IES302 2011/1 Part I.2 Dr.Prapun

5 Probability Foundations

Constructing the mathematical foundations of probability theory has proven to be a long-lasting process of trial and error. *The approach consisting of defining probabilities as relative frequencies in cases of repeatable experiments leads to an unsatisfactory theory.* The frequency view of probability has a long history that goes back to **Aristotle**. It was not until 1933 that the great Russian mathematician A. N. **Kolmogorov** (1903-1987) laid a satisfactory mathematical foundation of probability theory. He did this by taking a number of axioms as his starting point, as had been done in other fields of mathematics. [20, p 223]

We will try to avoid some technical details in this class. Therefore, the definition given below is not the "complete" definition. Some parts are modified or omitted to make the definition easier to understand.

P(A)

Definition 5.1. Kolmogorov's Axioms for Probability [9]: A **probability measure** is a real-valued (set) function[11] that satisfies

P1 ***Nonnegativity***:

$$P(A) \geq 0.$$

P2 ***Unit normalization***:

$$P(\Omega) = 1.$$

[11] A real-valued set function is a function the maps sets to real numbers.

P3 ***Countable additivity*** or ***σ-additivity***: For every countable sequence $(A_n)_{n=1}^{\infty}$ of disjoint events,

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

disjoint
$P(A \cup B) = P(A) + P(B)$
disjoint
$P(A \cup B \cup C) = P(A) + P(B) + P(C)$

- The number $P(A)$ is called the ***probability*** of the event A
- The entire sample space Ω is called the ***sure event*** or the ***certain event***.
- If an event A satisfies $P(A) = 1$, we say that A is an ***almost-sure event***.
- A ***support*** of P is any set A for which $P(A) = 1$.

Axioms do not determine probabilities. Axioms enable us to calculate the probabilities of some events from the knowledge of the probabilities of other events.

From the three axioms above, we can derive many more properties of probability measure. These properties are useful for calculating probabilities.

5.2. $P(\emptyset) = 0$.

5.3. Finite additivity: If $A_1, \ldots, A_n$ are disjoint events, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Special case when $n = 2$: ***Addition rule*** (Additivity)

$$\text{If } A \cap B = \emptyset, \text{ then } P(A \cup B) = P(A) + P(B). \tag{4}$$

5.4. If A is countable, then

$\{a_1, a_2, \ldots\}$
$\{a_1\} \cup \{a_2\} \cup \cdots$

$$P(A) = \sum_{n=1}^{\infty} P(\{a_n\}).$$

Example:
$\Omega = \{a, b, c, d\}$
$P(\{a\}) = 0.1$
$P(\{b\}) = 0.3$
$P(\{c\}) = 0.5$
$P(\{d\}) = 0.1$

$P(\{1,2\}) = P(\{1\}) + P(\{2\})$

$A = \{a, b\}$

$P(A) = P(\{a,b\}) = P(\{a\}) + P(\{b\})$
$= 0.1 + 0.3 = 0.4$

Similarly, if A is finite, then

$\{a_1, a_2, \ldots, a_{|A|}\}$

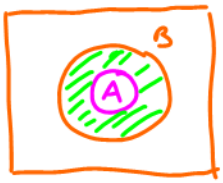
$$P(A) = \sum_{n=1}^{|A|} P(\{a_n\}).$$

$B = \{b, c, d\}$

$P(B) = 0.3 + 0.5 + 0.1 = 0.9$

5.5. ***Monotonicity***: If $A \subset B$, then $P(A) \leq P(B)$

$P(A \cap B) = P(\{b\})$
$= 0.3$



$\Downarrow$

$B = A \cup (B \setminus A)$
↑ disjoint union

$P(B) = P(A) + P(B \setminus A)$

$P(B) - P(A) = P(B \setminus A) \geq 0$

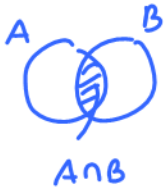
Example 5.6. Let A be the event to roll a 6 and B the event to roll an even number. Whenever A occurs, B must also occur. However, B can occur without A occurring if you roll 2 or 4.

