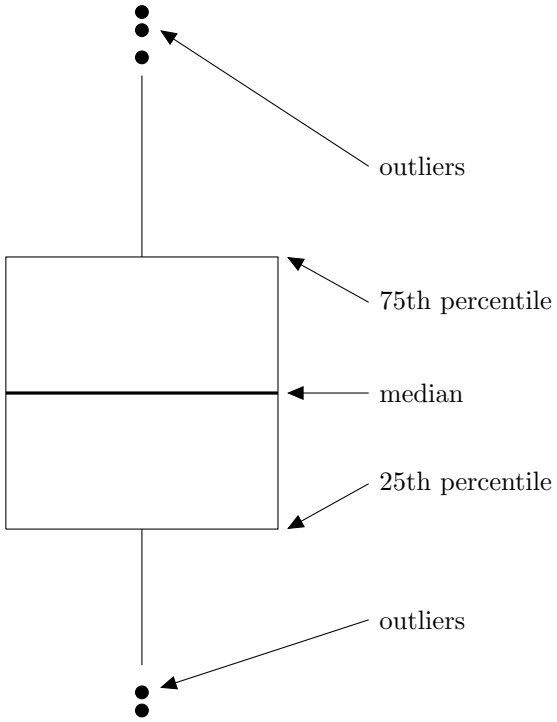
Figure 2.1.4: Density plots of the log of claim amount (left) and square root of claim amount (right) in the `persinj` dataset.

Comparing Figures 2.1.3 and 2.1.4, we can see that the density plots have the same shape as the corresponding histograms and can be interpreted in the same way. Do note that the vertical axis of the density plots shows "density" rather than "count" and the area under a density curve is always 1, just as a probability density function does.

Boxplots. An alternative to a histogram is a ***boxplot***, a.k.a. a *box plot* and a box-and-whiskers plot, which visualizes the distribution of a numeric variable by placing its 25% quantile, the median, and the 75% quantile in a "box," (hence the name of the plot) with the rest of the data points constituting the "**whiskers**." The amount of spacing between different parts of a boxplot reflects the degree of dispersion and skewness of the variable's distribution. "**Outliers**," defined here as data points that are above or below 1.5 times the inter-quartile range from either edge of the box, are shown as large dotted points.

Q2.6

Here is a prototypical boxplot.



Although boxplots do not directly show the actual shape of the variable's distribution (there is some loss of information relative to histograms), they offer a useful graphical summary of the key numeric statistics and allow for a visual comparison of the distributions of different numeric variables (in particular, the relative magnitude of their median and dispersion) or the distribution of the same numeric variable across different levels of another categorical variable. We will see how this works in Section 2.2.

Exercise 2.1.2. (What can we get from a boxplot?) Determine whether each of the following distributional quantities can be obtained (at least approximately) from a boxplot.

(a) Mean (b) Median (c) Mode

(d) Inter-quartile range (e) Standard deviation

Solution. Among the five quantities above, only the median (b) and inter-quartile range (d) can be read off a boxplot. □

In `ggplot2`, boxplots are constructed by the `geom_boxplot()` function, which takes the `y` aesthetic representing the numeric variable of interest (the `x` aesthetic is optional, but can be added to achieve splitting, as we will see in the next section). Use CHUNK 8 to draw a boxplot for each of claim amount and the log of claim amount (Figure 2.1.5).

```
# CHUNK 8
p1 <- ggplot(persinj, aes(y = amt)) +
  geom_boxplot()
p2 <- ggplot(persinj, aes(y = log(amt))) +
  geom_boxplot()
grid.arrange(p1, p2, ncol = 2)
```

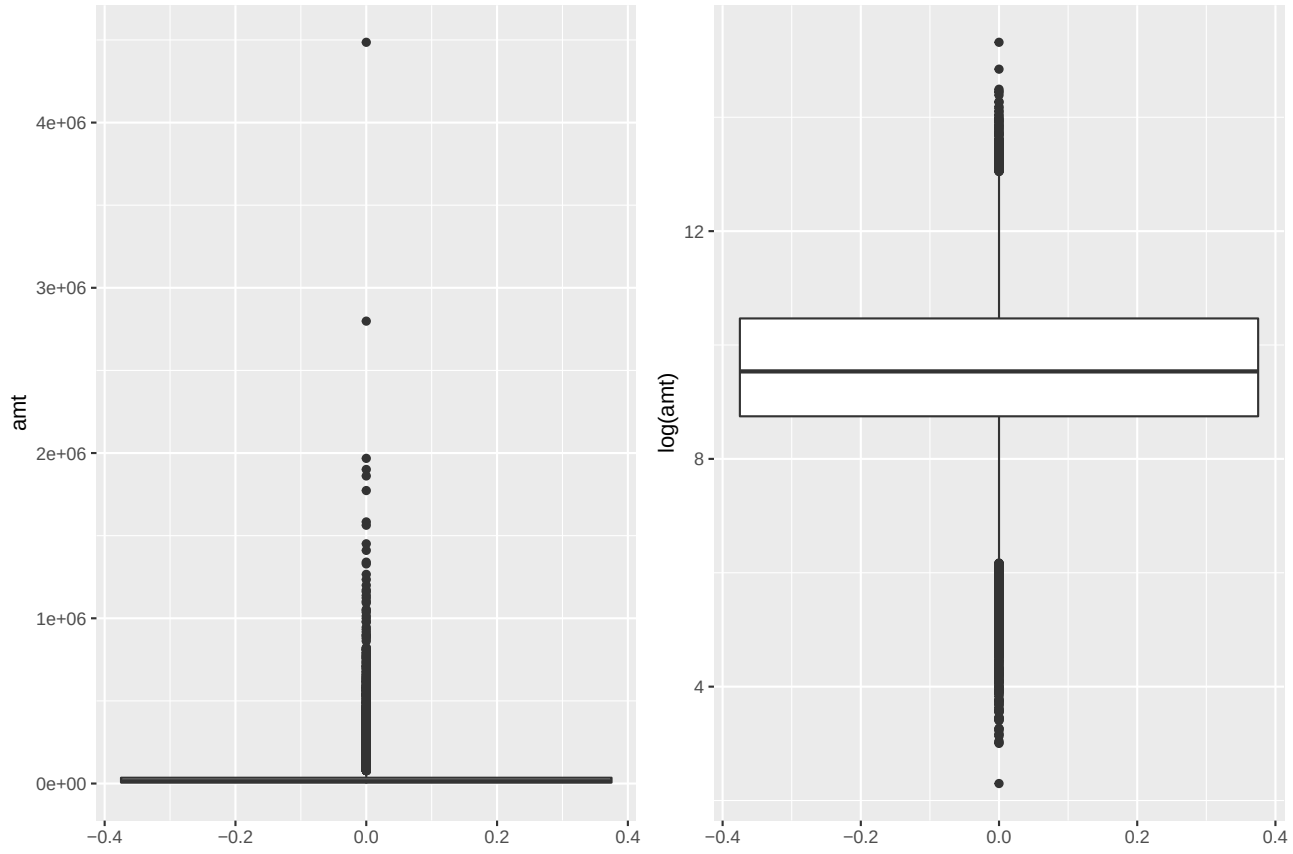


Figure 2.1.5: Boxplots of claim amount (left) and the log of claim amount (right) in the `persinj` dataset.

While most of the raw claim amounts are so close that the 25% percentile, median, and 75% percentile all degenerate to essentially the same line, the log transformation corrects for the skewness and re-positions the data points for much easier visual inspection. Still there are quite a number of "outliers," which are shown as large dotted points.

2.1.2 Categorical Variables

Let's move on to statistical (tabular) and graphical representations for categorical variables.

Case 1: Given raw data (the easier and more common case in Exam PA)

Suppose that we have access to data identifying the levels of a categorical variable for a set of individual observations. Here is the general form of the data:

Observation	Category Level
1	A
2	C
3	B
⋮	⋮

Descriptive statistics: Counts. Recall from Section 1.1 that **categorical variables**, even when coded as numbers, do not always have a natural numeric order, so statistical summaries like the mean and median may not make sense. To understand the distribution of a categorical variable, we can look at the number (or percentage) of observations in each of its levels through a *frequency table*, constructed by the `table()` function in R.

The `persinj` dataset has two categorical variables, injury code (`inj`) and legal representation (`legrep`). Recall from Table 2.1 that `inj` has seven levels while `legrep` is binary. In CHUNK 9, we make two frequency tables for `inj`, one showing the raw counts and one showing the percentage counts of the seven levels of `inj`.

```
# CHUNK 9
table(persinj$inj)

##
##     1     2     3     4     5     6     9
## 15638  3376  1133   189   188   256  1256

table(persinj$inj)/nrow(persinj)

##
##           1           2           3           4           5           6
## 0.709656925 0.153203848 0.051415865 0.008576874 0.008531494 0.011617353
##           9
## 0.056997640
```

We can see that the predominant group of injuries is those of injury code 1, followed by codes 2, 9, and 3. The other three groups have minimal observations in comparison.

Graphical displays: Bar charts. When the number of levels of a categorical variable increases, a frequency table becomes more and more difficult to read. In most cases, the frequencies *per*

se are not that important; what truly matters is their relative magnitude. In this regard, ***bar charts*** turn the numeric counts in a frequency table into bars whose heights (or lengths) are proportional to the number of observations in each level of the variable. Looking at a bar chart, we can easily tell which levels are the most popular and which ones have minimal observations on a relative basis. Q2.7

A bar chart can be created in `ggplot2` by the `geom_bar()` function, which takes the `x` aesthetic representing the categorical variable of interest. Let's run CHUNK 10 to produce two bar charts for injury code corresponding to the two frequency tables in CHUNK 9 (see Figure 2.1.6).[3]

```
# CHUNK 10
# first convert inj and legrep to factors (original data type is integer)
persinj$inj <- as.factor(persinj$inj)
persinj$legrep <- as.factor(persinj$legrep)

p1 <- ggplot(persinj, aes(x = inj)) +
  geom_bar(fill = "blue")
p2 <- ggplot(persinj, aes(x = inj)) +
  geom_bar(fill = "blue", aes(y = ..prop.., group = 1))
grid.arrange(p1, p2, ncol = 2)
```

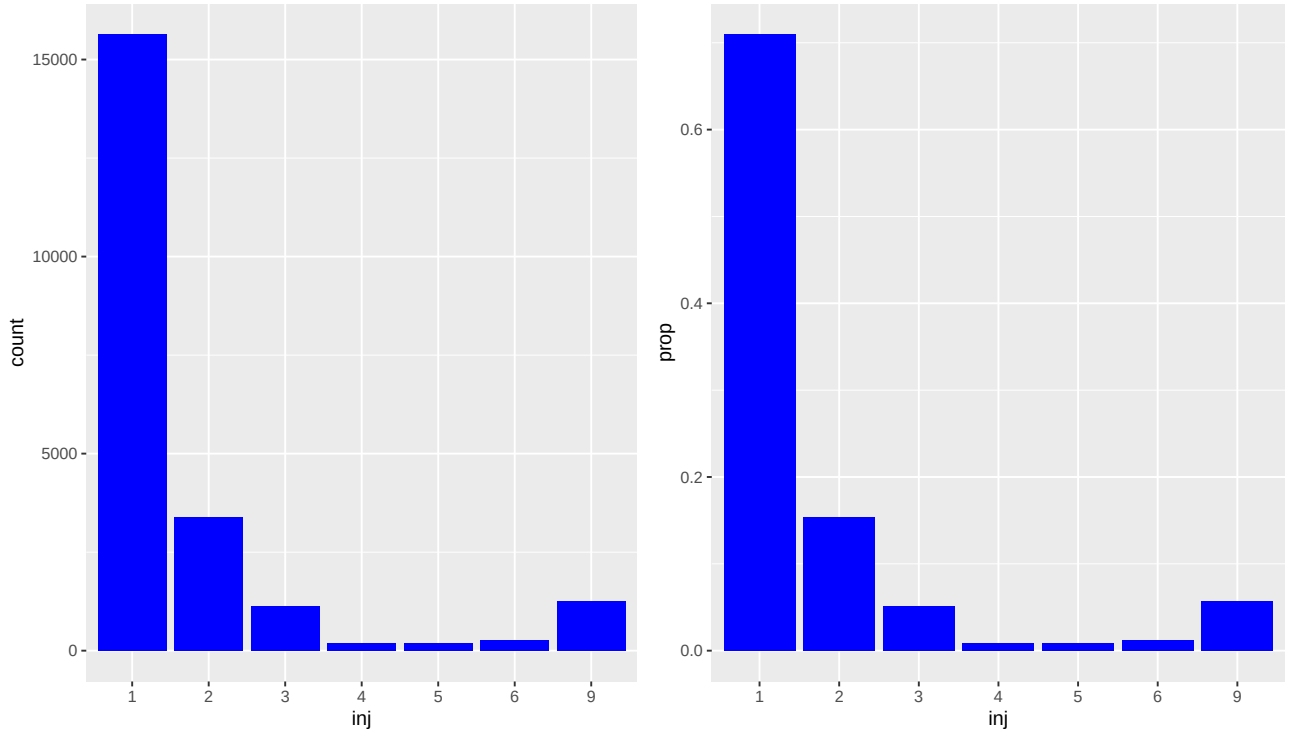


Figure 2.1.6: Bar charts of injury code in the `persinj` dataset.

