Business Math I, Project I
Project Outline

William Flack and Wayne Hacker
Copyright ©Wayne Hacker 2006. All rights reserved.
Contents

1 Introduction to insurance
   1.1 Hazards .................................................. 3
   1.2 How do we get these numbers? ......................... 5

2 The project ................................................. 5
   2.1 The Sad Tale of Dan the Insurance Man ............... 6

3 The assignment ........................................... 7
   3.1 The Excel file ........................................... 7
      3.1.1 What makes a good spreadsheet? ................. 7
      3.1.2 Naming data ranges .............................. 8
   3.2 The written answers .................................. 9
   3.3 The presentation ....................................... 9

4 Calculating the fair value ................................ 10
   4.1 Factors affecting cancellation ....................... 10
   4.2 The problem ........................................... 10
   4.3 The assumptions ....................................... 11
   4.4 Independence ......................................... 11
   4.5 The calculations .................................... 13
In this project, you will use conditional probability and expected value to solve an insurance problem. Before we describe the project, we’ll discuss some basics of insurance.

1 Introduction to insurance

Suppose you own a house. There is a 1% chance that it will burn down in any given year. It would cost you $100,000 to replace it. You can’t do without a house, and you can’t afford a $100,000 loss.

One approach is to do nothing and hope that your house doesn’t burn down. However, if you’re in the unlucky 1%, you will find yourself living under a bridge.

Another approach is to pool risk with other people in a similar situation. If you look at a large number of people with houses like yours, 1% will suffer a $100,000 loss, and 99% will lose nothing. The average loss will be

\[
\text{Average Loss} = (0.01)(100,000) + (0.99)(0) = 1000
\]

Suppose that everyone in such a large group agrees to pay $1000 at the beginning of the year, with the money used to cover the losses of those whose houses burn down during the year. Then everyone in the group sustains a certain loss of $1000, but no one suffers a catastrophic loss of $100,000.

This is the essence of insurance. To avoid the small possibility of a large loss that you can’t afford, you accept the certainty of an expense that you can afford.

In practice, there will be other expenses than paying benefits to people whose houses burn down. Someone must be paid to collect and manage the money, to keep records of who’s insured, to investigate and pay claims, and to defend against lawsuits by claimants who’ve been refused. In reality, the insurance on the house might cost $1100 per year, with the extra $100 going to cover such additional expenses.

Let’s define some insurance terms, using our example. A person paying for insurance on his house is a policyholder. The organization that provides the insurance, whether it is a public company or a mutual insurance company (owned by the policyholders) is the insurer. The payment of $1100 for a year’s insurance is the premium. The $100,000 paid to people whose houses burn down is the benefit. The fair value of the policy is $1000: the expected value of the payout, i.e. the amount that the insurer would expect to pay to the average policyholder.

The premium must be larger than the fair value to cover expenses other than benefit payments. That means that the average policyholder will lose money. In our example, everybody pays the $1100 premium. 1% of the policyholders receive a benefit of $100,000. Then the average policyholder’s yearly loss is

\[
(1)(1100) - (0.01)(100,000) = 100
\]
That doesn’t mean that insurance is a scam or a swindle. The policyholder isn’t trying to make money; he’s trying to avoid catastrophic loss. In considering whether an insurance policy is a wise purchase, you should consider two things:

- How catastrophic would the loss be? If you can afford it, it might be better not to buy insurance. If you’ve got several million dollars in reasonably liquid assets, it might not make sense to insure a $100,000 house. If you’ve got several hundred thousand in assets, you might not want to buy a collision policy on a car worth $5000. And it almost never makes sense to buy the extended warranty on a $100 DVD player, whatever the salesman might say.

- How likely is the loss, and how much risk can you tolerate? It’s probably a good idea to insure your house against a 1% chance that it’ll burn down. On the other hand, it would probably be unwise to spend $100 on a policy that only pays if your house is demolished by a meteorite.

1.1 Hazards

So far, we’ve assumed that all policyholders are alike: that everyone’s house has an equal 1% chance of burning down. What if that’s not true?

Suppose that half of an insurer’s customers for house-fire policies live in Des Moines, and the other half live in Detroit. Let’s also suppose that the likelihood that a policyholder’s house in Des Moines will burn down in a particular year is 0.5%, but that for a house in Detroit the figure is 1.5%. Since the overall probability that a policyholder’s house will burn down is 1%, the fair value of a $100,000 house policy is $1000, and the company sells such policies for $1100.

