# Exploring Risk Using Data

Tim Erickson, Epistemological Engineering

Risk is part of our everyday lives, and the lives of our students. How should we do "risk education?" We want to help students reduce "serious" risk: from drinking and driving, smoking, or risky sexual behavior. Alas, "serious" risk is earnest and adult, laden with societal rules. More to the point for this conference, it won't expose them to the interesting mathematics of risk and how ubiquitous risk can be.

Risk is a natural part of any phenomenon with a cost and an uncertain outcome. There are a number of ways to study choices, for example, using cost-benefit analysis. But let's simply look at *gambling* as the prototypical risky activity.

In gambling, there's a winning payoff $W$ with a probability $P$, and a cost for each play $C$. That means that on the average, you will win $WP$ on every play. We call $WP$ the *expected value* of your winnings. If $WP > C$, playing the game is a winning proposition; if $WP < C$, the game will lose you money; and if $WP = C$, we call it a *fair game*. But expected value is a slippery concept; it's not obvious that this "average" win (which never actually happens) is really the right quantity to use in analysis.

But let's move on for now. In gambling, $P$ is usually small and $W$ is positive, that is, it's unlikely that you will win, but you pay the cost $C$ hoping to "get lucky." Many other real-life situations are similar to gambling, but inverted: the low-probability outcome is negative—i.e., to be avoided—and we risk it to get the small benefit $C$ that we receive every time we "play."

We saw an example of this attending a conference in Singapore (ICOTS 5, 1998), where fares are not collected on the public buses; you're supposed to buy a ticket and have it with you. We spoke to a rider who had essentially done the calculation: if you don't buy a ticket, you save the fare—let's call it €1. The fine for not having a ticket is €50. If they check tickets fewer than one ride in 50, it pays to cheat. (Despite Singapore's vaunted law-and-order reputation, the rider had no ticket.)

## Insurance

Another real-life phenomenon that involves risk is *insurance*. We pay a premium $C$; if the unlikely dreaded even occurs with probability $P$ (a fire, an accident, theft, the ship sinks), we're covered for our loss $W$. Students, for the most part, have never dealt with insurance, so an activity focusing on the principles of insurance not only gives students an application for expected value but also insight into part of their likely financial futures. What will it teach them about risk? We'll see.

A task exploring insurance might go something like this:

> A truck loaded with tacks has overturned at the beginning of winter on a frozen highway in North Dakota. For various reasons, it will be impossible to clean up the tacks before Spring, and the highway must remain open. As a consequence, there are many more flat tires on this stretch of road than usual. Your take on the role of an insurance company. You sell insurance to drivers, agreeing to

pay for a tire ($100) is they get a flat. The number of policies you will sell depends on the price you set, according to this formula:

$$N = 2000(1 - C) \text{ where } C \in [0,1]$$

where, as before, $C$ is the cost of a policy. That is, if you give the policies away ($C = 0$), you will "sell" 2000, but if you charge a dollar or more ($C \geq 1$), no one will buy them; and the demand curve is linear in between.

The probability that a vehicle gets a flat is $P = 0.002$. What price should you charge for insurance to maximize your profit?

At this point, the astute and experienced reader will recognize that there is a quadratic relationship between the price and the profit; in fact, if $P < 0.01$, the maximum of that profit curve lies in the range $0 < C < 1$. In particular, given the demand formula above, we can compute the expected income $I$:

$$I = NC = 2000(1 - C)C$$

and because we expect $NP$ flats, the expense $E$:

$$E = 100NP = 100(2000)(1 - C)P$$

which means that the expected profit (yield) $Y$ is

$$\begin{aligned} Y &= I - E \\ &= \left[2000(1 - C)C\right] - \left[100(2000)(1 - C)P\right] \\ &= 2000(1 - C)(C - 100P) \end{aligned}$$

If $P = 0.002$, (an average of one flat per 500 vehicles), this becomes

$$Y = 2000(1 - C)(C - 0.20)\,.$$

This function is quadratic in $C$, opening down. It has zeros at $C = 1$ and $C = 0.2$, so has a maximum halfway between, at C = 0.60. At that price, you sell 800 policies, for an income of $480. You expect only 1.6 flats in 800 policies, a cost of $160—yielding an expected profit of $320.