[3]The command `y = ..prop.., group = 1` in the bar chart in CHUNK 10, as its name indicates, computes proportions rather than raw counts and relies on the so-called stat function associated with a geom function. This concept is rather involved, but is of limited use in Exam PA. If you are interested in what it is, please read pages 80 and 81 of the *Data Visualization* book.

Case 2: Given summarized data

In real applications, it is not uncommon that the data we have has been grouped, or summarized, to show the number of observations in each level of a categorical variable. Here is the general data structure:

Category Level	Number of Observations
A	n_{A}
B	n_{B}
C	n_{C}
⋮	⋮

CHUNK 11 below produces a version of `persinj` dataset, called `persinj_by_inj`, that shows the number of observations for each level of `inj`. The code involves the `tidyverse` package, which will be covered in Exam ATPA. Instead of worrying about the somewhat convoluted code syntax, try to pay attention to the output, which is far more important in the new exam format.

```
# CHUNK 11
# Uncomment the next line the first time you use the tidyverse package
# install.packages("tidyverse")

library(tidyverse)
persinj_by_inj <- persinj %>%
  group_by(inj) %>%      # grouped by inj
  summarize(count = n()) # count the no. of observations for each level of inj
persinj_by_inj
```

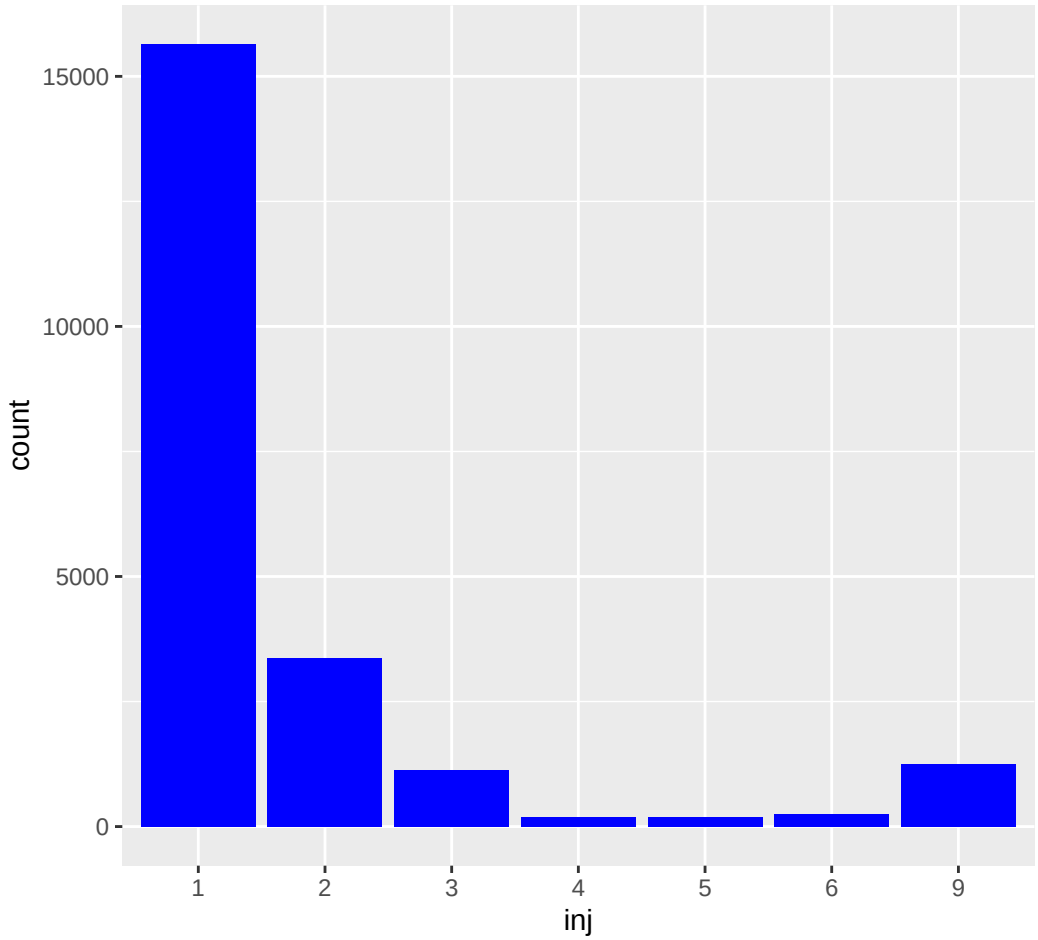
```
## # A tibble: 7 x 2
##   inj   count
##   <fct> <int>
## 1 1     15638
## 2 2      3376
## 3 3      1133
## 4 4       189
## 5 5       188
## 6 6       256
## 7 9      1256
```

The output shows, for example, that there are 15,638 observations with an injury code of 1, consistent with the frequency table in CHUNK 9.

If all we have is the `persinj_by_inj` dataset (we no longer have access to the original `persinj` dataset), how can we display the counts for each injury code? The `geom_bar()` function will not work well. If we map the `inj` variable to the `x` aesthetic of the `geom_bar()` function, then the function will keep track of how many times each distinct value of `inj` has occurred. It will

(faithfully!) treat `inj` as a variable with 7 distinct values, 1, 2, 3, 4, 5, 6, and 9, each of which appears once and only once, which is definitely not what we want. Instead, the `geom_col()` function will suit our purpose. When `inj` is mapped to the `x` aesthetic and `count` to the `y` aesthetic, the function will display the counts for each value of `inj`, as CHUNK 12 shows.

```
# CHUNK 12
ggplot(persinj_by_inj, aes(x = inj, y = count)) +
  geom_col(fill = "blue")
```



This bar chart is identical to the one in the left panel of Figure 2.1.6 based on the original version of the `persinj` dataset.

In summary:

> The `geom_bar()` function is for visualizing the distribution of a categorical variable given individual (raw) data, while the `geom_col()` function can serve the same purpose, given grouped (summarized) data.

2.2 Bivariate Data Exploration

Data exploration becomes even more intriguing and challenging when two or more variables are analyzed together rather than in isolation. This has the important advantage of revealing relationships, patterns, and outliers which become apparent only when variables are considered in combination with one another. This section therefore focuses on *bivariate data exploration*, where pairs of variables are investigated numerically and/or graphically to identify potentially interesting relationships. Note that:

- Due to the sheer number of possible pairs of variables in a real dataset, it won't always be feasible to examine every bivariate relationship as part of data exploration, e.g., if there are $p = 10$ (a reasonably small number) variables, then there are already a total of $\binom{p}{2} = 45$ pairs to examine. In practice, we tend to focus on relationships between the target variable and each predictor, and other potentially interesting relationships between the predictors that you hypothesize will be important in a given business problem.
- As you will see, bivariate exploration tools, if used appropriately, also lend themselves to multivariate relationships, i.e., relationships among three or more variables.

There are three types of bivariate combinations, depending on the type of variables under consideration.

2.2.1 Combination 1: Numeric vs. Numeric

Descriptive statistics: Correlations. An easy way to summarize the strength of the *linear* relationship between two numeric variables is through the ***correlation coefficient***, or simply *correlation*,[4] which is a unit-free metric on a scale from -1 to $+1$, as we learned from Exam P. The fact that the correlation is bounded makes it an especially interpretable and comparable metric.

Case 1. If the correlation is $+1$, then the two variables are perfectly *positively* correlated, and all pairs of values of the two variables lie exactly on an upward sloping straight line.

Case 2. If the correlation is 0, then the two variables are *uncorrelated*. (A straight line fitted to the pairs of values of the two variables has a zero slope.)

Case 3. If the correlation is -1, then the two variables are perfectly *negatively* correlated, and all pairs of values of the two variables lie exactly on a downward sloping straight line.

Although these extreme correlation values almost never arise in real datasets, they provide useful benchmarks for judging the size of a typical correlation. The larger the correlation in magnitude (i.e., the closer they are to $+1$ or -1), the stronger the degree of linear association between the two variables.

[4]To be precise, in Exam PA we are looking at *sample* correlations computed from data rather than population correlations computed from bivariate probability distributions and studied in Exam P. It is unthinkable that you had to do double integration in Exam PA!

In CHUNK 13, we use the `cor()` function in R to compute the correlation between the claim amount and operational time, and between the log-transformed claim amount and operational time in the `persinj` data.

```
# CHUNK 13
cor(persinj$amt, persinj$op_time)

## [1] 0.3466114

cor(log(persinj$amt), persinj$op_time)

## [1] 0.6070667
```

The two variables are moderately positively correlated on the original scale and the correlation becomes noticeably stronger when the claim amount is on the log scale. As much as correlation is a compact summary of the extent to which two numeric variables move in tandem, it can only capture ***linear*** relationships. A zero correlation only means that two variables are not linearly related, but they may be related in more subtle ways. As the following exercise shows, correlations may fail to reflect more complex, non-linear relationships (e.g., quadratic), which can be revealed more effectively by graphical displays.

Exercise 2.2.1. (Motivated from an old Exam P sample question: What can you say about two variables with a zero correlation?) Consider the following dataset with two variables:

```
# CHUNK 14
X <- seq(-10, 10)  # consecutive integers from -10 to 10
Y <- X^2           # the square of X
```

Determine whether X and Y are:

(a) Correlated or uncorrelated (b) Dependent or independent

Solution. In the rest of CHUNK 14, we compute the (sample) correlation between X and Y:

```
# CHUNK 14 (Cont.)
cor(X, Y)

## [1] 0
```

The zero correlation suggests that X and Y are uncorrelated, or *linearly* unrelated. However, the two variables are perfectly dependent via the quadratic relationship $Y = X^2$. **(Answer: (C))** □

Remark. This toy example illustrates the pitfall of using the correlation to detect useful predictors. If Y is the target variable and X is a potential predictor, then X may seem to be a predictor with limited predictive power based on its zero correlation with Y. To appreciate

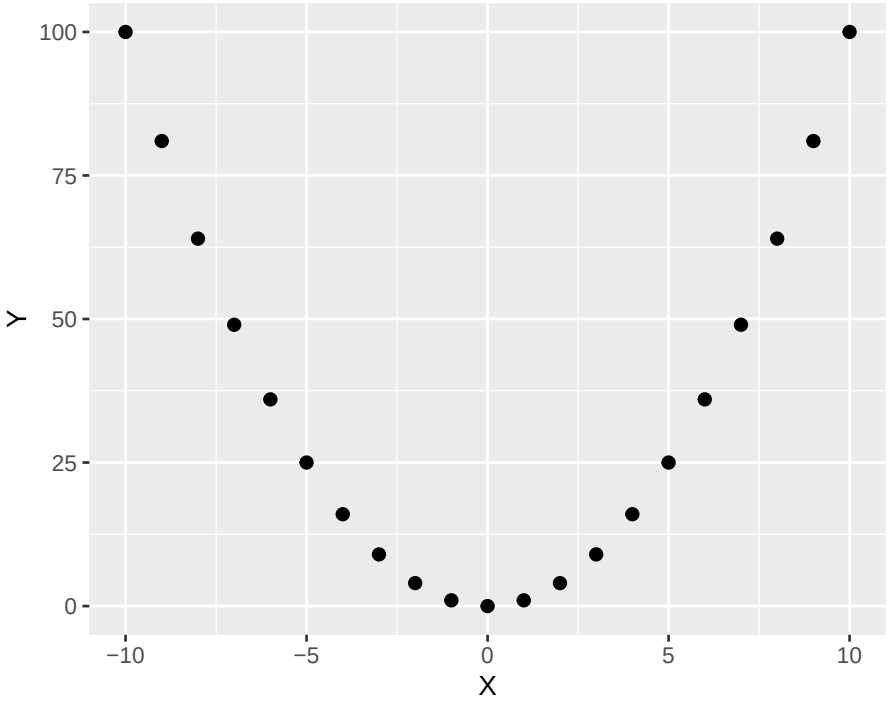
the relationship between the two variables fully, it is important to look at data in visual form using a graphical tool like a scatterplot, which we discuss next, instead of relying solely on summary statistics.