5.7. If $A \subset B$, then $P(B \setminus A) = P(B) - P(A)$

$P(\Omega) = 1$

5.8. $P(A) \in [0, 1]$. $A \subset \Omega \Rightarrow P(A) \leq P(\Omega) = 1$

5.9. $P(A \cap B)$ can not exceed $P(A)$ and $P(B)$. In other words, "the composition of two events is always less probable than (or at most equally probable to) each individual event."



$A \cap B \subset A$ $\quad$ $A \cap B \subset B$

$P(A \cap B) \leq P(A)$ $\quad$ $P(A \cap B) \leq P(B)$

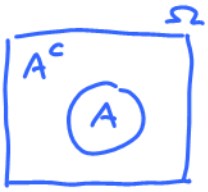
Example 5.10. Let us consider Mrs. Boudreaux and Mrs. Thibodeaux who are chatting over their fence when the new neighbor walks by. He is a man in his sixties with shabby clothes and a distinct smell of cheap whiskey. Mrs.B, who has seen him before, tells Mrs. T that he is a former Louisiana state senator. Mrs. T finds this very hard to believe. "Yes," says Mrs.B, "he is a former

state senator who got into a scandal long ago, had to resign, and started drinking." "Oh," says Mrs. T, "that sounds more likely." "No," says Mrs. B, "I think you mean less likely."

Strictly speaking, Mrs. B is right. Consider the following two statements about the shabby man: "He is a former state senator" and "He is a former state senator who got into a scandal long ago, had to resign, and started drinking." It is tempting to think that the second is more likely because it gives a more exhaustive explanation of the situation at hand. However, this reason is precisely why it is a less likely statement. Note that whenever somebody satisfies the second description, he must also satisfy the first but not vice versa. Thus, the second statement has a lower probability (from Mrs. Ts subjective point of view; Mrs. B of course knows who the man is).

This example is a variant of examples presented in the book *Judgment under Uncertainty* by Economics Nobel laureate Daniel Kahneman and co-authors Paul Slovic and Amos Tversky. They show empirically how people often make similar mistakes when they are asked to choose the most probable among a set of statements. It certainly helps to know the rules of probability. A more discomforting aspect is that the more you explain something in detail, the more likely you are to be wrong. If you want to be credible, be vague. [15, p 11–12]

5.11. Complement Rule:



$$P\left(A^c\right) = 1 - P\left(A\right).$$

$A \cup A^c = \Omega$

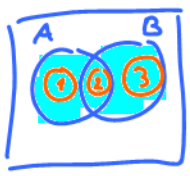
$P(A) + P(A^c) = P(\Omega) = 1$

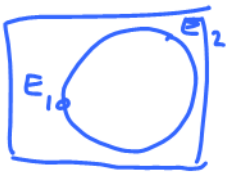
- "The probability that something does not occur can be computed as one minus the probability that it does occur."
- Named "probability's Trick Number One" in *Taking Chances: Winning with Probability*, by British probabilist Haigh.

More general union

5.12. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A) + \left(P(B) - P(A \cap B)\right)$





Combining two errors E_1, E_2

$P(E_1 \cup E_2) \approx P(E_2)$ if $P(E_2) \gg P(E_1)$

- $P(A \cup B) \leq P(A) + P(B)$.
- Approximation: If $P(A) \gg P(B)$ then we may approximate $P(A \cup B)$ by $P(A)$.

Example 5.13. In his bestseller *Innumeracy*, John Allen Paulos tells the story of how he once heard a local weatherman claim that there was a 50% chance of rain on Saturday and a 50% chance of rain on Sunday and thus a 100% chance of rain during the weekend. Clearly absurd, but what is the error?

Answer: Faulty use of the addition rule (4)!

If we let A denote the event that it rains on Saturday and B the event that it rains on Sunday, in order to use $P(A \cup B) = P(A) + P(B)$, we must first confirm that A and B cannot occur at the same time ($P(A \cap B) = 0$). More generally, the formula that is always holds regardless of whether $P(A \cap B) = 0$ is given by 5.12:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

The event "$A \cap B$" describes the case in which it rains both days. To get the probability of rain over the weekend, we now add 50% and 50%, which gives 100%, but we must then subtract the probability that it rains both days. Whatever this is, it is certainly more than 0 so we end up with something less than 100%, just like common sense tells us that we should.