Then competition shows up. Another company starts selling similar policies, but only to Des Moines residents. Since the likelihood that a Des Moines house will burn down is only 0.5%, the fair value of a $100,000 policy is

\[(0.005)(100,000) = 500\]

and the new company sells the policies for $600.

Very soon, the first company is going to see a drop in its sales, as its Des Moines policyholders switch over to the new company. It’s also going to have to raise its premium, because more than half of its customers will have high-risk Detroit houses. Eventually, it’s going to have only Detroit policyholders left, and the fair value of its $100,000 policies will be

\[(0.015)(100,000) = 1500\]

The alternative for the first company is to distinguish between Des Moines and Detroit policyholders. It could set a premium of $600 for the former, and $1600 for the latter.
The difference in price is based on the different probabilities that it will have to pay a benefit to a particular policyholder.

A condition that increases the probability or the expected severity of a loss is called a hazard. In our example, living in Detroit is a hazard.

We can express the effects of hazards using the language of conditional probability. Let $B$ be the event that a particular house burns down in a given year. Let $M$ be the event that the house is in Des Moines, and $T$ the event that the house is in Detroit. The probability that a house located in Des Moines burns down is

$$P(B|M) = 0.5\% = 0.005$$

and the probability that a house in Detroit burns down is

$$P(B|T) = 1.5\% = 0.015$$

Remember that half of the houses in our example are in Des Moines, and the other half are in Detroit. In equation form

$$P(M) = P(T) = 50\% = 0.5$$

The events $M$ and $T$ partition the set of all policyholders’ houses, since every house is either in Des Moines or Detroit, and no house is in both. Thus the total probability that a house will burn down is

$$P(B) = P(B|M)P(M) + P(B|T)P(T) = (0.005)(0.5) + (0.015)(0.5) = 0.01 = 1\%$$

To be competitive, our insurer should base premiums not on the total probability $P(B)$, but on the conditional probabilities $P(B|M)$ and $P(B|T)$. The fair value for a $100,000 policy for a Des Moines policyholder is

$$($100,000)$P(B|M) = ($100,000)(0.005) = $500$$

and the fair value for a Detroit policyholder is

$$($100,000)$P(B|T) = ($100,000)(0.015) = $1500$$

The location of the house might not be the only hazard that should be taken into account. Suppose that the presence of small children also increases the chance of fire. Then the insurer might also want to consider that when setting premiums. Let’s call $K$ (for “kids”) the event that there are small children in the house, and $K^C$ the event that there aren’t any. (Remember that the $C$ superscript stands for “complement”. If $E$ is the event that something happens, $E^C$ is the event that it doesn’t happen.) Then the event that someone both lives in Des Moines and has children in the house is the intersection $M \cap K$. The probability that a house burns down if it’s in Des Moines and has small children living in it is $P(B|M \cap K)$. 


1.2 How do we get these numbers?

In our example, we said that the probability of a house in Detroit burning down is 1.5%. We made that number up.

Real-world insurers don’t have that option. They need to know the probabilities of various events, and the effects of various hazards, as accurately as they can.

Unfortunately, the probability of an event can’t be measured directly. Insurers can’t hook up an instrument to a house and read the probability that it will burn down, or test an eight-year-old’s DNA to get the probability that he’ll torch his parents’ house while playing with matches.

In real life, actuaries for insurance companies have to estimate the probabilities of events based on historical records. For example, the tax records for Detroit might show the number of houses there; the Fire Department’s records would show the number of them that burn down in a given year. Using these records as our sample space \( \Omega \), we can estimate

\[
P(B|T) = \frac{\#(B \cap T)}{\#(T)} = \frac{\text{number of houses in Detroit that burn down}}{\text{total number of houses in Detroit}}
\]

Similarly, we might be able to find records giving the total number of American households with children, and the number of such houses that burn down. We can use those records to estimate \( P(B|K) \).

Finding \( P(B|T \cap K) \) might be more complicated. The records for houses that burn in Detroit might not show whether there are children present; and the records that note the presence or absence of children might not show the city in which the fire occurred. In that case, we have to make some assumptions and do some calculations to extract the number we want from incomplete data. This is the kind of thing that you’ll be doing in this project. We will explain the math after we’ve laid out the basics of the project.