Graphical displays: Scatterplots. The relationship between two numeric variables (one of which is usually the target variable) is typically visualized by a ***scatterplot*** (a.k.a. a scatter plot), which plots each ordered pair of values of the two variables on a two-dimensional plane. If we plot variable Y against variable X, that means values of Y are mapped to the vertical axis and those of X to the horizontal axis. Such a plot visualizes the relationship between the two numeric variables across a wide range of their values. Besides linear relationships, it can also reflect non-linear, more complex ones (e.g., polynomial, periodic) and yield insights that correlations alone cannot provide.

Q2.8

Here is the scatterplot produced by the `geom_point()` function in the `ggplot2` package for the two variables in CHUNK 14 above: (If you are interested, see Subsection R.5.1 to learn more about how to customize a scatterplot produced by the `ggplot2` package.)

```
# CHUNK 14 (Cont.)
Df <- data.frame(X = X, Y = Y)
ggplot(Df, aes(x = X, y = Y)) +
  geom_point(size = 2)
```



The scatterplot lays bare the perfect square relationship between the two variables that is completely undetected by the correlation. This explains why scatterplots are one of the most commonly used graphical displays in data exploration in practice.

In CHUNK 15, we make two scatterplots, one for the untransformed claim amount and one for the log of claim amount, both against operational time, in the `persinj` dataset; see Figure 2.2.1. Due to the large number of overlapping observations, we have set the `alpha` argument to a very small value to incorporate a large amount of transparency (see Subsection R.5.1 for details).

Both plots exhibit an increasing relationship, but the scatterplot for the log of claim amount displays a much more conspicuous upward sloping trend, indicating that the log of claim amount is approximately positively linear in operational time. This is a further manifestation of the merits of the log transformation introduced in Subsection 2.1.1 in uncovering relationships that would otherwise go unnoticed.

```
# CHUNK 15
p1 <- ggplot(persinj, aes(x = op_time, y = amt)) +
  geom_point(alpha = 0.05) +
  geom_smooth(method = "lm", se = FALSE)
p2 <- ggplot(persinj, aes(x = op_time, y = log(amt))) +
  geom_point(alpha = 0.05) +
  geom_smooth(method = "lm", se = FALSE)
grid.arrange(p1, p2, ncol = 2)
```

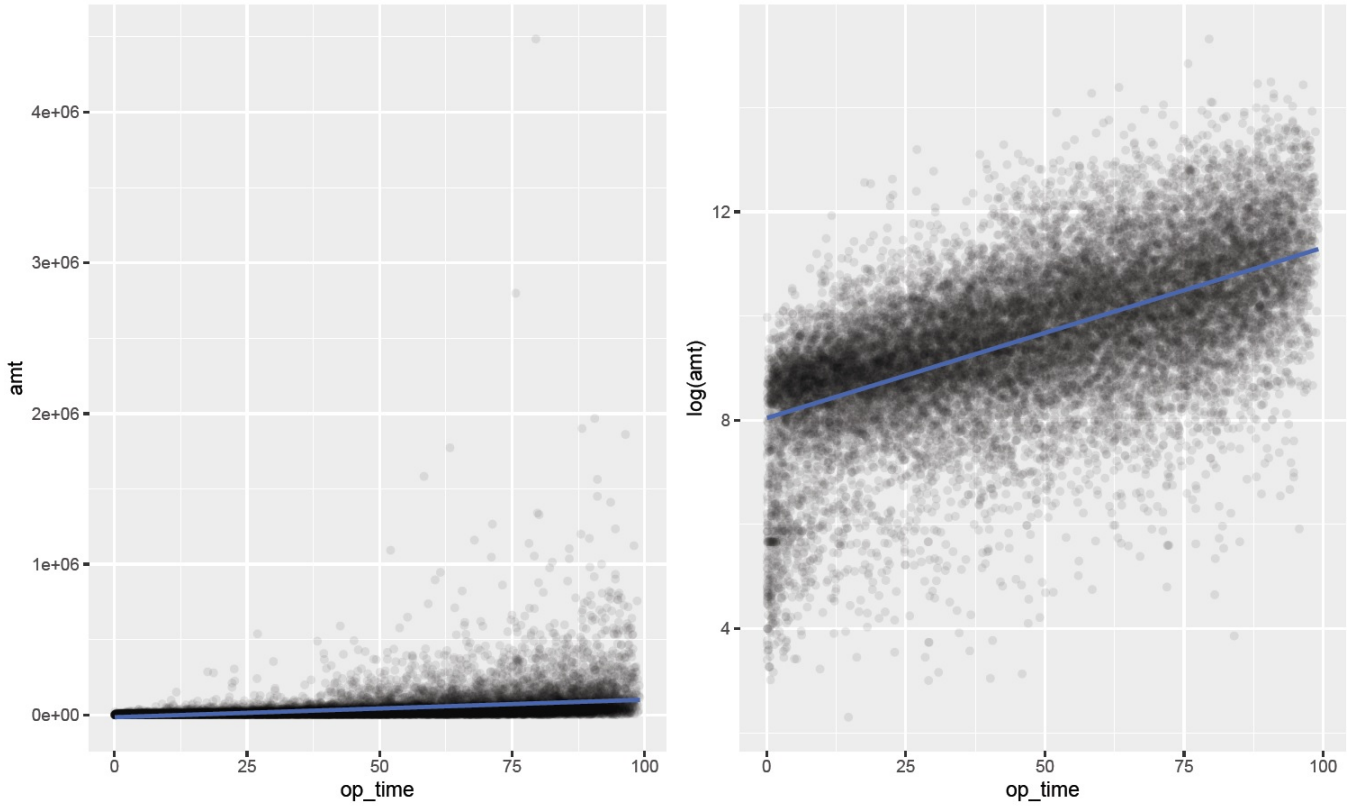


Figure 2.2.1: Scatterplots of claim amount (left) and the log of claim amount (right) against operational time in the `persinj` dataset.

Using scatterplots to explore three-way relationships. Although a scatterplot itself is confined to depicting the relationship between only two numeric variables, the effect of a third, categorical variable can be incorporated and investigated by decorating the observations by color, shape, or other visual elements according to the levels assumed by this third variable. This way, we can visually inspect whether the relationship between the two numeric variables varies with the levels of the third, categorical variable. In statistical language, this phenomenon is known as ***interaction***, which we will study in Section 3.1, and is an important modeling issue to keep in mind when constructing an effective predictive model.

Now run CHUNK 16 to make a scatterplot for the log of claim amount against operational time, with the observations colored according to legal representation (note that the `color` aesthetic is mapped to `legrep`); see Figure 2.2.2.

```
# CHUNK 16
ggplot(persinj, aes(x = op_time, y = log(amt), color = legrep)) +
  geom_point(alpha = 0.25) +
  geom_smooth(method = "lm", se = FALSE)
```

Figure 2.2.2: Scatterplot of the log of claim amount against operational time colored by legal representation in the `persinj` dataset.

The scatterplot shows that the two smoothed lines corresponding to the two levels of `legrep` have markedly different slopes and intercepts (keep in mind that we are on the log scale, so a small change in the intercept and slope can matter a lot on the original scale). In other words, the linear relationship between the log of claim amount and operational time depends materially on whether legal representation is present or not. We can roughly tell the effect of legal representation on the (log of) claim amount:

> Injuries with legal representation (i.e., those with `legrep = 1`) tend to produce higher claim amounts, unless operational time is extraordinarily large (90 or higher).

In Section 4.2, we will formally assess the extent of interaction and construct a generalized linear model that properly takes the interaction effect into account.

2.2.2 Combination 2: Numeric vs. Categorical

To understand the interplay between a numeric variable and a categorical variable, it is best to investigate the distribution of the former indexed (or "split") by each possible level of the latter. In effect, we are looking at the conditional distribution of the numeric variable given different levels of the categorical variable.

Descriptive statistics: Conditional means. To summarize the association between a numeric variable and a categorical variable, we can partition the data into different subsets, one subset for each level of the categorical variable, and compute the mean (or median) of the numeric variable there. These conditional means varying substantially suggest a strong relationship between the two variables.

In CHUNK 17 below, we produce a table of the mean and median of the log-transformed claim amount (it is also fine to use the untransformed claim amount variable) split by different levels of each of `inj` and `legrep`, which are categorical.

```
# CHUNK 17
library(tidyverse)
persinj %>%
  group_by(inj) %>%
  summarize(
    mean = mean(log(amt)),
    median = median(log(amt)),
    n = n()
  )
```

```
## # A tibble: 7 x 4
##   inj    mean median     n
##   <fct> <dbl>  <dbl> <int>
## 1 1      9.37   9.36 15638
## 2 2     10.3   10.3   3376
## 3 3     10.7   10.9   1133
## 4 4     11.0   11.2    189
```

```
## 5 5     10.8   10.6    188
## 6 6      9.68   9.07   256
## 7 9      8.35   8.57  1256
```

```
persinj %>%
  group_by(legrep) %>%
  summarize(
    mean = mean(log(amt)),
    median = median(log(amt)),
    n = n()
  )
```

```
## # A tibble: 2 x 4
##   legrep  mean median     n
##   <fct>  <dbl>  <dbl> <int>
## 1 0       9.18   9.32  8008
## 2 1       9.77   9.64 14028
```

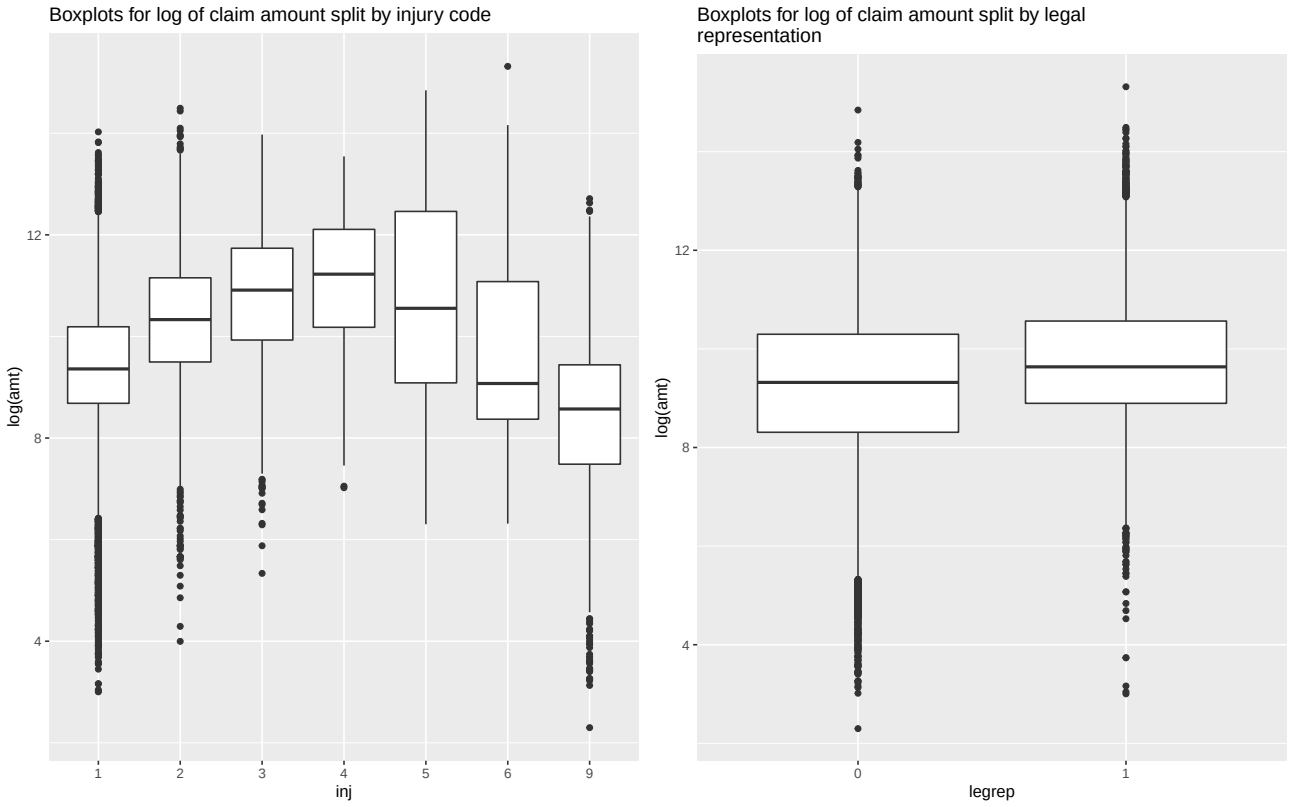
It is clear that the claim amount on average increases from injury code 1 to injury code 4, then decreases down to injury code 9. The two means split by `legrep` are in agreement with what we observed in Figure 2.2.2, which shows that larger claim amounts are associated with the use of legal representation.

