Statistical Risk Analysis
Chapter 1: Probability Theory and Statistics

Shaobo Jin

Department of Mathematics

What is Risk

Risk is a quantity derived from

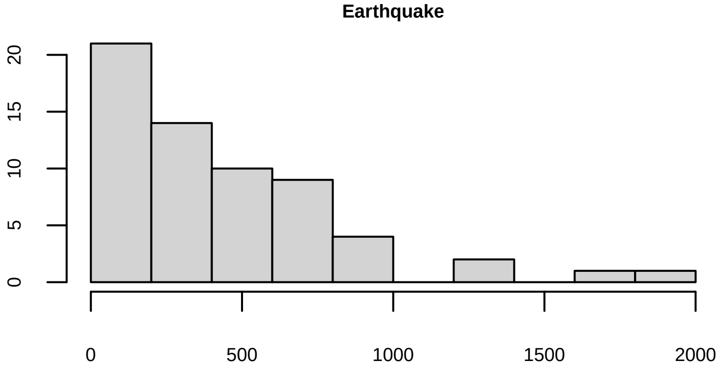
1. the probability that a particular hazard will occur,
2. and the magnitude of the consequence of the undesirable effects of that hazard.

Risk analysis is an approach to analyzing risk. We need to

1. identify failure or damage scenarios
2. state chances for these scenarios and their consequences.

Example: Earthquakes

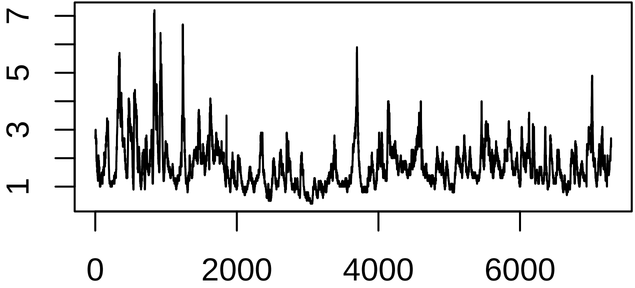
The time intervals in days between 63 successive serious earthquakes world-wide have been recorded.



- How often can we expect a time period longer than 5 years?
- How many earthquakes can happen during a certain period of time, e.g. 1 year?

Example: Significant Wave Height

We have recoreded the wave height near a station situated in Pacific.



- We need to determine the so-called 100-year significant wave: a level that height will exceed on average only once over 100 years.
- We need to estimate durations of storms (time periods with high values) and calm periods to arrange transportation of large cargos.

Rules for Probability

We will use rules for probability a lot. From previous courses,

1. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.
2. Let A_1, A_2, ... be an infinite sequence of events (countable sequence) such that at most one of them can be true, then
$$\mathrm{P}(\text{at least one of } A_i \text{ is true}) = \mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$
3. A collection of events A_1, A_2, ..., A_n is called a partition of the sample space $\mathcal{S}$, if at most one of the events can be true, and the collection is exhaustive, i.e., $\bigcup_{i=1}^{n} A_i = \mathcal{S}$.
4. If A_1, A_2, ..., A_n is a partition of $\mathcal{S}$, then
$$\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathrm{P}(A_i) = 1.$$

Conditional Probability

Definition (Definition 1.4)

The conditional probability of event B given event A such that $\mathrm{P}(A) > 0$ is defined as

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}.$$

Theorem (Theorem 1.1, Law of total probability)

Let A_1, A_2, ..., A_n be a partition of $\mathcal{S}$. Then for any event B,

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) P(A_i).$$

Example Electrical Power Supply

We are interested in the risk of failure of an electric power supply in a house. Define the following events

- $A_1 = \{\text{A day with thunder storm}\}$
- $A_2 = \{\text{A day with blizzard}\}$
- $A_3 = \{\text{A day with other weather}\}$
- $B = \{\text{Error in electricity supply during a day}\}$

We know that errors in supply occur on average once per 10 thunder storms, once per 5 blizzards, and once per 100 days without any particular weather related reasons. We have on average 20 days with thunder storms and 2 days with blizzards during a year. Find $\mathrm{P}(B)$.