2 The project

You are working for the Schadenfreude Insurance Company. The company is launching a new product: one-time cancellation insurance for concerts.

Each policy will cover one concert, and will pay the promoter a benefit if that concert must be cancelled. Remember that if an event is cancelled, the ticketholders generally receive full refunds. However, the promoter is still liable for many of the costs associated with the concert. Depending on the circumstances of the cancellation, he may have non-refundable expenses for things like advertising and promotion, rental of the performance site, hiring of security personnel, and performance fees for the bands. Since most concert promoters don’t have a lot of reserve cash, the cancellation of a single major event could be catastrophic.
The premium will be based on three pieces of information about the event to be insured: the location, the musical genre (country, bluegrass, jazz, etc.), and the benefit to be paid if the event is cancelled. You have two databases on concert cancellations: one by location, and one by genre. From these, you need to calculate the probability that your concert will be cancelled; and from that number, the fair value of the insurance policy. After you’ve calculated the fair value, you need to add a certain amount to cover your company’s overhead and allow for a profit. This amount will be $200, plus 10% of the fair value. Adding that to the fair value will give the premium.

2.1 The Sad Tale of Dan the Insurance Man

You are not the first person to work on this project. It was originally assigned to Dan, the company’s blue-chip employee. Dan had a reputation for accurate and reliable work; and he had a strong math background, having gone through a pre-actuarial program as an undergrad.

Unfortunately, Dan’s undergrad years were a long way behind him. Indeed, his fortieth birthday was only a memory. Unwilling to accept that his youth was no more, Dan attempted to halt Time’s wingèd chariot by spending his days on the elliptical trainer and his nights in pursuit of bad beer and worse women.

One Friday evening at the bars, Dan met a trio of women who expressed admiration for his rock-hard abs, and who subsequently returned to his apartment with him. There, they informed him that all they needed to make their erotic arousal complete was a big bucket of fried chicken. On the way to Popeye’s, Dan met some friends, whom he gloatingly informed of his forthcoming adventures. Those friends were the last people known to have seen him.

When police entered the apartment several days later, they found no one home and chicken strewn across the floor. What became of Dan is unknown. It may be that his new acquaintances whacked him on the head, looted the apartment, then left the corpse in a convenient dumpster. Or it may be that he flung down the bucket of chicken upon coming home to find his place empty of girls and valuables, then left town to avoid the merciless teasing he’d get from his friends. Several pieces of chicken were missing, supporting the former theory. On the other hand, there have been unconfirmed sightings of Dan panhandling in Chicago, passed out on a bar stool in Oakland, and working at a car wash in Baltimore.

From Schadenfreude’s perspective, the critical thing is that Dan and his laptop are missing. You have been assigned to reconstruct and complete Dan’s work.
3 The assignment

On the afternoon of his fateful Friday, Dan delivered a report, a template Excel file, and some sample results to the boss. The report explains the mathematics and describes how the Excel file works. The boss was to read the report over the weekend. On the following Monday, Dan was supposed to use the template file to show the boss how to construct the calculation, and answer any questions regarding the report. Unfortunately, that meeting was never to be. The boss was left with a complete report, an Excel file with the formulas missing, and a number of unanswered questions.

Your boss has given you the report and the template file. The report includes one of the sample results: for the case of a polka show, scheduled in Alaska, with a benefit of $30,000. He has also given you a list of questions that he came up with after reading the report. The boss is holding back some of the results to make sure that you don’t fudge your data. This way, he can test your Excel file against Dan’s results.

You need to complete the Excel file by putting in formulas. You must also turn in written answers to your boss’s list of questions. Finally, you must deliver a PowerPoint presentation to your boss and other Schadenfreude executives. Details of all these follow.

3.1 The Excel file

You must write formulas for the appropriate cells on the template spreadsheet. Notice that Dan’s report includes a description of what each formula cell should do. Your formulas should match these descriptions.

You can use the Alaska-polka-$30,000 case to check your formulas. If you’ve constructed them correctly, you should get the numbers that appear in Dan’s report.