Graphical displays: Split boxplots. The conditional distribution of a numeric variable given a second, categorical variable can be visualized by a ***split boxplot***, which consists of a series of boxplots of the numeric variable, one corresponding to each level of the categorical variable. If the level and/or size of the boxes vary remarkably across the levels of the categorical variable, then that is a pointer to a strong association between the two variables.

Q2.9

Run CHUNK 18 to construct two split boxplots for the log of claim amount, one split by injury code and one split by legal representation. As in univariate exploration, you will enter the numeric variable in the `y` aesthetic, but the categorical variable that is used to split the numeric variable will play a role in the `x` aesthetic. This will generate a collection of boxplots of the numeric variable for each level of the categorical variable. The two boxplots further support the findings based on the summary statistics above, but turn them more powerfully into diagrams.

```
# CHUNK 18
p1 <- ggplot(persinj, aes(x = inj, y = log(amt))) +
  geom_boxplot() +
  labs(title = "Boxplots for log of claim amount split by injury code")
p2 <- ggplot(persinj, aes(x = legrep, y = log(amt))) +
  geom_boxplot() +
  labs(title = "Boxplots for log of claim amount split by legal
representation")
grid.arrange(p1, p2, ncol = 2)
```



Exercise 2.2.2. **(Motivated from page 116 of PA Module 2: Using boxplots to visualize a three-way relationship)** CHUNK 19, shown overleaf, produces a series of boxplots for the log of claim amount split by injury code (the `x` aesthetic), followed by legal representation (the `fill` aesthetic) within each injury code.

Describe three conclusions concerning the relationships among two or more of the variables revealed by the boxplots. (Try to come up with your own points in your head before peeking at the solution!)

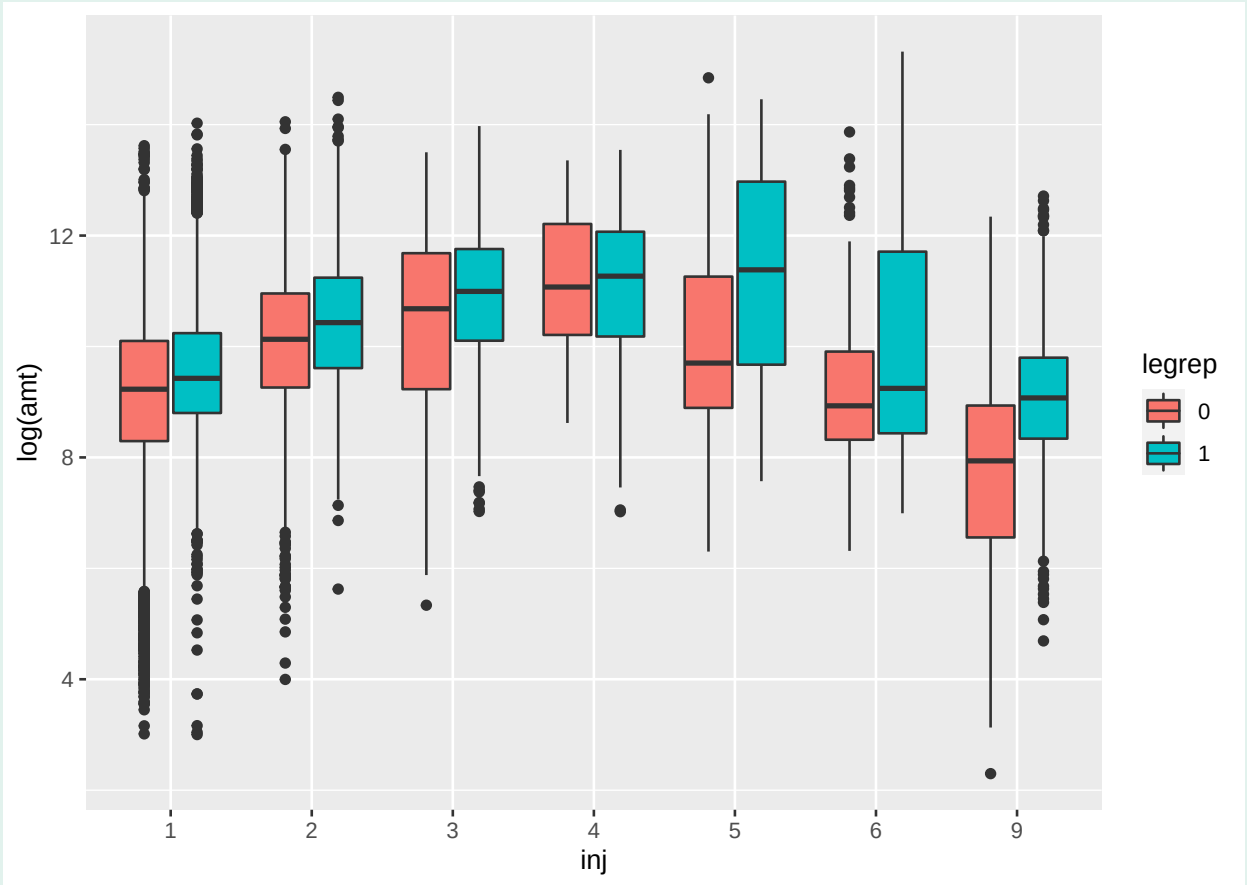
Solution. Here are three conclusions you can draw: (There are other points you can propose.)

- Based on the levels of the centers of the boxplots (which represent the medians of the

`log(amt)`) across the injury codes, the claim amount tends to increase from injury codes 1 to 4, then decrease thereafter.

- Because the center of the pinkish red boxplot tends to be lower than that of the greenish blue boxplot within each injury code, we can see that larger claim sizes tend to come from injuries with legal representation.
- The positive effect of legal representation is the most prominent for injuries of codes 5 and 9. In other words, there may be an interaction between `inj` and `legrep` in affecting `log(amt)`. □

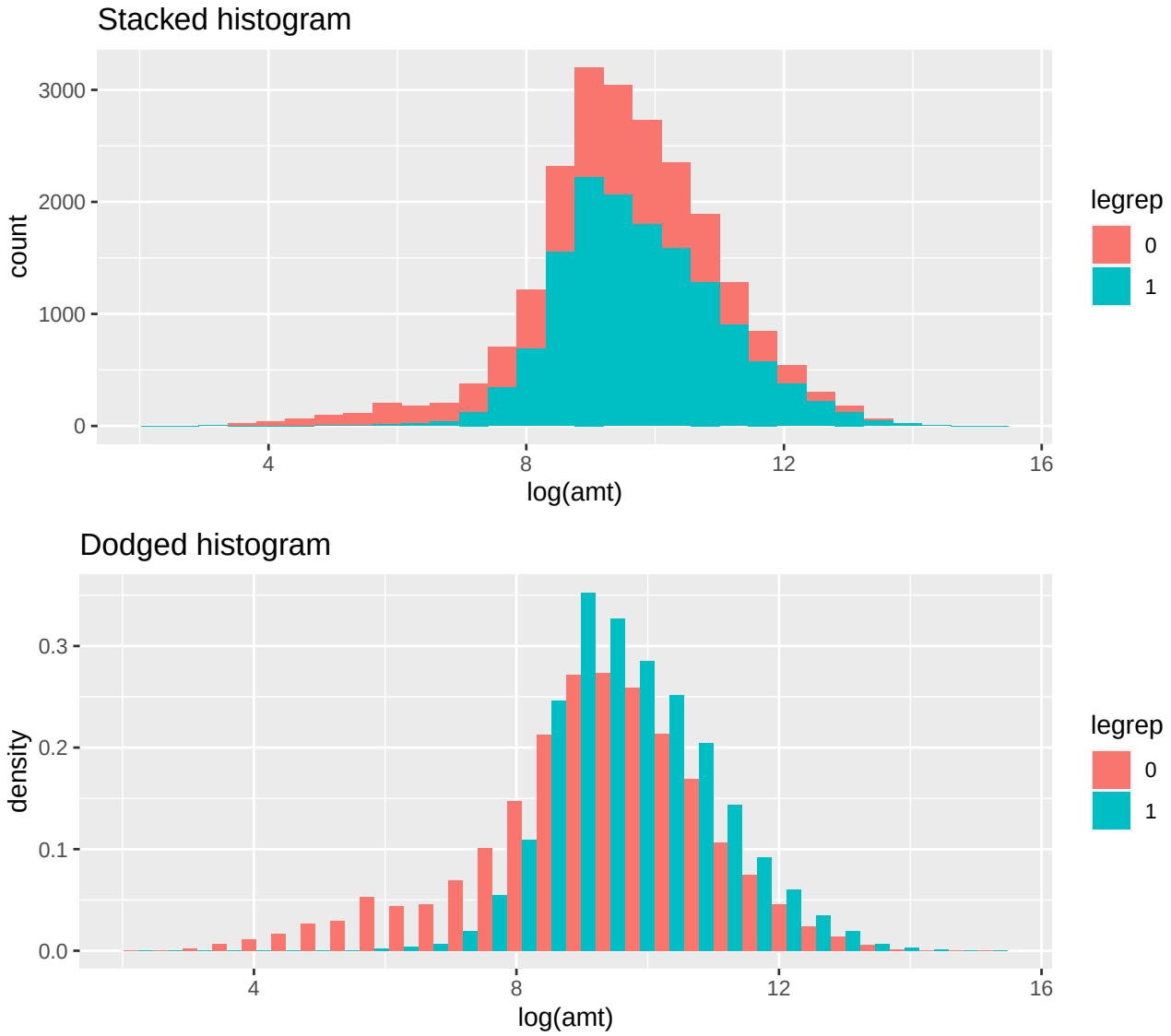
```
# CHUNK 19
ggplot(persinj, aes(x = inj, y = log(amt), fill = legrep)) +
  geom_boxplot()
```



Alternative graphical displays: Stacked and dodged histograms. Although not as effective as a split boxplot (in my opinion!), a modified histogram can also be adapted to visualize the distribution of a numeric variable split by a categorical variable. They can either be histograms **stacked** on top of one another (using the `fill` aesthetic) to highlight the contribution of each categorical level to the overall distribution of the numeric variable, or **dodged** histograms, with each bin placed side by side for comparison (note the argument `position = "dodge"`).

Run CHUNK 20 to produce both types of histograms for the log of claim amount split by legal representation. For the dodged histogram, it is necessary to specify `y = ..density..` so that the histogram shows the density rather than raw counts; raw counts are misleading because there are a lot more injuries with legal representation than those without. As expected, both histograms suggest that larger claims (e.g., those with `log(amt)` greater than 9) tend to be those with legal representation.

```
# CHUNK 20
p1 <- ggplot(persinj, aes(x = log(amt), fill = legrep)) +
  geom_histogram() +
  labs(title = "Stacked histogram")
p2 <- ggplot(persinj, aes(x = log(amt), y = ..density.., fill = legrep)) +
  geom_histogram(position = "dodge") +
  labs(title = "Dodged histogram")
grid.arrange(p1, p2)
```



2.2.3 Combination 3: Categorical vs. Categorical

Descriptive statistics: Joint counts. When examining a pair of categorical variables, it is often useful to construct a two-way frequency table showing the counts, or the number of observations, for *every combination* of the levels of the two variables, again using the `table()` function. When two arguments are supplied to the `table()` function, the first argument will correspond to the rows of the two-way frequency table while the second argument will correspond to its columns.