You may wonder what the weatherman would have said if the chances of rain had been 75% each day. [15, p 12]

5.14. If a (finite) collection $\{B_1, B_2, \ldots, B_n\}$ is a partition of Ω, then

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

Similarly, if a (countable) collection $\{B_1, B_2, \ldots\}$ is a partition of Ω, then

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i)$$

5.15. Connection to classical probability theory: Consider an experiment with **finite** sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ in which each outcome ω_i is **equally likely**. Note that $n = |\Omega|$.

$P(\{\omega_i\}) = \alpha$

$= \frac{1}{n}$

$P(\Omega) = P(\{\omega_1\}) + P(\{\omega_2\}) + \cdots + P(\{\omega_n\})$

$1 \qquad \alpha \qquad \alpha \qquad \alpha$

$1 = n\alpha$

We must have

$$P(\{\omega_i\}) = \frac{1}{n}, \quad \forall i.$$

Now, given any event A, we can write A as a disjoint union of singletons:

$$A = \bigcup_{\omega \in A} \{\omega\}.$$

After applying finite additivity from 5.3, we have

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} \frac{1}{n} = \frac{|A|}{n} = \frac{|A|}{|\Omega|}.$$

same formula from classical probability

We can then say that the probability theory we are working on right now is an extension of the classical probability theory. When the conditons/assumptions of classical probability theory are met, then we get back the defining definition of classical classical probability. The extended part gives us ways to deal with situation where assumptions of classical probability theory are not satisfied.

6 Event-based Independence and Conditional Probability

Example 6.1. Diagnostic Tests.

6.1 Event-based Conditional Probability

Definition 6.2. ***Conditional Probability***: The conditional probability $P(A|B)$ of event A, given that event $B \neq \emptyset$ occurred, is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{5}$$

P(A)

- Read "the (conditional) probability of A given B".
- Defined only when $P(B) > 0$.
- If $P(B) = 0$, then it is illogical to speak of $P(A|B)$; that is $P(A|B)$ is not defined.

6.3. ***Interpretation***: Sometimes, we refer to $P(A)$ as

- **a priori probability** , or
- the **prior probability** of A, or
- the **unconditional probability** of A.

P(A|B)
a posteriori probability
posterior probability
conditional probability

It is sometimes useful to interpret $P(A)$ as our knowledge of the occurrence of event A *before* the experiment takes place. Conditional probability $P(A|B)$ is the ***updated probability*** of the event A given that we now know that B occurred (but we still do not know which particular outcome in the set B occurred).

The term **a posteriori** is often used for $P(A|B)$ when we refer to $P(A)$ as a priori.

Example 6.4. Roll a fair dice. Let X be the outcome.

X can be 1, 2, 3, ..., 6

Let A be the event that $X = 2$

Let B be the event that you saw

$B = \{3, 5\}$

$P(A) = P[X = 2] = \frac{1}{6}$

$A = \{2\}$

$P(A|B) = \frac{P(A \cap B)}{P(B)} = 0$

Let C be the event that $X = 5$

$P(C) = \frac{1}{6}$

$P(C|B) = \frac{P(C \cap B)}{P(B)} = \frac{P(\{5\})}{P(\{3,5\})} = \frac{1/6}{\frac{1}{6}+\frac{1}{6}} = \frac{1}{2}$

Example 6.5. In the diagnostic tests example, we learn whether we have the disease from test result. Originally, before taking the test, the probability of having the disease is 0.01%. Being tested positive from the 99%-accurate test ***updates*** the probability of having the disease to about 1%.

More specifically, let D be the event that the testee has the disease and T_P be the event that the test returns positive result.

- Before taking the test, the probability of having the disease is $P(D) = 0.01\%$. $= 10^{-4}$
- Using 99%-accurate test means

$$P(T_P|D) = 0.99 \text{ and } P(T_P^c|D^c) = 0.99.$$

- Our calculation shows that $P(D|T_P) \approx 0.01$.

Note also that although the symbol $P(A|B)$ itself is practical, it phrasing in words can be so unwieldy that in practice, less formal descriptions are used. For example, we refer to "the probability that a tested-positive person has the disease" instead of saying "the conditional probability that a randomly chosen person has the disease given that the test for this person returns positive result."

