Basic Training 4: Risk Analysis

Teacher Workbook

A Program of The Actuarial Foundation

Modeling the Future Challenge

The Modeling The Future Challenge

As part of the Scenario Phase of the MTFC, teams will be demonstrating and applying their mathematical analysis skills to a scenario response paper as well as identifying a potential project and writing a proposal. The Actuarial Process Guide to be an invaluable resource.

- The Actuarial Process Guide

How to Use this MTFC Risk Analysis Scaffolding Guide

When all of the potential topics in the world are at your fingertips, identifying a topic, identifying possible risks, finding sources, mathematically modeling, and studying risks can seem overwhelming to begin. This guide will help scaffold the process and guide participants through the process of risk analysis in the MTFC Project Proposal. Each task refers to a specific section of The Actuarial Process Guide for more in-depth information.

Content: The process is scaffolded into 1 explanatory lesson and 2 total tasks.

Suggested pacing: 1 task per week for 2-3 weeks. Completing the *Basic Training 3. Mathematical Modeling* resource first will provide context to the scenario and scaffolding procedure.

Common Core Standards for Mathematical Practice

The Common Core Standards for Mathematical Practice. The MTFC Project Proposal specifically addresses the following standards:

- ❏ CCSS.MATH.PRACTICE.MP1 **Make sense of problems and persevere in solving them.**
- ❏ CCSS.MATH.PRACTICE.MP2 **Reason abstractly and quantitatively.**
- ❏ CCSS.MATH.PRACTICE.MP3 **Construct viable arguments and critique the reasoning of others.**
- ❏ CCSS.MATH.PRACTICE.MP4 **Model with mathematics.**

4.1 Fundamentals of Characterizing & Quantifying Risk

Risk and Loss

The Oxford English Dictionary defines risk as "a situation involving exposure to danger." However, actuaries and other insurance professionals use the term in a narrower sense. A **risk** is the *uncertain* possibility of something harmful happening. This might involve the loss, theft, or damage of property, or it may involve a person being injured or dying.

A **loss** occurs when an event anticipated as a risk takes place. A risk is a potential for a loss. A loss is the realization of that negative potential. A risk is playing in traffic. A loss is getting hit by a car while doing that. Not all risks result in losses, and not all losses result from risks.

What makes a loss insurable?

1. It is a **pure risk** (or event risk), which refers to an uncertain situation where the opportunity for loss is present but the opportunity for financial gain is absent. Speculative risks refer to uncertain situations that might produce a profit or a loss, such as business ventures, market trading, or gambling transactions. Speculative risks are almost never insurable.

2. The loss must be **accidental**, meaning that it is the result of an unintended action and is unexpected in its exact timing and impact. A landlord burning down his or her own building is not insurable. Losses caused by a divorce are not insurable.

3. One must be able to demonstrate a **definite** proof of loss, and the amount of the loss must be a **measurable** amount. If the existence of a loss cannot be definitively established, or the extent of the loss cannot be calculated, then the loss is not insurable.

4. The loss must be **statistically predictable**, meaning it must be possible to use statistical and mathematical techniques to estimate how frequently a loss might occur and the severity of the loss. Lack of statistical predictability is one reason why earthquakes are generally considered non-insurable events.

5. The loss must be **non-catastrophic**. Nuclear war is not insurable.

6. The number of exposures to any specific event must be **sufficiently large** to allow the insurer to make a reasonable prediction about the loss related to that event. Furthermore, the number of exposures must be **independent** (often this can only be partially accomplished) and encompass a **statistically random** sample of the overall population.

7. The loss must result in **non-trivial economic hardship**. If it does not, then there is no reason to insure against the loss. You cannot insure against breaking your house key in the lock and needing the services of a locksmith.

8. The insurer and the insured must share a **common understanding** of the risk, which is why insurance contracts are so lengthy (which often ironically results in the insured NOT understanding the risk).

This material was adapted from resources created by Alberto L. Dominguez.

Quantifying Risk

In the MTFC, risks must be quantified in some meaningful way. Not all risk has an inherent or measurable value, but assigning a description to the way that risk is being identified and quantified will go hand-in-hand with the assumptions made in your Math Modeling section.

Some values that can be used to quantify an identified risk:

- expected value
- standard deviation & variance
- confidence intervals
- histograms
- projecting trends (possibly linear regression)
- distribution of results

Qualifying reports do not leave risk analysis in general, qualitative forms but instead measure and describe risk in quantitative ways. But how do you actually do this?

1. **Identify Risks:** the first step in your risk analysis section is to identify what risks there may be. It is fine to start off by simply writing these out. Once you have a good list of potential risks, you can then analyze the numbers and quantify the risks.

   For example, consider an example project in the fictional city of Farmland, USA. This city's population is primarily made up of pecan farmers. Imagine in our project we can identify two key things about water access in the coming decades: (1) Farmland, USA will have 20% less annual precipitation by 2050 than it does today, and (2) Farmland, USA will have a 20% increased chance of having a devastating severe summer storm than it does today – increasing from 1.2% chance to 1.44% chance.