As an example, run CHUNK 21 to make a two-way frequency table for legal representation crossed with injury code.

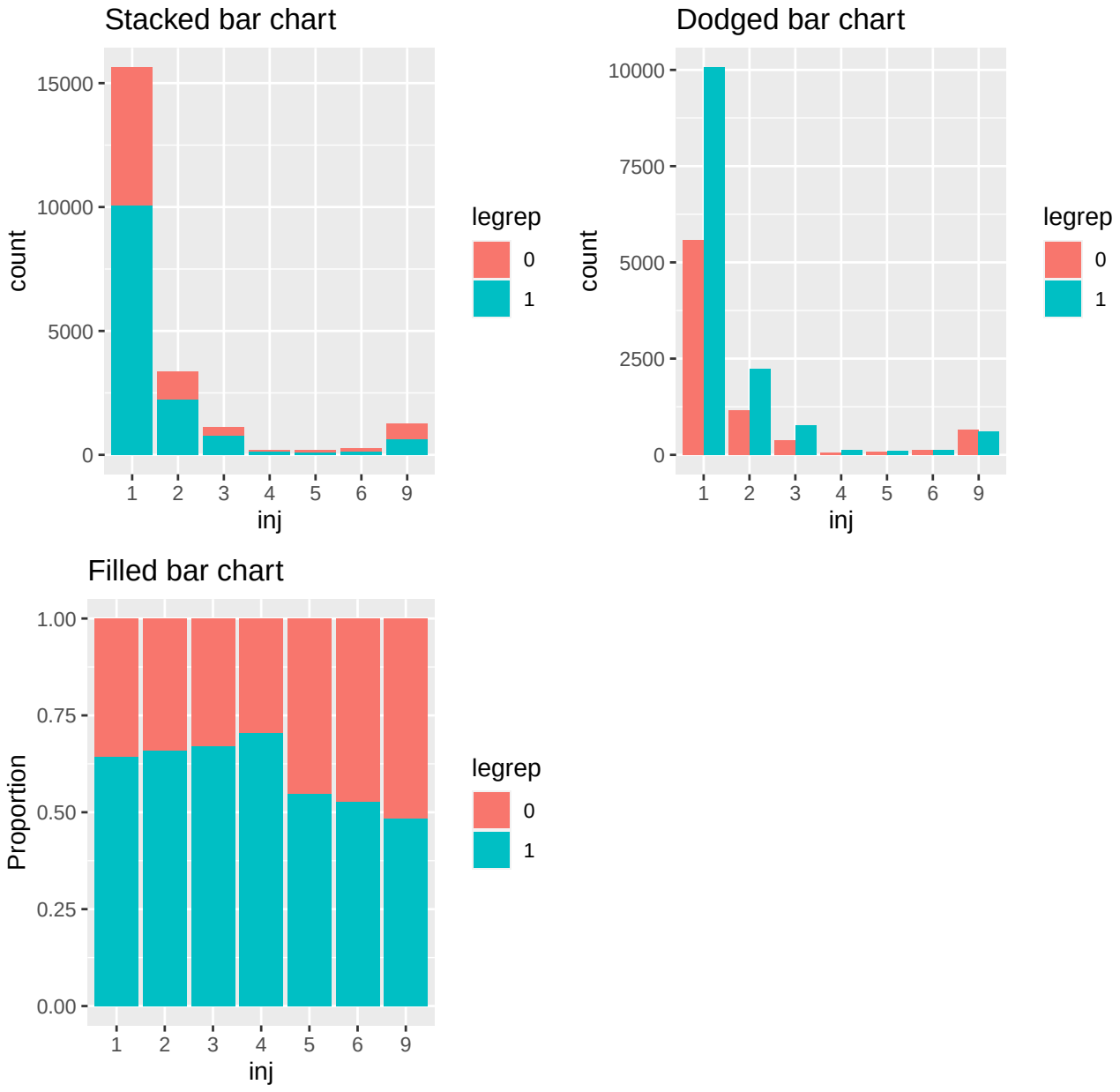
```
# CHUNK 21
table(persinj$legrep, persinj$inj)
```

```
##
##         1     2     3     4     5     6     9
##   0  5571  1152   374    56    85   121   649
##   1 10067  2224   759   133   103   135   607
```

Graphical displays: Bar charts. The counts or proportions in a frequency table are useful statistics, but a visual display often expresses these statistics much more powerfully and makes for easier comparison interpretation, especially when the two categorical variables have a lot of levels. To visualize the distribution of a categorical variable split by another categorical variable effectively, *split bar charts* can be of use. These charts come in different versions, and they are illustrated in CHUNK 22 for injury code split by legal representation:

- *Stacked:* The stacked bar chart has counts within each injury code colored by legal representation; notice that the `fill` aesthetic is set to `legrep`. In other words, each bar is broken down proportionally into injuries with legal representation (colored in greenish blue) and those without legal representation (colored in pinkish red), and the two bars are "stacked" on top of each other.
- *Dodged:* Showing the same information as the stacked bar chart, the dodged bar chart has counts within each injury code separated according to legal representation and placed side by side for visual comparison; notice the option `position = "dodge"` in the `geom_bar()` function. This is very much like how a dodged histogram works.
- *Filled:* The filled bar chart displays the relative proportions (not the raw counts) of injuries with and without legal representation within each injury code due to the option `position = "fill"` (not to be confused with the `fill` aesthetic). In the language of probability, it shows the conditional distribution of the binary legal representation variable given the injury code, and is, in effect, a rescaled version of the stacked bar chart, with the bars within each injury code adjusted so that their total height is 1.

```
# CHUNK 22
p1 <- ggplot(persinj, aes(x = inj, fill = legrep)) +
  geom_bar() +
  labs(title = "Stacked bar chart")
p2 <- ggplot(persinj, aes(x = inj, fill = legrep)) +
  geom_bar(position = "dodge") +
  labs(title = "Dodged bar chart")
p3 <- ggplot(persinj, aes(x = inj, fill = legrep)) +
  geom_bar(position = "fill") +
  labs(title = "Filled bar chart", y = "Proportion")
grid.arrange(p1, p2, p3, ncol = 2)
```



In the face of these three bar charts, an inevitable question is:

Which one is the best and in what way?

The truth is that the three charts convey subtly different information and each have their own specialties. There is no unequivocal answer as to which chart is superior, and it all depends on your focus and needs.

- *Proportions ACROSS injury codes*

  Although all of these charts serve to depict the association between two categorical variables, filled bar charts visualize this most effectively via rescaling and is most suitable when the categorical variable in the `fill` aesthetic is the target variable (and the variable in the `x` aesthetic is a predictor). A cursory glance 👁 at the filled bar chart in CHUNK 22 shows how the proportion of injuries with legal representation varies *across* the injury code—higher for codes 1 to 4 than for codes 5, 6, and 9. Such information, strictly speaking, is also revealed in the stacked and dodged charts, but can be easily obscured by the number of observations in some of the injury codes. Given the stacked or dodged chart in CHUNK 22, one might need to zoom in 🔍 on, for example, codes 5 and 6, to determine which code has a higher proportion of injuries with legal representation, but the filled chart makes this rather apparent. Q2.10

  On the downside, the filled bar chart does not show the number of injuries in each code, but the stacked and dodged charts do. As we discussed in Section 1.4, factor levels with more observations are generally considered more reliable in predictive modeling and more likely to produce statistically significant results than sparse levels, but we wouldn't be able to tell from the filled bar chart alone.

  Perhaps a good strategy in practice is to pair filled bar charts up with stacked or dodged chart, using the former to see the variation of the level proportions easily and using the latter to identify populous and sparse levels.

- *Proportions WITHIN each injury code*

Q2.11

If you are interested in comparing the proportions of injuries with and without legal representation *within* (not across) each injury code, then the dodged bar chart will suit your needs better than the stacked and filled charts. By aligning the two proportions within each injury code and putting them side by side on a *common baseline* (both bars start from 0), the dodged bar chart makes it much easier to determine visually which proportion is higher based on the bar heights. Shown below is a simple graphical illustration in a general case.

(a) Two bars sharing the same common base. Clearly, B is taller than A.

(b) The same two bars in (a), except that they are unaligned. Now it is much less clear whether B is taller than A.

Let's return to CHUNK 22. Looking at the stacked or filled chart, you may have difficulty judging which proportion within injury code 9 is higher, but the dodged chart shows more clearly that the greenish blue bar is marginally higher than the pinkish red bar.

When the categorical variable that enters the `fill` aesthetic has three or more levels, a dodged bar chart will ease the comparison of the proportions of this variable within each level of the primary variable (i.e., the one in the `x` aesthetic) even more effectively.

Epilogue: How to make a graph clearer and more effective? The comparison between the stacked and dodged bar charts above suggests, from a broader perspective, that there are better and worse ways of visually representing even the same data. Some research indicates that when people perceive 👁 graphs, or visual representations of data, they tend to compare values within the graph, judging them mostly on a *relative* basis (rather than an absolute basis). Here are some general tips to make this visual comparison easy and, more generally, make a graph communicate the intended information more effectively:

- Use appropriate visual elements to display numeric values. According to the *Data Visualization* book, human perception of relative values differs widely by the visual elements that are adopted. Here are some of these elements ranked from more to less effective:

$$\boxed{\begin{array}{c}\text{heights (lengths)}\\ \text{on a } \underline{\text{common}} \text{ scale}\end{array} > \begin{array}{c}\text{heights (lengths)}\\ \text{on } \underline{\text{unaligned}} \text{ scales}\end{array} > \begin{array}{c}\text{angles or areas}\\ \text{(e.g., in a pie chart)}\end{array}}\ .$$

  The first inequality is precisely the reason why a dodged bar chart (where the bars are put on a common scale) may be preferable to a stacked or filled bar chart (where the bars are unaligned), and the second inequality explains why some data scientists frown upon ☹ seeing *pie charts*.[5] 🥧 A pie chart encodes numeric quantities by angles or areas, which are usually more prone to misjudgment and unnecessarily complicated compared to lengths. Q2.12

- Choose a color palette that reflects the structure of your data. This means:
  - ▷ For an unordered (nominal) categorical variable such as "sex" or "color," use distinct, contrasting colors that are easy to distinguish.
  - ▷ For a numeric variable or an ordered (ordinal) categorical variable such as "level of education," use a graded or sequential color scale that reflects the magntude or order of the variable.
- For numeric quantities, clearly labeled axes are essential for communicating the type of scaling (e.g., the original scale vs. the log scale) as well as the range of values (does the axis start from 0, or is it truncated?).
- It will also help to add an informative title that precisely describes what the graph shows and makes it stand on its own, at least largely.

  (Note that a title is not always necessary, and the plots in the rest of this manual and in the PA modules don't always have one, as long as the conveyed information is clear.)

[5] *(For mathematically inclined students)* Pie charts can be viewed as bar charts in polar coordinates.

Conceptual Review Questions for Chapter 2

Q2.1. Explain two typical uses (or applications) of exploratory data analysis.

(Answer: See page 71.)

Q2.2. Explain the pros and cons of using descriptive statistics and graphical displays for exploring data.

(Answer: See page 72.)

Q2.3. Explain how a histogram can be used to visualize the distribution of a numeric variable.

(Answer: See page 76.)

Q2.4. Explain the problems with using a right-skewed variable in predictive modeling.

(Answer: See page 81.)

Q2.5. Suggest two functional transformations that can be used to handle a right-skewed variable.

(Answer: See page 82.)

Q2.6. Explain how a boxplot can be used to visualize the distribution of a numeric variable.

(Answer: See page 85.)

Q2.7. Explain how a bar chart can be used to visualize the distribution of a categorical variable.

(Answer: See page 88.)

Q2.8. Explain how a scatterplot can be used to visualize the relationship between two numeric variables.

(Answer: See page 93.)

Q2.9. Explain how a split boxplot can be used to visualize the relationship between a numeric variable and a categorical variable.

(Answer: See page 97.)

Q2.10. Identify a bivariate visualization that is suitable for representing the variation of the level proportions of a categorical variable across the levels of another categorical variable.

(Answer: See page 103.)

Q2.11. Identify a bivariate visualization that is suitable for representing the level proportions of a categorical variable within each level of another categorical variable.

(Answer: See page 104.)

Q2.12. Explain why dodged bar charts often do a better job of conveying relative counts than stacked bar charts and pie charts.