This is the "official" solution using algebra. To produce this solution, a student has to understand expected value pretty well, and must construct the appropriate income and expense functions. We can give these students interesting follow-up tasks, for example, to explain the two zeros. (The one at C = 0.20 represents the zero-profit point when your expected pay-outs equal your income from premiums; the one at C = 1.0 gives zero profit because you don't sell any policies.)

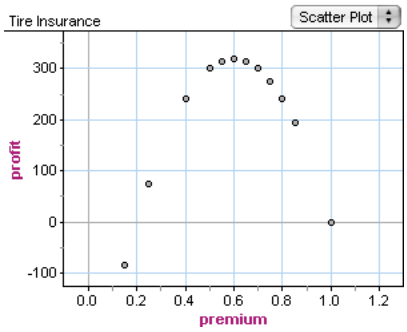
This approach, however, is very "mathematical" and idealized. It's also quite hard. To be sure, it lets students apply the theory of expected value, but it may not prepare them well for real-life decision-making. Let's add some data.

## A Data-Oriented Risk Activity: Floyd's of Fargo

Let's look at an activity we have recently developed as part of the DataGames project (Finzer 2009). This activity is based on a "manual" game of the same name (Erickson 1985).

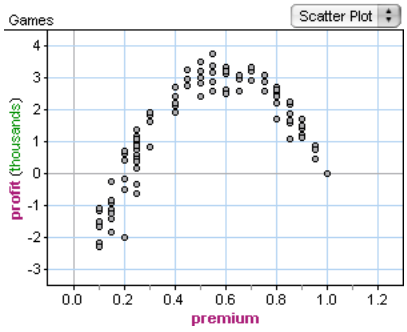
The back-story about the tacks is the same, and so is the model, but instead of knowing all the relevant parameters ahead of time, students play a game on the computer where they set premium prices and see the results of their decisions: how many people bought policies at that price, and how many of those got flats. Their goal is to make as much money as possible in ten turns.

This lets students approach the problem *empirically*. Instead of doing symbol manipulation, they do repeated trials at different prices, and learn from experience that no one buys your policy if you charge too much, and that if you charge too little, you lose money from having to buy too many tires. They can estimate the optimum by inspecting a graph of profit as a function of premium price. In this graph, it looks like the best price is near 60 cents:




Hard-nosed teachers might object that this defeats the purpose of all that lovely math. You could, if you wish, think of the empirical game as an introduction to theoretical techniques, designed to help confused students ground their formulas in some practical understanding. But that's not the only reason to use a game with data instead of simply presenting the classic problem.

For one thing, the game lets us include *variation*. In our initial implementation of this game, the demand model—the number of policies sold—is as described above, but the number of flats is not set deterministically to the expected value as in the previous graph. Instead, it follows the appropriate binomial distribution based on the probability. Here is a typical graph from a simulation:

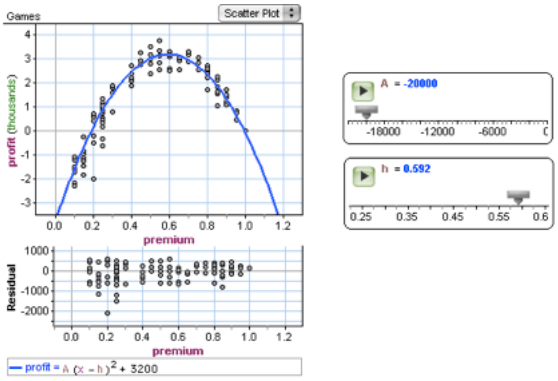



The variation has several important curricular consequences:

- First, variation makes the data seem more realistic even though they are clearly coming from a computer simulation. This may fluster some students who expect all the points to "line up," but it may also make the activity more appealing and less "mathy."
- Second, variation lets us ask deeper questions about the graph. We can ask students to explain what the "stacks" of points mean in the context of the tires and tacks. This is not trivial for many of our students, and gets at the root of understanding what a graph is saying. For example, several statistics students in a field test looked at a graph like the one above and said, "it's a normal distribution." They confused the hump in this bivariate plot with the one in a univariate distribution —an example of not really thinking about what the graph was showing them.
- Next, we can ask what gives rise to the variation; that is, students need to explain and understand that you get different numbers of flats on different runs even with a constant probability.

- Finally, with more sophisticated students, we can ask about the nature of the variation. Does the distribution for a given price seem to follow the expected binomial? Are there outliers? How does the variation vary with the premium price? How does it affect your estimate of the optimum price? How many trials do you need at a given price to get the information you need?

Interestingly, faced with the graph of their data, many field-test students (20 17–18-year-olds in a non-AP statistics class) wanted to fit a quadratic to the data—because it *looked* quadratic, not because of any analysis. They were using *Fathom* (Finzer 2000), so they did so using various techniques, including using variable parameters as described in Erickson (2008). They tended to use vertex form, with a result like the illustration:




When we asked them what their best estimate for the optimum price was, they could read it directly from their function, in this case, $0.592, the number in the "vertex" place in the function. (This is the slider **h** in the illustration.)

These students are a long way, of course, from doing the formal analysis we described above. They have not developed symbolic expressions for income or expenses, or combined them into the quadratic for profit. They have not figured out *why* the relationship is quadratic.

Left to their own devices in our classroom trial, they did not even do what we might do first: they did not estimate the probability of a flat. To prompt this gently, the default state of the game is simply "watching": you watch 2000 vehicles drive the road, and see how many get flats. When, at the end of class, we asked students what the probability was, they estimated it quickly, but the fact that this did not occur to them earlier is interesting. Why didn't they? One possibility is that since they were able to make a model without it—albeit a purely empirical model—they saw no need.

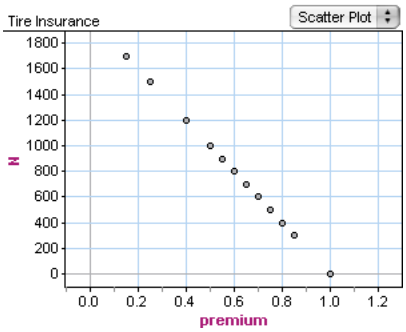
These students had just been learning about expected value, so when we then asked them how much income they would get at a particular price, and how much, on average, they expected to pay out, they did these calculations immediately (using that estimate for $P$). They had not developed the theoretical model, but given their facility with quadratic functions, they were clearly capable of it. So what can we conclude from their approach to the situation?

We can't know the details without further inquiry, but here are some conjectures:

- If you aren't given the probability in a problem, you're less likely to use it. Students are not trained to estimate empirical probabilities to be used later to solve a problem.
- Having to do the observations to estimate a probability is an additional barrier: after all, real observations give varied results. Compare a "book" version: "10,000 vehicles drove through the tack zone and 20 got flats. How much would you charge as an insurance premium if a tire costs $100?" Here the experienced student *knows* that this means $P = 0.002$. Variation doesn't come into it.

- Without variability in the data, data points would line up perfectly, and students would be more certain of the values of the parameters in their models, and more likely to wonder where those values came from. With variation, a model that fits is good enough. Students accept its parameters and don't worry about why the parameters have those values.

As curriculum designers, what should we do? One course of action is to prompt students with tasks that lead to our desired analysis: e.g., have them estimate the probability before they go on to the rest of the activity; or insist that they plot the number of policies sold against premium price (as in the illustration). We could ask other leading questions as well, but should we? Here are some observations:




- Having the tools, skills, and inclination to make a quadratic model, students did so successfully.
- Their empirical model gave them an "answer"—the optimum price—that was as good as any we would get.
- It did not seem to bother them that they had no reason why their quadratic model might make sense.
- The quadratic result we so want students to discover depends on the artificial and deterministic linear demand curve.