3.1.1 What makes a good spreadsheet?

The spreadsheet will be used by other employees of the company to calculate premiums for concert-cancellation policies. It’s very important that it be clearly laid out and easy to use, so that those employees don’t have to spend unnecessary time and effort figuring it out. The written explanation must also be very clear, so that other employees can easily check your work or modify it later.

To use the spreadsheet, it should only be necessary to change the values in three cells: one for the region, one for the genre, and one for the benefit. The results (the fair value and the premium cost) are in a conspicuous position, and are unambiguously labelled.

The calculation in each cell should be relatively simple. In principle, you could calculate the fair value in a single cell, with one massive formula. However, it would be difficult indeed for others to understand and proofread such a formula. It’s better to proceed
in small steps, so that others can understand what you’re doing, and so that errors can
more easily be detected.

Notice that Dan’s spreadsheet uses names for the cells that make up the databases. There are several reasons for this. First of all, it’s easy to make a mistake when typing a
range reference like ‘Data1!A1:B13865’, especially if you have to do it repeatedly. If you
accidentally type ‘13685’ instead of ‘13865’, you probably won’t get an error message;
but your calculation will ignore the last 180 records, which will probably produce errors
in your results. If you give the range an easy name like ‘data1’, you’re not going to make
this kind of mistake; and if you do mistype the name, you will probably get a #NAME?
error, which will tip you off that something’s wrong. Second, suppose that you later want
to add additional data to the databases. If you’ve used straight cell references, you’ll
have to go through and find every formula that contains ‘Data1!A1:B13865’ and change
the reference. If you’ve used a name, then all you have to do is redefine the name once so
that it includes the new data; then, every formula that uses the name will automatically
be recalculated. Third, names are easier to understand. If someone else is trying to
figure out your Excel file, or if you’re working with it again after a long hiatus, it’s easier
to read formulas if they include names like ‘data1’ than if they use monstrous strings of
characters like ‘Data1!A1:B13865’.

If you don’t know about creating and using names for ranges, read the following section.

3.1.2 Naming data ranges

Excel allows you to give a name to a cell, or to a range of cells, and to use that name as
a reference in formulas.

The procedure is fairly simple. Highlight the cell or range that you want to name; then
go to the menu bar and click Insert > Name > Define. That will open a small window with
a top line in which you can type the new name; right below that will be a list of all the
names you’ve already used. Type the name and return or click “OK”.

Names can consist of up to 255 characters. They have to begin with a letter or an
underscore; after that, they can contain letters, numbers, underscores, and periods. You
can put uppercase letters in a name, but they’re not case-sensitive: for example, Excel
would regard ‘fiscalyear2006’ and ‘FiscalYear2006’ as the same thing. Names can’t be
the same as cell references: for example, you can’t use the name ‘FY2006’, because that’s

Once you’ve inserted a name, you can use it like you’d use the cell references. For
example, we gave the name ‘data1’ to the range ‘Data1!A1:B13865’. That means that
instead of typing

\[=\text{DCOUNT}(\text{Data1!A1:B13865}, \text{H16:I17})\]

we can use

\[=\text{DCOUNT}(\text{data1}, \text{H16:I17})\]
The latter is easier to type, and less subject to error.

To check existing names, there are two things you can do. If you think you’ve named a cell or range but can’t remember the name, highlight that cell or range. Above the worksheet, at the left end of the formula bar, is a small box called the name box. If the cell or range you’ve highlighted has a name, that will appear there. (Try it: go to the Project 1 Excel file and highlight Data1!A1:B13865; see if ‘data1’ appears in that box.) Alternatively, you can put the cursor anywhere, then go to the menu bar and click Insert>Name>Define. When the small window opens up, go to the large box with the list of existing names. Click on one of those, and the range that the name refers to should appear in the bottom line of the window, under ‘Refs to:’

To change an existing name, click Insert>Name>Define. When the small window opens, click on the name you want to change in the list of names. The ‘Refs to:’ box at the bottom of the window will show the range that the name refers to. Click in that box and change the range to the new value.

3.2 The written answers

Your boss has given you a list of questions for which he wants written answers. Several of these are taken from sample results that Dan computed before his disappearance. The boss has given you the genre, location, and benefit, and expects you to come up with the fair value and premium. There may also be questions on the mathematics involved. Your answers to his questions should be written in a Word file and handed in with the completed Excel spreadsheet.