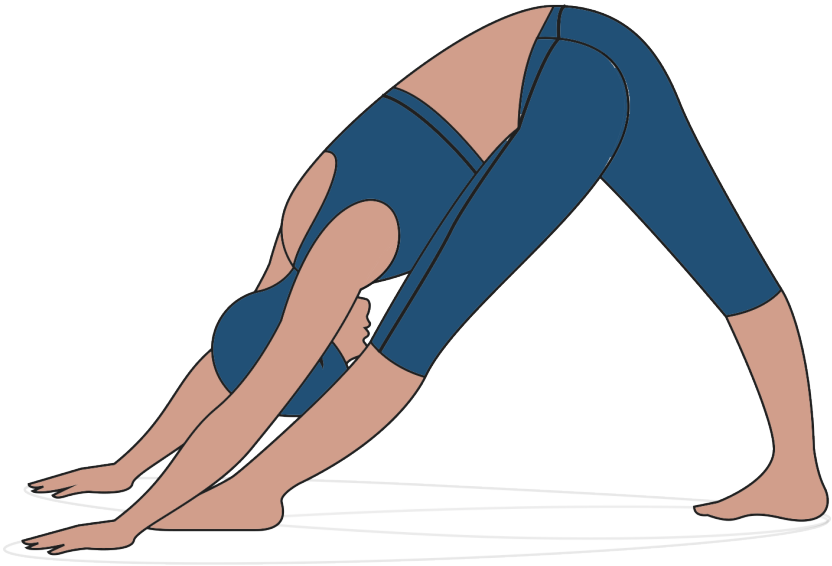
(Answer: See page 105.)

STUDY BREAK

Do you have any fun ideas for future Study Breaks?

Share your inspiration with **#actexstudybreak**

Exercise is a proven way to boost your verbal memory, thinking, and learning. [1]



Pyramid with Arms Extended

Yoga - Sun Salutations

*Hold each pose for three long breaths, breathing in through your nose and out through your mouth.

Note that we are mathematicians, not yogi: attempt poses at your own risk!

[1]: https://www.health.harvard.edu/blog/regular-exercise-changes-brain-improve-memory-thinking-skills-201404097110

ACTEX Learning

Learn Today. Lead Tomorrow.

Part II

Theory of and Case Studies in Predictive Analytics

Chapter 8

Practice Exams

Introduction

Having been well *trained* on the core of this study manual (Chapters 1 to 6) and past PA exams (Chapter 7), you need to be exposed to unseen *tests* to avoid overfitting and identify areas in which you need more *training*. To this end, please make good use of the two substantially updated practice exams in this chapter. Each exam comes with the following resources:

(1) A project statement describing a business problem, a data dictionary, and a series of tasks you have to complete

According to the PA exam syllabus,

> "A hardcopy of the problem statements will *not* be available at Prometric testing centers. The statements will be available for the entirety of the exam on-screen."

This is rather unfortunate because when I took the exam in December 2019, having a printed statement to look at helped quite a lot.

(2) *(Available for download on Actuarial University as a separate file)* A Microsoft Word document with spaces labeled as "**ANSWER:**" for you to write your responses to each specific subtask when you practice

On the real exam, this Word document and the project statement are the same file. In other words, you will enter your responses directly in the project statement, similar to FSA written-answer exams. This is the only file you will submit for grading.

(3) Detailed illustrative solutions with sample responses and related learning outcomes from the PA exam syllabus identified[1]

[1]Not all subtasks conveniently fit into the learning outcomes in the syllabus, but they still lie within the general scope of Exam PA.

⚠ NOTE ⚠

In addition to Practice Exams 1 and 2, we have introduced a graded mock exam product with completely different questions, which is available for separate purchase. Please refer to page xxix of the preface or check out https://www.actexlearning.com/exams/pa/exam-pa-mock-exam for more details.

What are these two practice exams like?

Designed taking the new exam format effective from April 2023 and the style of recent PA exams into account, these two practice exams give you a holistic review of the entire PA exam syllabus and have the following characteristics:

- They consist of 7 to 9 tasks,[2] with a total of **70 points**. Some tasks are longer and some shorter. Almost all tasks are further broken down into a few subtasks. Exam points are provided for each subtask, so you have some idea of how much you should write. As I mentioned in the preface of this manual, you should spend about 3 minutes per exam point.
- Following the exam format effective from the December 2021 sitting, different tasks are mutually independent and can be answered in any order (unless you have a special preference, you may simply start with Task 1). Even if you struggle in a certain task, you can proceed to the next task and start anew. You will not make data preparation or modeling decisions that affect the rest of the project. There are also no tasks about comparing the performance of models constructed in different tasks. (You may have to rank models and select the best model *within the same task*, however.)
- Like recent exams, there are a large number of conceptual or descriptive tasks testing your prior understanding of predictive analytic concepts (look for the verbs "Describe" and "Explain" in the question prompt). You can complete these tasks without looking at any R code or output, or referring to the business problem.
- *(New!)* Starting from the April 2023 sitting, R and RStudio will not be available on the exam, but as the PA exam syllabus says,

  > "all code and output relevant to the tasks will be provided as part of the exam materials."

  The two practice exams embrace this new format. In quite a few tasks (e.g., Tasks 1, 2, 3, 6, and 8 of Practice Exam 1), you are given some R output and asked to use the output to answer the questions. Sometimes R code is also provided (e.g., Task 5 of Practice Exam 1). You are expected to know what the code does to address some subtasks adequately. This is what the R-based case studies in the manual are for.

[2] Although recent exams seem to have more and more tasks (e.g., the October 2024 exam has 10 and the April 2024 exam has 12) of a silo nature, the two practice exams remain valuable and perfectly applicable resources as some tasks can be easily broken down into shorter tasks. The total points are still fixed at 70.

- They strike a good balance between easy items testing topics regularly featured in past exams and harder, more unfamiliar items. As comprehensive as this study manual is, each PA exam will likely have a small number of unfamiliar tasks designed to identify the candidates that thrive on new, unseen exam tasks and are not overfitted to past exams. The harder items in the practice exams are in a similar vein.

  (In fact, I am not surprised if members of the PA exam committee have access to this manual and deliberately test obscure things I did not discuss at length!)

How to use these practice exams?

To make the most of these practice exams, here are my recommendations:

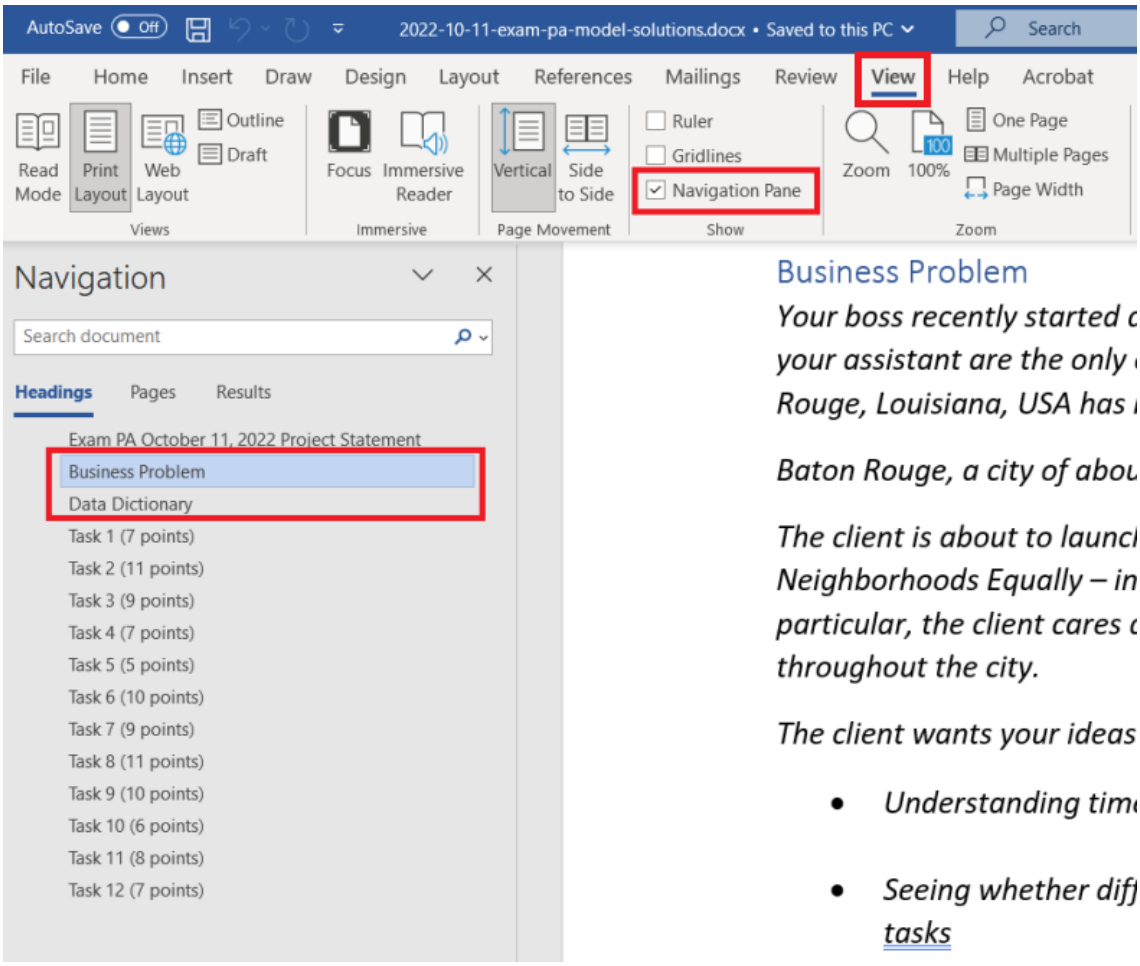
- Attempt them only when you have finished reading the study manual and studied recent PA exams (at least the October 2024, April 2024, October 2023, and April 2023 exams). Working on the practice exams when you are not fully ready defeats their purpose.
- Set aside exactly 3 hours 30 minutes and work on each exam in a simulated exam environment detached from distractions. Put away your manual, notes, and phone—no Facebook, Instagram, Twitter, or Snapchat. You can only have your calculator with you (you may also use Excel, which is available on the exam, to do calculations if you prefer). When you are finished, compare your responses with my suggested solutions and see how well you have done.
- Be sure to read the task statement and the Business Problem section carefully. Almost always, the Business Problem section has something useful for answering a few subtasks, and a seemingly minor point mentioned there can make a huge difference.
- Budget your time wisely. Don't spend a disproportionate amount of time on a single subtask, no matter how difficult it seems. As I mentioned in the preface, you should spend 3 minutes per exam point on average.

⚠ NOTE ⚠

Don't feel too frustrated if you find these two exams hard and long—they are probably (a bit) harder than the real exam! It is better to see something more difficult when you practice than to be defeated on the real exam, right? ☻

A Note on Navigation

During the exam, you may want to scroll back to the Business Problem and Data Dictionary at the beginning of the Microsoft Word file, then continue to work on different tasks. To navigate back and forth efficiently, press `Ctrl+F`, or click `View > Navigation Pane`.



This may save you some time and trouble on the exam, where every second counts!

ACTEX PA Manual Practice Exam 1 Project Statement

IMPORTANT NOTICE – THIS IS THE PROJECT STATEMENT OF THE FIRST PRACTICE EXAM. IF YOU ARE NOT READY FOR IT YET, LEAVE IMMEDIATELY AND RETURN LATER.

General Information for Candidates

This examination has 9 tasks numbered 1 through 9 with a total of 70 points. The points for each task are indicated at the beginning of the task, and the points for subtasks are shown with each subtask.

Each task pertains to the business problem described below. Additional information on the business problem may be included in specific tasks—where additional information is provided, including variations in the target variable, it applies only to that task and not to other tasks. For this exam there is no data file or .Rmd file provided. Neither R nor RStudio are available or required.

The responses to each specific subtask should be written after the subtask and the answer label, which is typically ANSWER, in the separate Word document.[3] Each subtask will be graded individually, so be sure any work that addresses a given subtask is done in the space provided for that subtask. Some subtasks have multiple labels for answers where multiple items are asked for—each answer label should have an answer after it.

Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. The answer should be confined to the question as set. No response to any task needs to be written as a formal report. Unless a subtask specifies otherwise, the audience for the responses is the examination grading team and technical language can be used.

Prior to uploading your Word file, it should be saved and renamed with your five-digit candidate number in the file name. If any part of your exam was answered in French, also include "French" in the file name. Please keep the exam date as part of the file name.