We have to ask ourselves how much we care about the theoretical solution. Is it important that students understanding how that modeling works? Will they truly understand the situation or just the algebra? Furthermore, do students benefit from understanding the situation through data even if they never understand the details of the model? It might be enough for many of our students.

## The Individual Perspective

Even if we were to answer those questions to our satisfaction, we would still miss part of our original goal: to help students assess risk, and, ideally, use mathematical understanding to make more informed decisions in risky situations. The problem is that we have taken on the role of the insurance company—a whole-system viewpoint—rather than that of the individual driver.

Suppose we let students be the drivers in our simulation. If they can choose whether they should buy insurance, most of the time they would not get any flats. In fact, in a class of 20 students each playing 10 turns with $P = 0.002$, there is a 2/3 chance that no one will have a flat at all. With that experience, buying insurance seems foolish—to everyone except the rare loser who gets a flat.

From a curriculum design point of view, we have a problem, and as educators of future citizens, we see its importance. How can we help students make good *individual* decisions when they seem to go opposite the best *policy* decisions?

We are still working on this problem, but here is an idea for a direction to try:

## Have Students Invent Stories

In the "policy" perspective, we simulate many experiences so that we can see the aggregate result. The problem is that the individual cases—such as the few people with flat tires—fade into the background, and students are not invested in them. But we could ask students to find individual cases and tell their stories; I predict that students will preferentially pick outliers because they are the interesting cases.

Asking students to tell stories about simulated people may seem odd, but we have seen them do it spontaneously, and have incorporated that idea into other activities (e.g., Turning Numbers into People, Erickson 2000).

This requires two things of the software. First, we need to simulate cases where the personal decision goes both ways, that is, some people buy insurance and others do not. Second, it might help to dress up the data, adding information about the drivers. We could do that with Census data, but it may be enough simply to give them names.

One can imagine doing this with more meaningful problems and choices than buying tack insurance. For example, we hear that texting while driving increases your risk of an accident by a factor of 23 (Box 2009). We can program the simulation to use accident rates of (say) 1:23,000 and 1:1,000 per day. An individual driver is still unlikely to suffer consequences of texting in the short run, but we can simulate 100 drivers for a year, and let half of them text.

That will give us dramatic graphs of aggregate data, but they might not get students to look seriously at consequences of their decisions. If we let students tell a story of someone from the data set, however—Amelia, who has two kids, works as a hospital administrator, and had an accident while texting—the students will identify with them; the former statistic becomes a person.

It would be interesting to see whether a program of data-oriented simulations using these techniques could lead to (or interfere with) a better mathematical understanding of risk and expected value. We also wonder what range of students might be interested in exploring this mathematics and, of course, whether that exploration has any impact on choices they actually make in the course of their lives.

## References


Box, Sherry. *New Data from VTTI Provides Insight into Cell Phone Use and Driving Distraction*. Virginia Tech Transportation Institute. www.vtti.vt.edu.

Erickson, Tim. 2008. "A Pretty Good Fit." *The Mathematics Teacher*. **102(4)**, 256–262.

Erickson, Timothy E. 2000. *Data in Depth: Exploring Mathematics with Fathom*. Emeryville, CA: Key Curriculum Press. ("Turning Numbers into People" begins on p. 4.)

Erickson, Tim. 1985. *Floyd's of Fargo*. Presentation at the Annual Meeting of the California Mathematics Council, Southern Section, Long Beach, California.

Finzer, Bill. 2009. Activities under development for *DataGames*. Project funded by the National Science Foundation.

Finzer, Bill. 2000. *Fathom Dynamic Data Software*. Emeryville, CA: Key Curriculum Press.