3.3 The presentation

Your presentation will be delivered to a group of Schadenfreude executives, after they’ve had the opportunity to read your report. The presentation itself should be brief. It should emphasize the big picture: what the new product is, and how you’re calculating fair value for the policy. You shouldn’t spend the presentation explaining the fine details of the mathematics. However, there will be a question-and-answer period at the end of the presentation, during which you must be able to field questions about anything in the report or the spreadsheet, including the math. At this time, it is highly likely that your boss will give you further combinations of genre, location, and benefit; you will use your spreadsheet on the spot to determine the fair value and premium. Failure to come up with the correct numbers may adversely affect your future employment.
4 Calculating the fair value

Suppose a particular concert has probability $p$ of being cancelled; and if it is cancelled, the promoters receive a benefit of $b$. The fair value $v$ is the benefit that the average policyholder would receive. From the insurer’s standpoint, the fair value is the premium for which the amount of money taken in in premiums would equal the amount paid out in benefits. The policyholder has a probability $p$ of receiving a benefit of $b$; so the fair value is

$$v = pb$$

You are given the benefit $b$; so to calculate the fair value, you need to figure out the probability that the concert will be cancelled, $p$.

4.1 Factors affecting cancellation

Concerts may be cancelled for a variety of reasons. Some of these reasons vary with the genre; others, with the region of the country in which they’re held.

Genre affects concert cancellations. Rap/hip-hop artists may be more likely to cancel because of drive-by shootings than lite-jazz performers; polka bands are more subject to bratwurst-related injuries than metal bands. Thus genre should be taken into account when calculating the probability of cancellation.

The location of the concert also matters. Tornadoes are more likely to cause cancellation in Kansas City than in Seattle. Tsunamis are a problem in Honolulu, but a non-issue in Indianapolis.

You have two databases in the Excel sheet. One, on the sheet ‘Data1’, lists a number of concerts; for each one, it gives the location and whether or not the show was cancelled. The other, on the sheet ‘Data2’, lists concerts, their genres, and whether or not they were cancelled.

To get the fair value of the policy for a particular concert, you want the conditional probability that the concert will be cancelled given the genre and location. If $G$ is the event that the concert is in a particular genre, $R$ is the event that the concert is to be held in a particular region of the country, and $C$ is the event that the concert is cancelled, you want to calculate

$$p = P(C|G \cap R).$$

4.2 The problem

It’d be great if we could go through a single database, count all the cancelled concerts for a particular genre and region, count the total number of concerts for the same genre
and region, and divide the first number by the second:

\[ P(C|G \cap R) = \frac{\#(C \cap G \cap R)}{\#(G \cap R)} = \frac{\# \text{ cancelled concerts in the genre and region}}{\# \text{ all concerts in the genre and region}} \]

The problem is that our databases come from two different sources, and contain two different sets of information. The worksheet ‘Data1’ was compiled by the Federal Minor Inconvenience Management Agency; it contains data on regions, but not on genres. The sheet ‘Data2’ was put together by a music-industry association; it has information on genres, but not on regions. This means that for any given concert, we can’t tell whether it’s a member of the set \( G \cap R \).

We can get around this problem, but it will take some calculation. It will also require us to make certain assumptions.

### 4.3 The assumptions

We’re going to make three basic assumptions:

1. The two databases describe similar concerts. Probabilities that we calculate from one database are good for the concerts described in the other.

2. The databases are a good sample. Probabilities that we calculate using relative frequencies in the databases are close to real-world probabilities.

3. The events \( R \) and \( G \) are independent, and this remains true if they’re conditioned on the event \( C \). If you’re having trouble with the concept of independence, read the next section. We’ll discuss this assumption a little further at the end.

### 4.4 Independence

The dependence or independence of two events is a fairly intuitive concept. Essentially, the events \( A \) and \( B \) are independent if knowing one doesn’t help you predict the other.

For example, let \( A \) be the event that a person is male. Let \( B \) be the event that the person lives in France. Assuming that the sex ratio in France is the same as it is in the rest of the world, then knowing the sex of a person doesn’t help us predict whether that person lives in France; and knowing that someone lives in France doesn’t give us any clues as to the sex. \( A \) and \( B \) are independent.

