Figure A-35 shows a version of the unnormalized insect data discussed in Example 1-3.

FarmID	FarmName	Field	Date	Visit	SampleID	Insect	Count
1	HighGate	F2	09-Feb-11	14	3	Beetle	2
1	HighGate	F2	09-Feb-11	14	2	Beetle	4
1	HighGate	F1	09-Feb-11	14	1	Beetle	4
1	HighGate	F1	18-Mar-11	15	1	Springtail	5
1	HighGate	F2	09-Feb-11	14	3	Springtail	3
1	HighGate	F2	09-Feb-11	14	2	Springtail	5
1	HighGate	F1	09-Feb-11	14	1	Springtail	6
1	High-Gate	F1	18-Mar-11	15	1	Beetle	7
2	Greyton	F2	09-Feb-11	16	1	Beetle	2
2	Greyton	F1	09-Feb-11	16	2	Beetle	4
2	Greyton	F1	09-Feb-11	16	2	Springtail	5
2	Greyton	F2	09-Feb-11	16	1	Springtail	3

Figure A-35. *Unnormalized insect data*

Excercise 8-3

a) What are some of the updating problems that could occur?

The most obvious problems are likely to arise from inconsistent data caused by repeated information. The name of each farm is recorded several times and has inevitably resulted in a spelling error for HighGate. The fact that Visit 14 is on February 9th is recorded several times and a data input error is possible. Because the data was stored in a spreadsheet there is no primary key, so the insertion and deletion problems don't arise.

b) Which of the following functional dependencies hold?

- `FarmID`→`FarmName`? Yes. If I know the ID I can tell you the name (or I should be able to do so if the spelling errors are removed).
- `FarmID`→`Visit`? No. Farm 1 has two associated visits (14 and 15) in the data shown in Figure A-35. If this functional dependency was true that would mean only one visit per farm—clearly not the intention.
- `Visit`→`Date`? This is certainly true for the data shown, but will it always be true? It would mean that a visit was on a single day—something to be checked with the client. We'll say yes for now.
- `Date`→`Visit`? No. That would mean only one visit a day. February 9th has two associated visits: 14 and 16.
- `Visit`→`FarmID`? The data shown support this. In general we need to confirm that a visit is for just one farm (i.e., it doesn't include a tour around the country). We'll say yes.

- `SampleID`→`Field`? If we know the sample we should surely know which field it comes from, right? However, some rows with `SampleID` having a value 2 have F1 for the `Field` value while others have F2. So, no—we can't tell the value of `Field` from knowing the value of `SampleID`. This is because the `SampleID` field doesn't uniquely identify a sample. There are several `SampleID` 2 s from different visits. To identify a sample we need to also know on which visit it was collected.

- (`SampleID, VisitID`)→`Field`? Yes—see preceding! The data supports this. Sample 2 for Visit 14 is associated with only F2 and Sample 2 for Visit 16 is associated only with F1.

- (`SampleID, Insect`)→`Count`? In an actual sample, if we specify the insect then we should come up with a unique count. But remember, the `SampleID` field doesn't uniquely identify an actual sample. Take a look at the rows where `SampleID` = 2, and `Insect` = Beetle. (What a coincidence!)

- `VisitID, SampleID, Insect`→`Count`? See preceding. Now that we include the `VisitID` with `SampleID`, we can identify the actual test tube (or whatever) and know the number of Beetles.

c) Is (`VisitID, SampleID, Insect`) a suitable key?

From question 2, let's summarize the functional dependencies (FD) that hold for this data.

FD1: FarmID -> FarmName

FD2: VisitID -> Date, FarmID

FD3: SampleID, VisitID -> Field

FD4: VisitID, SampleID, Insect -> Count

What field values can I determine if I have values for (VisitID, SampleID and Insect)?

We know Count (FD4), Field (FD3), Date, and FarmID (FD2), and because I now know FarmID I can tell you FarmName (FD1). That's everything, so this set is a key. It is a primary key because if you remove any of the fields then you won't know everything (I'll leave that to the really keen readers to verify).

d) Decompose the data in Figure A-35 into a set of tables in 3rd Normal form.

Here are the fields (with the primary key underlined):

(FarmID, <u>VisitID, SampleID</u>, Date, FarmName, Field, <u>Insect</u>, Count)

Is this in 1st Normal Form?

Yes. We are not trying to put more than one piece of information in a cell.

Is this in 2nd Normal Form, or equivalently do the non-key fields depend on the whole key?

No. We can determine FarmName by just knowing FarmID (FD1), so it should be removed into another table with what it depends on. This leaves us with:

(FarmID, VisitID, SampleID, Date, ~~FarmName~~, Field, Insect, Count)
(FarmID, FarmName) (new table)

Date and FarmID can be determined by knowing just Visit (FD2). Removing these two fields leaves us with:

(~~FarmID~~, VisitID, SampleID, ~~Date~~, ~~FarmName~~, Field, Insect, Count)
(VisitID, FarmID, Date) (new table)

Field can be determined by knowing just SampleID and VisitID (FD3). So this needs to be removed.

(~~FarmID~~, VisitID, SampleID, ~~Date~~, ~~FarmName~~, ~~Field~~, Insect, Count)
(VisitID, SampleID, Field) (new table)

Finally we have the following four table (keys underlined, foreign keys italicized).

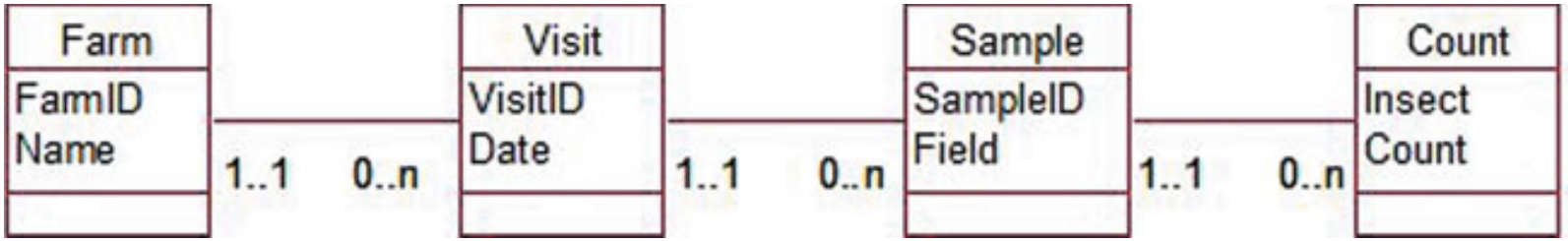
(FarmID, FarmName)

(VisitID, *FarmID*, Date)

(*VisitID*, SampleID, Field)

(*VisitID*, *SampleID*, Insect, Count) (remains of original table)

These tables would have been derived if we had started with the class diagram in Figure A-36.



ure A-36. *Class diagram for insect data*

The class diagram essentially says a farm gets many visits. Each visit has a number of associated samples and for each sample we can record the numbers of insects. I think you will see that starting with the class diagram would have been a great deal less trouble and would also have made us think about some other possible classes, such as `Field` and `Insect`. However, the two approaches are equivalent ways of arriving at similar outcomes.