The Word file that contains your answers must be uploaded before the five-minute upload period time expires.

[3] As mentioned before, you will write your responses directly in the project statement on the real exam.

Business Problem

You work at ABC, a large actuarial consulting firm, and have been asked to assist School Wiz, a group dedicated to providing remedial education to troubled students. School Wiz has heard about you the legend, because you are among the very few who got Grade 10 in Exam PA, and wants to explore using your services to advance their business goals. They have collected preliminary data[4] *of past high school students. They would like to be able to identify which of the incoming high school students have a high tendency to fail, before they enter their high school year. These students will receive remedial services in time.*

School Wiz has determined that out of the three grade variables in the data, `G1`*,* `G2`*, and* `G3`*, they would like you to just focus on building predictive models based on* `G3`*. A student who receives a grade of 10 or more will pass. Your goal is to use the available data to construct two models that will predict if a student will pass (rather than the overall grade). One model should be GLM-based and one should be tree-based.*

School Wiz has provided the following data dictionary.

[4]This practice exam is based on the setting of the Student Success sample project (available from pages 8 and 9 of the June 2021 PA exam syllabus) and turns it into a much more useful task-based project consistent with the current exam format. The dataset for this sample project in turn is adapted from the Student Performance Data Set contributed by Paulo Cortez to the UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Data Dictionary

Name	Description	Variable Values
sex	Student's sex	Binary: F (female) or M (male)
age	Student's age	Integer from 15 to 22
Medu	Mother's education	Integer from 0 (none) to 4 (higher education)
Fedu	Father's education	Integer from 0 (none) to 4 (higher education)
Mjob	Mother's job	Factor: at_home, health (health care related), other, services (civil services, administrative or police), teacher
Fjob	Father's job	Same levels as Mjob
studytime	Weekly study time	Integer from 1 (very short) to 4 (very long)
failures	Number of past class failures	Integer from 0 to 3
schoolsup	Extra educational support	Binary: yes or no
famsup	Extra family supplement	Binary: yes or no
paid	Extra paid classes	Binary: yes or no
activities	Extra-curricular activities	Binary: yes or no
internet	Internet access at home	Binary: yes or no
romantic	Has a romantic relationship	Binary: yes or no
famrel	Quality of family relationships	Integer from 1 (very bad) to 5 (excellent)
freetime	Free time after school	Integer from 1 (very low) to 5 (very high)
goout	Going out with friends	Integer from 1 (very low) to 5 (very high)
Dalc	Weekday alcohol consumption	Integer from 1 (very low) to 5 (very high)
Walc	Weekend alcohol consumption	Integer from 1 (very low) to 5 (very high)
absences	Number of absences in high school year	Integer from 0 to 75
G1	First trimester grade in high school year	Integer from 0 to 20
G2	Second trimester grade in high school year	Integer from 0 to 20
G3	Third trimester grade in high school year	Integer from 0 to 20
pass	Pass indicator	0 if a student fails and 1 if a student passes

Task 1 (7 points)

The following is the correlation matrix for **G1**, **G2**, and **G3**:

```
          G1        G2        G3
G1 1.0000000 0.8821056 0.8301591
G2 0.8821056 1.0000000 0.9151279
G3 0.8301591 0.9151279 1.0000000
```

(a) (*2 points*) Describe one strength and one weakness of a correlation matrix as a bivariate data exploration tool.

ANSWER:

(b) (*2 points*) Based on the correlation matrix, explain why basing pass or fail entirely on **G3**, as requested by School Wiz, may be a sensible decision.

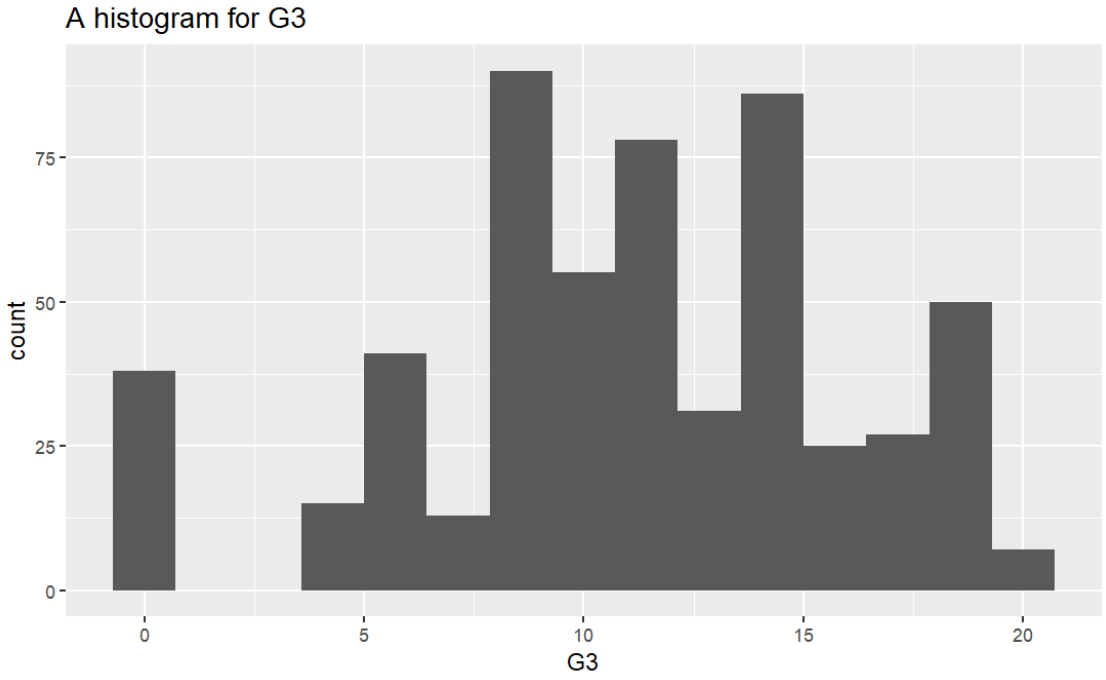
ANSWER:

(c) (*3 points*) Describe how principal components analysis can provide an alternative method for determining whether a student will pass or not based on all of **G1**, **G2**, and **G3**.

ANSWER:

Task 2 (5 points)

An alternative to modeling **pass** is to treat **G3** as the target variable and model it directly to determine pass or fail. To explore this alternative, your assistant has produced the following histogram for **G3**:



(a) (*2 points*) Describe the distributional characteristics of **G3**.

ANSWER:

(b) (*3 points*) Discuss one advantage and one disadvantage of modeling **G3** as the target variable over modeling **pass** for School Wiz from a GLM perspective.

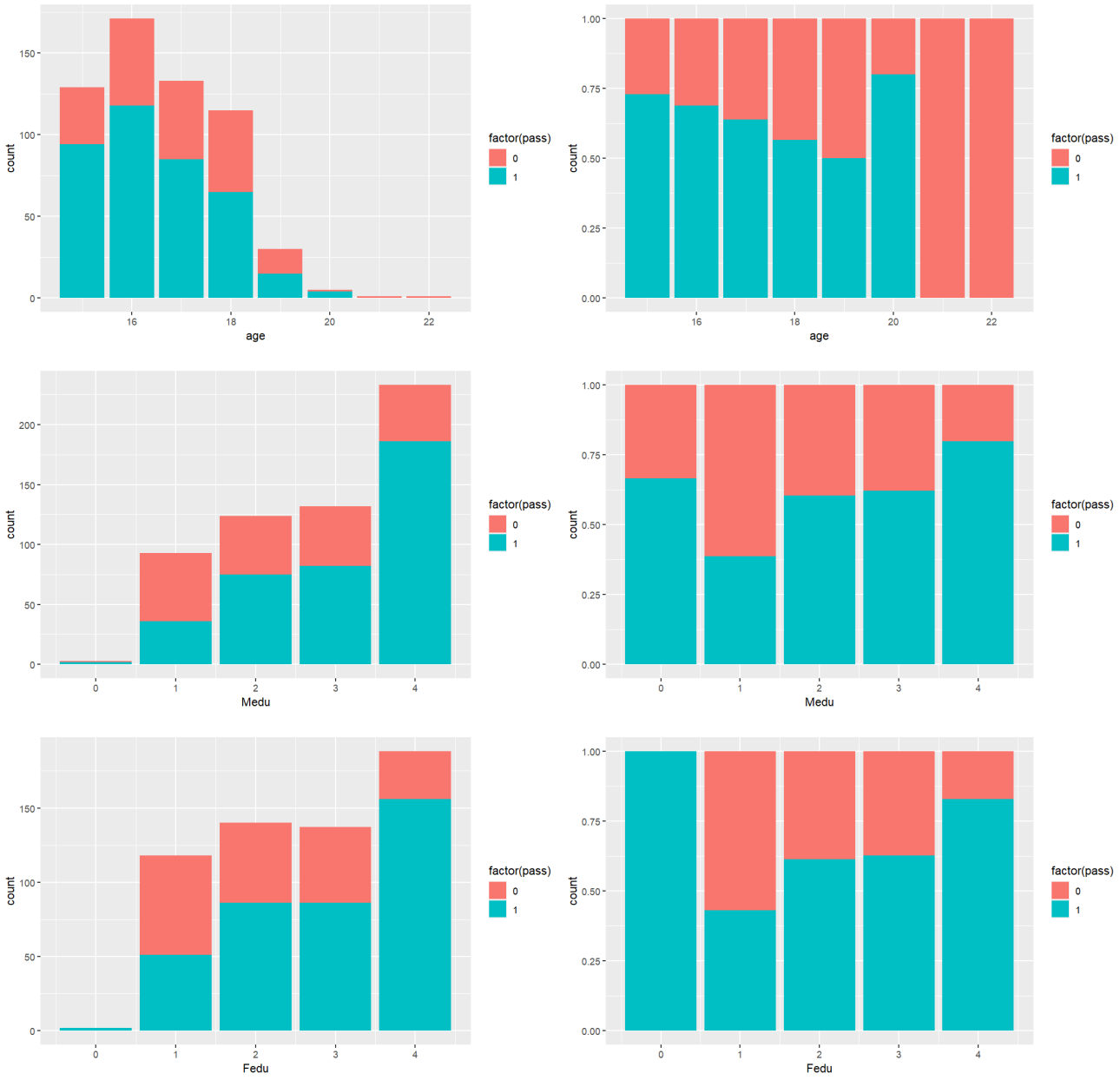
ANSWER:

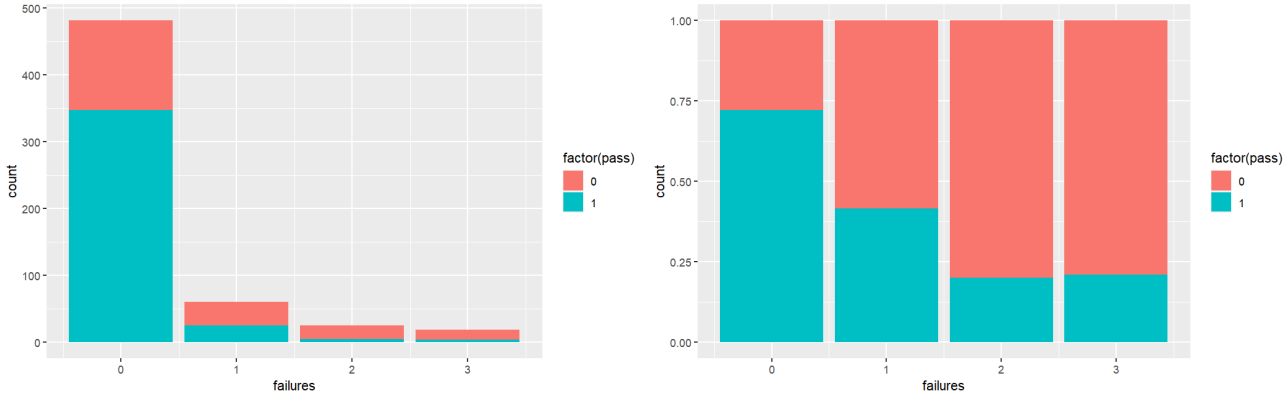
Task 3 (7 points)

You have asked your assistant to investigate the variables in the data.

(a) (*2 points*) Explain the problem with using **absences** for predicting **pass**.

Your assistant has conducted exploratory data analysis for **pass**. Here is part of the output:





(b) (*3 points*) Describe two anomalies of the data revealed by the output above.