On the other hand, if \( A \) is the event that a person speaks fluent French, and \( B \) is again the event that the person lives in France, then the events are dependent. Knowing whether one is true can help you predict the other. If \( A \) is true, \( B \) is more likely to be true as well, and vice versa.
The events are also dependent if $A$ is the event that a person speaks fluent Chinese, and $B$ is the event that the person lives in France. This may seem like the opposite situation: now, if $A$ is true, then $B$ is more likely to be false, and vice versa. Again, though, knowing one helps you predict the other.

If $A$ and $B$ are independent, we can write three equations:

$$
P(A|B) = P(A) \quad P(B|A) = P(B)
$$

$$
P(A \cap B) = P(A)P(B)
$$

(4.1)

The first of these tells us that knowing $B$ doesn’t help us predict $A$. The second tells us that knowing $A$ doesn’t help us predict $B$. The third equation is one that we will use in this project: it tells us, for example, that the probability that a random person is male ($A$) and lives in France ($B$) is the product of $P(A)$ and $P(B)$.

Events can also be dependent or independent when they’re conditioned on other events. Suppose $A$ is the event that a person is male; $B$ is the event that the person goes to college; and $C$ is the event that the person lives in France. Assuming that men and women in France are equally likely to go to college, we can say that $A$ and $B$ are independent when conditioned on $C$; and we can write a conditional version of equation (4.1):

$$
P(A \cap B|C) = P(A|C)P(B|C)
$$

(4.2)

Conditioning can create or destroy independence. In this example, $A$ and $B$ are probably not independent: over the whole world, men are probably more likely to go to college than women. Knowing $A$ can help us predict $B$, and

$$
P(A \cap B) \neq P(A)P(B)
$$

Thus two dependent events $A$ and $B$ become independent when conditioned on a third event $C$.

The opposite is also possible: two independent events $A$ and $B$ might become dependent when conditioned on a third event $C$. You might try to think of an example of this.

Let’s consider our assumption that $R$ and $G$ are independent. In essence, we’re saying that knowing the genre of a concert doesn’t give us any clues as to where in the country it’s held, and that knowing the location doesn’t tell us anything about the genre. This might seem debatable. We’re accustomed to thinking of musical tastes as varying with the part of the country: people in Texas listen to country, people in Tennessee listen to bluegrass, people in Wisconsin listen to polka...

However, on second thought the assumption might not seem so incorrect. The influence of national radio and television networks has greatly reduced the difference in musical preference between different parts of the country. The tendency of Americans to make long-distance moves at least once in their lives has had the same effect. Our assumption may not be perfectly accurate; but it’s not unreasonable, either.
4.5 The calculations

Recall that $C$ is the event that a concert is cancelled. $G$ is the event that it belongs to a particular genre. $R$ is the event that it is scheduled in a particular region of the country. You need to calculate $\mathcal{P}(C \mid R \cap G)$.

Remember the formula to calculate a conditional probability:

$$\mathcal{P}(A \mid B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)} \quad (4.3)$$

Let’s apply this to $\mathcal{P}(C \mid R \cap G)$. Don’t be intimidated by the fact that instead of a simple $B$, you’ve now got $R \cap G$. The intersection of sets is associative, which is to say that you can write

$$C \cap (R \cap G) = C \cap R \cap G$$

We’ll substitute $C$ for $A$ and $R \cap G$ for $B$ in Equation (4.3):

$$\mathcal{P}(C \mid R \cap G) = \frac{\mathcal{P}(C \cap R \cap G)}{\mathcal{P}(R \cap G)}$$

By our third assumption, $R$ and $G$ are independent. Equation (4.1) allows us to rewrite the denominator:

$$\mathcal{P}(C \mid R \cap G) = \frac{\mathcal{P}(C \cap R \cap G)}{\mathcal{P}(R)\mathcal{P}(G)} \quad (4.4)$$

Our first and second assumptions allow us to calculate the two factors of the denominator. $\mathcal{P}(R)$ is the probability that a randomly selected concert will be in your region. ‘Data1’ contains information on regions, so we’ll treat that as our sample space $\Omega_1$. To find $\mathcal{P}(R)$, count all the concerts in ‘Data1’, then count the concerts that are in your region. Divide the second number by the first:

$$\mathcal{P}(R) = \frac{\#(R)}{\#(\Omega_1)} = \frac{\# \text{ concerts in your region}}{\# \text{ all concerts in ‘Data1’}} \quad (4.5)$$

In the same way, $\mathcal{P}(G)$ is the probability that a randomly selected concert will be in your genre. ‘Data2’ gives information on genres, so we’ll use that as the sample space $\Omega_2$.

$$\mathcal{P}(G) = \frac{\#(G)}{\#(\Omega_2)} = \frac{\# \text{ concerts in your genre}}{\# \text{ all concerts in ‘Data2’}} \quad (4.6)$$

That was the easy part. We still need to calculate $\mathcal{P}(C \cap R \cap G)$, the numerator in Equation (4.4). This will require a little cleverness.