   From these two facts, we might be able to identify these risks:
   - Farmland, USA pecan farmers may be at risk of not having enough water to irrigate their crops.
   - Farmland, USA pecan farmers may be at risk of losing crops due to increases in severe storms.

2. **Make and Validate Assumptions**: once you have identified a few potential risks, you may need to make some assumptions about how your data projections will impact the risks you have identified. For example, consider the following assumptions (remember that this is a VERY SIMPLIFIED example, and your real project may need to go into more depth with your assumptions):

   - We assume that the amount of precipitation directly correlates to the amount of water available for pecan farmers in Farmland, USA.
   - We assume that all severe storms calculated in our projections are equally severe an are above a threshold where they would result in a loss of 50% of farm production for the entire city of Farmland, USA.

   Make sure that you can validate your assumptions with some kind of background data or logical reasoning. It doesn't matter so much what assumptions you make, so long as you have sound data and reasoning behind them.

3. **Quantify Risks**: use other numbers to identify the severity and frequency of the risks you calculate. In our example, you may need to identify the price pecans sell for, or the change in output of pecan farms per acre, or other data about the economics of pecan farming. Let's assume the following:
   - Pecan prices are $10/bushel.
   - Farmland, USA includes 10,000 acres of pecan farms.
   - Each acre of pecan farms typically produces 250 bushels during a regular annual growing season.

   With these very basic assumptions and numbers we can make some simple quantifications of the risks:
   - With our assumption of a direct correlation between precipitation and water available for farming, by 2050, the projected trend means that there will be 20% less water available for farmers in Farmland, USA, and they may have to cut their production by 20%. This would result in a potential loss of $5,000,000 a year for the farmers of Farmland, USA.
   - The total annual value of Farmland, USA pecan crops is $25,000,000. With our assumption that the occurrence of a severe storm will result in a 50% loss of pecan crops, a 5% increase in these severe storms (from 1.2% chance to 1.26% chance) means that the expected value for pecan profits for Farmland, USA will go from $24,850,000 today, to $24,820,000 in 2050. Or a loss in expected value of $30,000.

Task 4.1: Identifying and Quantifying Risk

For this task, consider water quality in the USA as a whole (think broader than just the 5 suburbs in that initial scenario). The topic is considerably broad (and more reminiscent of what you will encounter when approaching your MTFC project) and leaves a lot of areas for identifying risk (potential for loss).

- Who or what is at risk when it comes to water quality in the USA?
- With each risk that is identified, how could it be measured, counted, or otherwise quantified?

Who/what is at risk?	How can that risk be quantified? What kind of data?

Task 4.2: Water Quality Scenario Risk Analysis

Continue referencing the water quality scenario and data introduced in Basic Training Resource 2.

Questions	Response
If the acceptable limit of average lead content in a child's blood is between 0 and 11 µg/dL, does the EPA need to take action in any of the suburbs? What are the risks of strictly adhering to these limits? Use data to justify your answer. Which suburbs have an average lead content above this level? What percentage of children in Montgomery are above the EPA's cutoff? Is there a risk in strictly using the 11 µg/dL marker to identify when the EPA should take action in a suburb? Use the data to explain and justify your answer.	Mason: 11.08714 µg/dL $P(A) = P(\text{blood content} > 11) = \frac{\#\text{ children with blood content}>11}{\text{Total \# of children}} = \frac{\sum_{i=1}^{150} I(\text{blood}>11)}{150}$ **= 39.33%** The mean lead content in blood for children in each suburb is: Mariemont: 8.160787 µg/dL Mason: 11.08714 µg/dL Montgomery: 10.97497 µg/dL Sherwood: 8.082522 µg/dL Terrace Park: 7.566373 µg/dL If the maximum acceptable average is 11 µg/dL, then the EPA would only view **Mason** as a problem because its average lead content is above the cutoff. However, this is a dangerous mindset to have for an agency designed to protect the health of citizens from environmental perils. The focus of evaluating lead content in blood should be based on not just the *majority* of citizens, but every citizen. The use of a mathematical mean does not align with this goal because it is based on the central tendency of the data instead of considering any upper outliers, and thus there is a risk of ignoring those who have extremely high lead concentrations in their blood. To exemplify this risk, look at the suburb of Montgomery. While the suburb is below the EPA average lead content cutoff with an average of 10.97497 µg/dL, a quick analysis by finding the probability that an individual child in this suburb is above the EPA cutoff using the logic below reveals that 39.33% of children in Montgomery are above the cutoff! $P(A) = P(\text{blood content} > 11) = \frac{\#\text{ children with blood content}>11}{\text{Total \# of children}} = \frac{\sum_{i=1}^{150} I(\text{blood}>11)}{150}$ As a whole, no, the suburb is not above the EPA cutoff and is therefore considered "safe". Even still though, this method analysis ignores the top end of the distribution and can cause children's health to be overlooked because they are part of a more general "average". Thus, as part of the duties as the EPA, a general summary statistic such as the mean should not be used when deciding whether to intervene in a suburb.