6.6. If the occurrence of B does not give you more information about A, then

$$P(A|B) = P(A) \tag{6}$$

and we say that A and B are ***independent***.

- Meaning: "learning that event B has occurred does not change the probability that event A occurs."
- Interpretation: "the occurrence of event A is not contingent on the occurrence (or nonoccurrence) of event B."

We will soon define "independence". Property (6) can be regarded as a "practical" definition for independence. However, there are some "technical" issues that we need to deal with when we actually define independence.

6.7. Similar properties to the three probability axioms:

(a) Nonnegativity: $P(A|B) \geq 0$

(b) Unit normalization: $P(\Omega|B) = 1$.

In fact, for any event A such that $B \subset A$, we have $P(A|B) = 1$.

This implies

$$P(\Omega|B) = P(B|B) = 1.$$

(c) Countable additivity: For every countable sequence $(A_n)_{n=1}^{\infty}$ of disjoint events,

$$P\left(\bigcup_{n=1}^{\infty} A_n \middle| B\right) = \sum_{n=1}^{\infty} P(A_n|B).$$

- In particular, if $A_1 \perp A_2$, $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$

6.8. Properties:

- $P(A|\Omega) = P(A)$
- If $B \subset A$ and $P(B) \neq 0$, then $P(A|B) = 1$.
- If $A \cap B = \emptyset$ and $P(B) \neq 0$, then $P(A|B) = 0$
- $P(A^c|B) = 1 - P(A|B)$ $P(A|B^c) = 1 - P(A|B)$
- $P(A \cap B|B) = P(A|B)$
- $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$.
- $P(A \cap B) \leq P(A|B)$

6.9. When Ω is finite and all outcomes have equal probabilities, classical case

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{|A \cap B| \,/\, |\Omega|}{|B| \,/\, |\Omega|} = \frac{|A \cap B|}{|B|}.$$

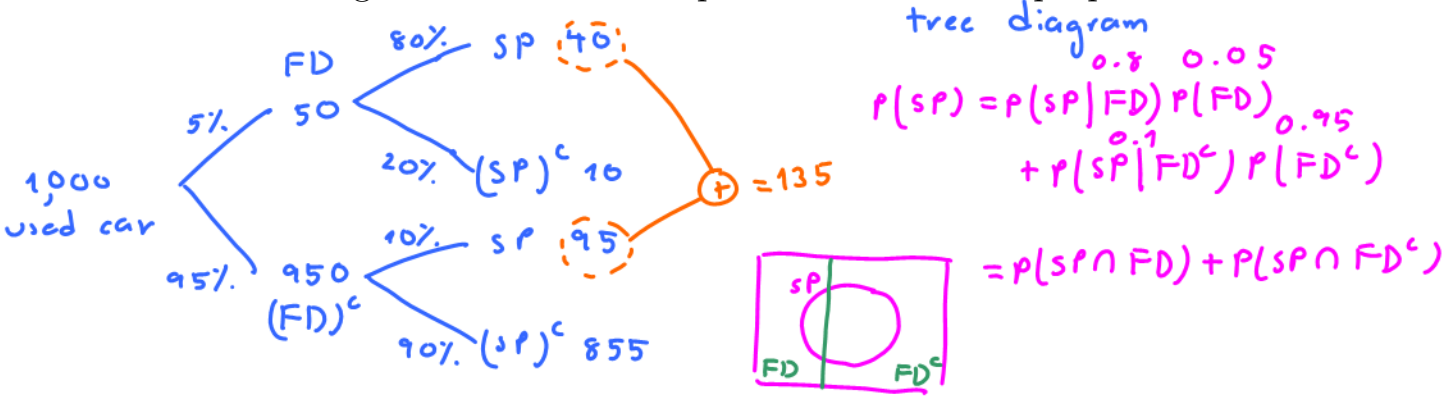
This formula can be regarded as the classical version of conditional probability.