We want the probability of the intersection of several events. That suggests that we might do well to look at conditional probabilities. Equation (4.3) gives us

$$\mathcal{P}(G \cap R \mid C) = \frac{\mathcal{P}(G \cap R \cap C)}{\mathcal{P}(C)} \quad (4.7)$$
Intersection of sets is commutative: \( A \cap B = B \cap A \). That means that \( \mathcal{P}(G \cap R \cap C) = \mathcal{P}(C \cap R \cap G) \), which is what we want. We can multiply both sides of Equation (4.7) by \( \mathcal{P}(C) \) to get
\[
\mathcal{P}(C \cap R \cap G) = \mathcal{P}(G \cap R \cap C) \mathcal{P}(C)
\] (4.8)

Now, we can use Assumption 3 again. Remember that we assumed that \( R \) and \( G \) were independent, and remained so when conditioned on \( C \). That means we can use Equation (4.2) to write
\[
\mathcal{P}(G \cap R \mid C) = \mathcal{P}(G \mid C) \mathcal{P}(R \mid C)
\]
Substituting this into Equation (4.8) gives us
\[
\mathcal{P}(C \cap R \cap G) = \mathcal{P}(G \mid C) \mathcal{P}(R \mid C) \mathcal{P}(C)
\] (4.9)

We can calculate all three factors on the right-hand side from the databases:
\( \mathcal{P}(G \mid C) \) is the probability that a cancelled concert will be in your genre. We’ll use ‘Data2’ as the sample space \( \Omega_2 \), and we’ll denote the cancelled concerts from ‘Data2’ as \( C_2 \). Then
\[
\mathcal{P}(G \mid C) = \frac{\#(G \cap C_2)}{\#(C_2)} = \frac{\text{# cancelled concerts in your genre in ‘Data2’}}{\text{# all cancelled concerts in ‘Data2’}}
\]
\( \mathcal{P}(R \mid C) \) is the probability that a cancelled concert will be in your region; calculate it using ‘Data1’ as the sample space \( \Omega_1 \), with the cancelled concerts from ‘Data1’ making up the set \( C_1 \):
\[
\mathcal{P}(R \mid C) = \frac{\#(R \cap C_1)}{\#(C_1)} = \frac{\text{# cancelled concerts in your region in ‘Data1’}}{\text{# all cancelled concerts in ‘Data1’}}
\]
\( \mathcal{P}(C) \) is the probability that a concert will be cancelled. We can get this information from both databases, so we’ll use the union of them as the sample space \( \Omega_{12} = \Omega_1 \cup \Omega_2 \). We’ll use the set of cancelled concerts from both databases as \( C_{12} = C_1 \cup C_2 \). Then we calculate
\[
\mathcal{P}(C) = \frac{\#(C_{12})}{\#(\Omega_{12})} = \frac{\text{# cancelled concerts in both databases}}{\text{# all concerts in both databases}}
\]
Substituting Equation (4.9) back into Equation (4.4) gives us the final formula:
\[
\mathcal{P}(C \mid R \cap G) = \frac{\mathcal{P}(G \mid C) \mathcal{P}(R \mid C) \mathcal{P}(C)}{\mathcal{P}(R) \mathcal{P}(G)}
\]
We can calculate all the probabilities in this formula from the databases.

Now, multiply the benefit \( b \) by the cancellation probability to get the fair value:
\[
 v = b \cdot \mathcal{P}(C \mid R \cap G)
\]
To calculate the premium, we need to add a certain amount to the fair value to cover costs other than benefit payments. As discussed on page 6, this amount is $200, plus 10% of the fair value. Thus the premium \( r \) is:
\[
r = v + 200 + 0.1v = 1.1v + 200
\]