What are some risks of dividing the data up into suburbs? Are there any solutions to these risks?	When choosing how to divide up data into arbitrary categories (such as ones based on political lines), there is the risk that false conclusions can be made based on summary statistics. This is similar to the political process of gerrymandering where the borders of counties (the categories used in voting) are manipulated in an effort to maximize votes for a particular party. The division of categories in this situation isn't done to intentionally show or hide specific results but can have a similar effect. For instance, if one pipeline was the true cause of lead poisoning in citizens across the state overall, but it affected two suburbs equally, then the resulting data would show the two suburbs with moderate averages and medians instead of one "area" reflecting an extremely high average and the other "area" having its true average. There are a few ways to mitigate this risk, although it is difficult to eliminate it altogether without completely removing the category variable altogether. One such solution would to be use smaller subsets of the data that more accurately groups data points based on more commonalities instead of broad groups like suburbs. Another possibility would be to graph the data on a map and visually outline different subsections. While this option is still arbitrary, it still has more of a basis in the actual data than simple city boundaries. Finally, as mentioned in question 11, more elaborate data analysis could be performed on these groups instead of mere summary statistics on each subset.
Is it dangerous to evaluate the situation based on the quantitative analysis of this data alone? What other data could make this analysis better?	There is a whole slew of data that could more accurately predict lead poisoning in children besides just the quality of the water in their household. Variables such as age, number of doctor's visits, age of house, or distance from the Ohio River would all be helpful in analyzing lead poisoning because it serves to paint a more precise picture of the environments each child is living in. Water is not the only source of lead in a child's life, and often times they can be poisoned from another source altogether, so a prediction of low lead poisoning based on low lead content in water is a narrow view of the situation.

After collecting the data, the statisticians realized that the samples were not randomly selected. What kind of problems can this create in our models?	Random selection is vital in data analysis because the mathematical conclusions made is reliant on this key trait. While the violation of an assumption of linear regression, for example, won't necessarily change the outcome, it does vastly increase the probability that the outcome is biased in some way. If every data point does not have an equal chance of being chosen, then our conclusions may falsely favor one result over another. If all of the data points selected draw water from the same well, for example, then the conclusion would place more weight on the water quality of that particular well over other wells in the suburb.
It is discovered that the 20 highest blood concentration samples from the Mason suburb observations were all from the same street, and all the houses were constructed in 1950 with lead-based paint that was proven to increase the lead concentration in blood of a child by 15 µg/dL. If these effects are taken into account, how will our analysis change? Cite specific numbers. EFFECTS ARE INDEPENDENT	The top 20 samples of Mason are data points: 299 154 250 161 197 243 286 202 258 212 169 207 277 292 203 214 184 300 189 185 177. Prior to adjustment, recall that the regression equation was: Expected lead concentration in blood = 0.2574*[Water Content] + 6.7351 The corresponding mean was 11.08714 µg/dL, and the variance was 17.98074 (µg/dL)2. The correlation coefficient, r, equals 0.4578843. After adjusting these 20 data points by subtracting 15 from the response variable, the following regression equation, mean, and variance were obtained: Expected lead concentration in blood = 0.09023*[Water Content] + 8.16187; mean = 9.687143, variance = 8.420913. The correlation coefficient = 0.1023522. It's clear that the adjustment of the data caused a large decrease in the overall correlation between water quality and lead concentration in blood, evidenced both by the reduction in the correlation coefficient from 0.4578843 to 0.1023522 and a reduction in $\beta 1$ from 0.2574 to 0.09023. This naturally lowers the mean as well from 11.09714 to 9.687143 as well as the variance. Overall, all of these changes just goes to show that there are a variety of factors that ultimately affect lead concentration in blood and cannot all be directly attributed to the water quality of a household alone.

How would an insurer and the EPA evaluate this situation differently? Be sure to mention acceptable levels of risk.	The goal of the Environmental Protection Agency is to "protect human health and the environment", as quoted from their mission statement. In this situation, their effective goal is to ensure that the population is not at risk from lead poisoning by evaluating lead concentrations in their blood. This means that their purpose is to eliminate the probability that any member of the population will suffer from lead poisoning and protecting the collective group from any occurrences, essentially eliminating the risk of lead poisoning altogether. This type of thinking differs from the goal of an insurance company, who seeks not only to protect policyholders from adverse financial consequences but also to make a profit through earning more in premiums and investment than they pay out in claims. To this effect, an insurance company would not innately want to eliminate the risk of lead poisoning in a population but would instead seek to minimize the difference between the expected loss of a policyholder and their actual loss. In practice, this would mean that an insurance company may also seek out to reduce the risk of lead poisoning in an effort to decrease the range wherein its true probability exists. This would help the insurance company better price their products and predict the true cost of their insurance product. Thus, while an insurance company may like for no lead poisoning to occur, their optimal risk is simply at a minimum level where they can accurately predict the occurrence of lead poisoning.