ANSWER:

(c) (*2 points*) Identify and explain which variable above appears to be the most important predictor of **pass**.

ANSWER:

Task 4 (5 points)

Your assistant suggested creating a new variable that flags any previous class failures and including this flag variable in your models, in addition to variables that already exist in the data. The value of the variable is 1 if **failures** is higher than or equal to 1, and 0 if **failures** is 0. Your assistant thinks that this variable may be a useful feature for predicting **pass**.

(a) (*3 points*) Explain the modeling impacts, if any, of adding the new flag variable when running a GLM.

ANSWER:

(b) (*2 points*) Explain the modeling impacts, if any, of adding the new flag variable when running a decision tree.

ANSWER:

Task 5 (9 points)

Your assistant has provided the following R code to perform a certain cluster analysis.

```
data.hc <- data.all[, c("Medu", "Fedu")]

data.hc$Medu <- scale(data.hc$Medu)
data.hc$Fedu <- scale(data.hc$Fedu)

hc <- hclust(dist(data.hc))
```

(a) (*2 points*) Explain how cluster analysis can be used to develop features for a predictive model.

ANSWER:

(b) (*3 points*) Explain what kind of cluster analysis is performed by your assistant.

ANSWER:

(c) *(2 points)* Explain what the **scale()** function in your assistant's code does and why it is important.

ANSWER:

In retrospect, your assistant thinks that the code should have included a random seed so that the same output will be obtained every time the code is run. He apologizes for this omission.

(d) (*2 points*) Critique your assistant's statement.

ANSWER:

Task 6 (12 points)

Having read the *ACTEX Study Manual for Exam PA*, your assistant knows that accuracy, sensitivity, specificity, and AUC are commonly used performance metrics for a classifier.

(a) (*3 points*) Define accuracy, sensitivity, specificity, and AUC for a general classifier.

ANSWER:

Accuracy:

Sensitivity:

Specificity:

AUC:

(b) (*4 points*) Describe how accuracy, sensitivity, specificity, and AUC vary with the cutoff of a classifier.

ANSWER:

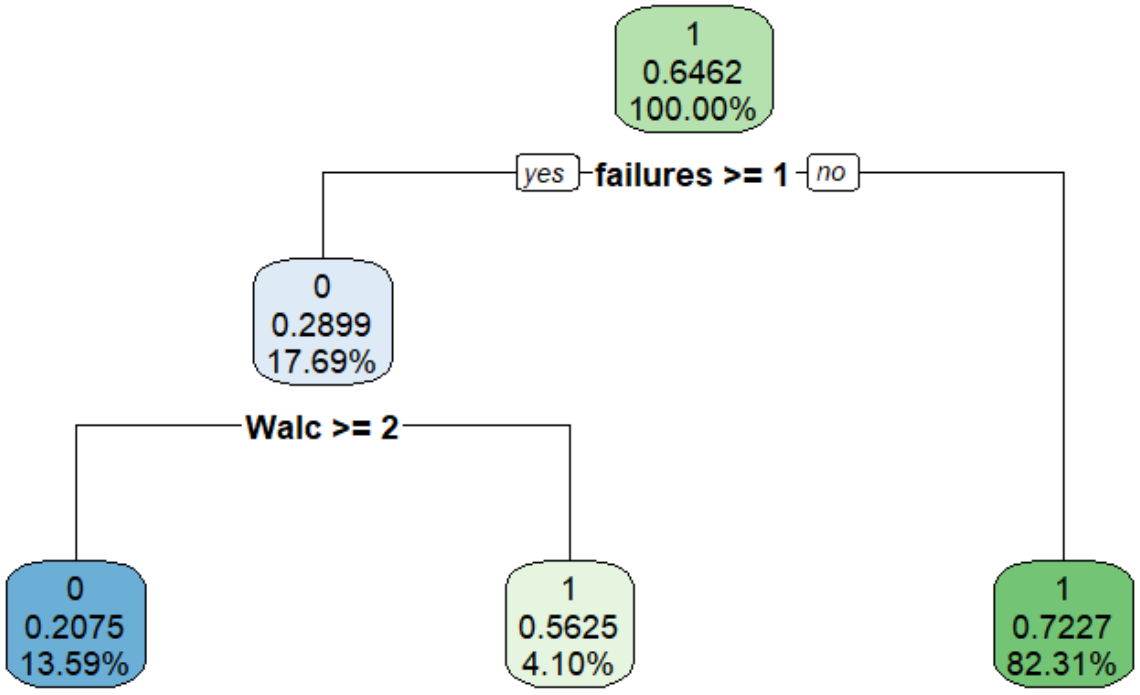
Accuracy:

Sensitivity:

Specificity:

AUC:

Your assistant has fitted a classification tree for **pass** on a subset of the data containing 390 observations, resulting in the following tree.



(c) (*5 points*) Fill in the following confusion matrix for the classification tree based on a cutoff of 0.5. Show your work.

	Reference	
Prediction	0	1
0	?	?
1	?	?

Task 7 (11 points)

Your assistant has provided code to split the data into the training (70%) and test sets (30%).

(a) (*2 points*) Describe the trade-off involved when selecting the percentages of data in the training and test sets.

ANSWER:

Your assistant has also set up code for fitting a regularized logistic regression model for **pass** on the training set.

(b) (*3 points*) Explain why **lambda** and **alpha** in an elastic net are hyperparameters and how these two parameters affect an elastic net.

ANSWER:

Why lambda and alpha are hyperparameters:

Lambda:

Alpha:

(c) (*2 points*) Describe the significance of using `alpha` equal to 1 in this business problem.

ANSWER:

The following shows the coefficient estimates of the variables selected in the resulting model:

```
                        s0
(Intercept)      0.49996348
Medu             0.23399344
Fedu             0.09891187
Mjobother       -0.14728207
failures        -0.68078016
famsupyes       -0.51504749
goout           -0.07769020
Walc            -0.02494307
```

(d) (*4 points*) Interpret the estimates of the intercept, and the coefficients for the categorical variable and the numeric variable with the most significant impact on **pass**.

ANSWER:

Intercept:

Coefficient for categorical variable:

Coefficient for numeric variable:

Task 8 (7 points)

Your assistant has run a boosted classification tree for **pass** on the training set.

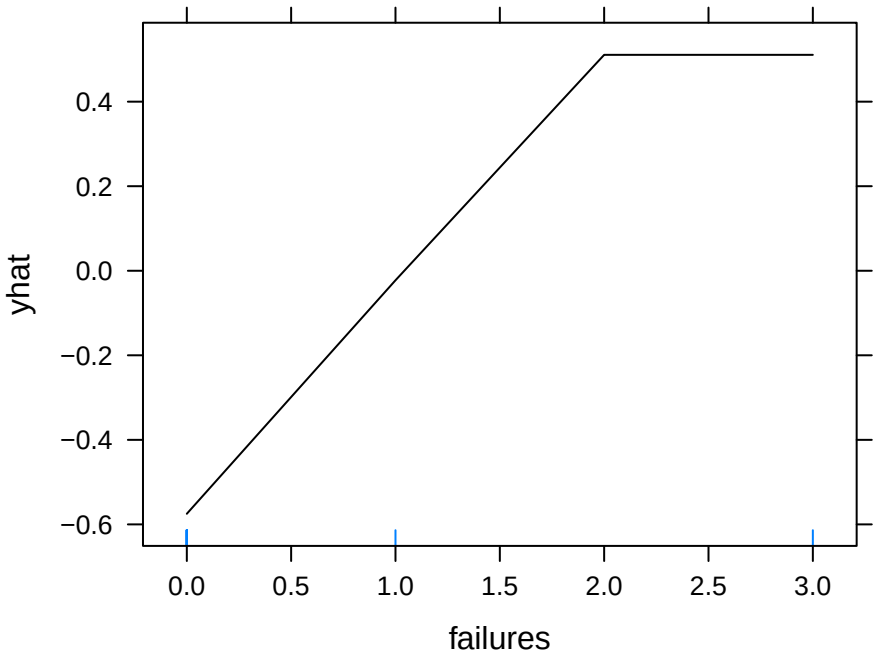
(a) (*3 points*) Explain the role played by the **eta** and **nrounds** parameters in a boosted tree, and the considerations for selecting these two parameters.

ANSWER:

eta:

nrounds:

The partial dependence plot for **failures** is provided below.



(b) (*2 points*) Provide an interpretation of the plot above.

ANSWER:

(c) (*2 points*) Describe the limitation of a partial dependence plot with respect to the interaction between variables.

ANSWER:

Task 9 (7 points)

To help School Wiz put the prediction performance of the models you have constructed or will construct in perspective, your assistant suggests fitting an intercept-only GLM for **pass** on the training set.

(a) (*2 points*) Explain how the intercept-only GLM can be used to assess the prediction performance of other models.

ANSWER:

(b) (*3 points*) Describe the characteristics of the prediction produced by this model and its ROC curve.

ANSWER:

(c) (*2 points*) Determine the test AUC of this model.

ANSWER:

****END OF PRACTICE EXAM 1****

Practice Exam 1 Suggested Solutions

⚠ NOTE ⚠

The following apply to both Practice Exams 1 and 2:

- Each practice exam has a fairly comprehensive coverage of the topics in the PA exam syllabus, ranging from the business problem, data exploration, data preparation, to modeling issues concerning GLMs and decision trees. Apart from conceptual and descriptive items, I made a deliberate attempt to include some subtasks that test your understanding of basic R code and hand calculation based on some R output. These subtasks may figure more prominently in the new exam format.
- The following "suggested" solutions are mainly for illustration purposes. Even though they are likely to be more detailed than what you can write in 3.5 hours, they are by no means perfect. Feel free to augment and refine my responses as you see fit. In many cases, there is a range of fully satisfactory approaches and there may be valid alternatives not discussed.
- Particularly important points in the solutions are underlined for easy identification. While these points (or phrases with similar meaning) can definitely enrich your responses, there is **no expectation that you cover all of them**. As the *Guide to SOA Exams* says,

  > *"...candidates do not always need to cover every possible aspect of the solution to receive full points..."*

- Some commentary and exam-taking strategies for specific subtasks are shown in *italics*. They are not part of the solutions.

Task 1 – Justify using G3 to determine pass or fail (7 points)

Ambrose's comments: This is an unseen, but not-so-demanding task specific to this business problem. It is about why using only one grade variable to determine pass or fail makes sense (though it may not be the optimal decision) with reference to the strong correlations among the three grade variables. A closely related topic is PCA.

Relevant PA exam learning outcomes:

2b) Identify the types of variables and terminology used in predictive modeling.

2f) Apply bivariate data exploration techniques.

3b) Apply principal components analysis to transform data.

(a) (*2 points*) Describe one strength and one weakness of a correlation matrix as a bivariate data exploration tool.

ANSWER:

Strength. A correlation matrix provides a convenient way to summarize the strength of the linear relationship between numeric variables, one pair at a time, by a set of metrics (the correlations), ranging from -1 to $+1$. Entries of the matrix that are close to $+1$ or -1 indicate strongly linearly related variables.

Weakness. Any of the following:

- A correlation matrix only captures linear relationships. Two numeric variables that have a nearly zero correlation can be related in many other regular ways (e.g., quadratic).
- A correlation matrix only captures the (linear) relationship between two numeric variables at a time. It may fail to represent relationships that exist among a group of numeric variables.
- A correlation matrix only works for numeric variables, not categorical (factor) variables.

(b) (*2 points*) Based on the correlation matrix, explain why basing pass or fail entirely on **G3**, as requested by School Wiz, may be a sensible decision.

ANSWER:

The correlation matrix indicates that `G1`, `G2`, and `G3` are strongly positively correlated with each other, with all pairwise correlations greater than 0.8, meaning that they tend to move in the same direction. This is perhaps not surprising given that they capture very similar information (grades in consecutive trimesters) about students, so simply using one of the three grade variables, say `G3`, to build a pass/fail classifier will not result in too much loss of information compared to using all of the three grade variables. The two modeling approaches (using `G3` only versus using all of `G1`, `G2`, and `G3`) will likely produce similar classifications for students.

(c) (*3 points*) Describe how principal components analysis can provide an alternative method for determining whether a student will pass or not based on all of **G1**, **G2**, and **G3**.

ANSWER:

Principal components analysis (PCA) is an analytic technique for summarizing high-dimensional data. It relies on the use of composite variables known as the principal components