Example 6.10. Someone has rolled a fair die twice. You know that one of the rolls turned up a face value of six. The probability that the other roll turned up a six as well is $\frac{1}{11}$ (not $\frac{1}{6}$). [20, Example 8.1, p. 244]

Example 6.11. You know that roughly 5% of all used cars have been flood-damaged and estimate that 80% of such cars will later develop serious engine problems, whereas only 10% of used cars that are not flood-damaged develop the same problems. Of course, no used car dealer worth his salt would let you know whether your car has been flood damaged, so you must resort to probability calculations. What is the probability that your car will later run into trouble?

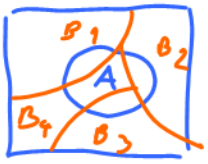
↳ $\frac{135}{1000} = 0.135$

You might think about this problem in terms of proportions.



If you solved the problem in this way, congratulations. You have just used the law of total probability.

6.12. ***Total Probability Theorem***: If a (finite or infinitely) countable collection of events $\{B_1, B_2, \ldots\}$ is a partition of Ω, then



$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + P(A \cap B_4)$

$$P(A) = \sum_i P(A|B_i)P(B_i). \tag{7}$$

$= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3) + P(B_4)P(A|B_4)$

This is a formula for computing the probability of an event that can occur in different ways.

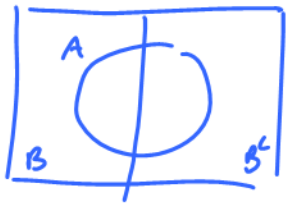
Example 6.13. The probability that a cell-phone call goes through depends on which tower handles the call.

The probability of internet packets being dropped depends on which route they take through the network.

6.14. Special case:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

This gives exactly the same calculation as what we discussed in Example 6.11.



Example 6.15. Diagnostic Tests:

$$\begin{aligned} P(T_P) &= P(T_P \cap D) + P(T_P \cap D^c) \\ &= P(T_P|D)\,P(D) + P(T_P|D^c)\,P(D^c). \\ &= (1 - p_{TE})p_D + p_{TE}(1 - p_D). \end{aligned}$$

6.16. Bayes' Theorem:

(a) Form 1:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad P(A|B) = \frac{P(A \cap B)}{P(B)}$$

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

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)}.$$

(b) Form 2: If a (finite or infinitely) countable collection of events $\{B_1, B_2, \ldots\}$ is a partition of Ω, then

$$P(B_k|A) = P(A|B_k)\frac{P(B_k)}{P(A)} = \frac{P(A|B_k)P(B_k)}{\sum_i P(A|B_i)P(B_i)}$$

Total probability theorem.

- Very simple to derive.
- Extremely useful for making inferences about phenomena that cannot be observed directly.
- Sometimes, these inferences are described as "reasoning about causes when we observe effects".

Example 6.17. Disease testing:

$$P(D|T_P) = \frac{P(D \cap T_P)}{P(T_P)} = \frac{P(T_P|D)\,P(D)}{P(T_P)} = \frac{(1-p_{TE})p_D}{(1-p_{TE})p_D + p_{TE}(1-p_D)}$$

6.18. Probability of compound events

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

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

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

When we have many sets intersected in the conditioned part, we often use "," instead of "∩".

Example 6.19. Most people reason as follows to find the probability of getting two aces when two cards are selected at random from an ordinary deck of cards:

(a) The probability of getting an ace on the first card is 4/52.

(b) Given that one ace is gone from the deck, the probability of getting an ace on the second card is 3/51.

(c) The desired probability is therefore

$$\frac{4}{52} \times \frac{3}{51}.$$

[20, p 243]

Example 6.20. In the early 1990s, a leading Swedish tabloid tried to create an uproar with the headline "Your ticket is thrown away!". This was in reference to the popular Swedish TV show "Bingolotto" where people bought lottery tickets and mailed them to the show. The host then, in live broadcast, drew one ticket from a large mailbag and announced a winner. Some observant reporter noticed that the bag contained only a small fraction of the hundreds of thousands tickets that were mailed. Thus the conclusion: Your ticket has most likely been thrown away!

Let us solve this quickly. Just to have some numbers, let us say that there are a total of $N = 100,000$ tickets and that $n = 1,000$ of them are chosen at random to be in the final drawing. If the drawing was from all tickets, your chance to win would be $1/N = 1/100,000$. The way it is actually done, you need to both survive the first drawing to get your ticket into the bag and then get your ticket drawn from the bag. The probability to get your entry into the bag is $n/N = 1,000/100,000$. The conditional probability to be drawn from the bag, given that your entry is in it, is $1/n = 1/1,000$. Multiply to get $1/N = 1/100,000$ once more. There were no riots in the streets. [15, p 22]

6.21. Chain rule of conditional probability [7, p 58]:

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

Example 6.22. Your teacher tells the class there will be a surprise exam next week. On one day, Monday-Friday, you will be told in the morning that an exam is to be given on that day. You quickly realize that the exam will not be given on Friday; if it was, it would not be a surprise because it is the last possible day to get the exam. Thus, Friday is ruled out, which leaves Monday-Thursday. But then Thursday is impossible also, now having become the last possible day to get the exam. Thursday is ruled out, but then Wednesday becomes impossible, then Tuesday, then Monday, and you conclude: There is no such thing as a surprise exam! But the teacher decides to give the exam on Tuesday, and come Tuesday morning, you are surprised indeed.

This problem, which is often also formulated in terms of surprise fire drills or surprise executions, is known by many names, for example, the "hangman's paradox" or by serious philosophers as the "prediction paradox." To resolve it, let's treat it as a probability problem. Suppose that the day of the exam is chosen randomly among the five days of the week. Now start a new school week. What is the probability that you get the test on Monday? Obviously 1/5 because this is the probability that Monday is chosen. If the test was not given on Monday. what is the probability that it is given on Tuesday? The probability that Tuesday is chosen

to start with is 1/5, but we are now asking for the conditional probability that the test is given on Tuesday, given that it was not given on Monday. As there are now four days left, this conditional probability is 1/4. Similarly, the conditional probabilities that the test is given on Wednesday, Thursday, and Friday conditioned on that it has not been given thus far are 1/3, 1/2, and 1, respectively.

We could define the "surprise index" each day as the probability that the test is not given. On Monday, the surprise index is therefore 0.8, on Tuesday it has gone down to 0.75, and it continues to go down as the week proceeds with no test given. On Friday, the surprise index is 0, indicating absolute certainty that the test will be given that day. Thus, it is possible to give a surprise test but not in a way so that you are equally surprised each day, and it is never possible to give it so that you are surprised on Friday. [15, p 23–24]

Example 6.23. Today Bayesian analysis is widely employed throughout science and industry. For instance, models employed to determine car insurance rates include a mathematical function describing, per unit of driving time, your personal probability of having zero, one, or more accidents. Consider, for our purposes, a simplified model that places everyone in one of two categories: high risk, which includes drivers who average at least one accident each year, and low risk, which includes drivers who average less than one.

If, when you apply for insurance, you have a driving record that stretches back twenty years without an accident or one that goes back twenty years with thirty-seven accidents, the insurance company can be pretty sure which category to place you in. But if you are a new driver, should you be classified as low risk (a kid who obeys the speed limit and volunteers to be the designated driver) or high risk (a kid who races down Main Street swigging from a half-empty $2 bottle of Boone's Farm apple wine)?

Since the company has no data on you, it might assign you an equal prior probability of being in either group, or it might use what it knows about the general population of new drivers and start you off by guessing that the chances you are a high risk

are, say, 1 in 3. In that case the company would model you as a hybrid–one-third high risk and two-thirds low risk–and charge you one-third the price it charges high-risk drivers plus two-thirds the price it charges low-risk drivers.

Then, after a year of observation, the company can employ the new datum to reevaluate its model, adjust the one-third and two-third proportions it previously assigned, and recalculate what it ought to charge. If you have had no accidents, the proportion of low risk and low price it assigns you will increase; if you have had two accidents, it will decrease. The precise size of the adjustment is given by Bayes's theory. In the same manner the insurance company can periodically adjust its assessments in later years to reflect the fact that you were accident-free or that you twice had an accident while driving the wrong way down a one-way street, holding a cell phone with your left hand and a doughnut with your right. That is why insurance companies can give out "good driver" discounts: the absence of accidents elevates the posterior probability that a driver belongs in a low-risk group. [11, p 111-112]