Role of Conditional Probability

For two events,

$$\mathrm{P}\left(A_1 \cap A_2\right) = \mathrm{P}\left(A_2 \mid A_1\right) \mathrm{P}\left(A_1\right).$$

For three events,

$$\mathrm{P}\left(A_1 \cap A_2 \cap A_3\right) = \mathrm{P}\left(A_3 \mid A_1 \cap A_2\right) \mathrm{P}\left(A_2 \mid A_1\right) \mathrm{P}\left(A_1\right).$$

In general

$$\mathrm{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathrm{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) \cdots \mathrm{P}\left(A_2 \mid A_1\right) \mathrm{P}\left(A_1\right).$$

Independent Events

If the events A_1, A_2, ..., A_n are independent, then

$$P\left(A_k \mid \bigcap_{i \neq k} A_i\right) = P\left(A_k\right), \text{ for any } k.$$

and

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

Bayes' Formula

Theorem (Theorem 2.1, Bayes' Formula)

Let A_1, A_2, ..., A_k *be a partition of* $\mathcal{S}$, *and* B *an event with* $P(B) > 0$. *Then,*

$$P(A_i \mid B) \quad = \quad \frac{P(A_i \cap B)}{P(B)} = \frac{P(B \mid A_i)\, P(A_i)}{P(B)},$$

where $P(B) = \sum_{j=1}^{n} P(B \mid A_j)\, P(A_j)$.

The function $L(A_i) = \mathrm{P}(B \mid A_i)$ is called the likelihood and measures how likely the observed event is under the alternative A_i.

Random Variable

For a random variable X,

$$F_X(x) = \mathrm{P}(X \le x), \qquad x \in \mathbb{R},$$

is the cumulative distribution function (cdf) or distribution function.

1. If X takes a finite or countable number of values, then the probability mass function (pmf) is $p_k = \mathrm{P}(X = k)$.
   - The pmf uniquely defines the cdf: $F(x) = \sum_{k \le x} \mathrm{P}(X = k)$.
2. If $F(x)$ is differentiable, then the derivative

$$f(x) = \frac{dF(x)}{dx}$$

is called the probability density function (pdf). Then,

$$\mathrm{P}(a < X \le b) = \int_a^b f(x)\,dx.$$

Examples: Probability Density Function

1. Exponential distribution:

$$f(x) = \frac{1}{a}\exp\left(-\frac{x}{a}\right), \quad x \geq 0.$$

2. Weibull distribution:

$$f(x) = cx^{c-1}\exp\left(-x^{c}\right), \quad x \geq 0.$$

3. Standard normal distribution:

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty.$$

It is denoted by $X \in N(0,1)$. Its cdf is often denoted by $\Phi(x) = P(X \leq x)$.

Quantiles

Definition (Definition 3.2)

The quantile x_α for a random variable X satifies $\mathrm{P}\left(X \leq x_\alpha\right) = 1 - \alpha$.

1. The median $x_{0.5}$ of a random variable X is the value such that

$$\mathrm{P}\left(X \leq x_{0.5}\right) = 0.5$$

2. The quartiles $x_{0.25}$ and $x_{0.75}$ are the values such that

$$\mathrm{P}\left(X \leq x_{0.25}\right) = 1 - 0.25, \qquad \mathrm{P}\left(X \leq x_{0.75}\right) = 1 - 0.75.$$

3. Some quantiles that we will often use:
   - λ_α: $\Phi\left(\lambda_\alpha\right) = 1 - \alpha$
   - $\chi^2_\alpha\left(f\right)$: the α quantile of a χ^2 distribution with f degrees of freedom.

Independent Random Variables

Definition (Definition 3.4, Independent random variables)

The variables X_1 and X_2 with distributions $F_1(x)$ and $F_2(x)$, respectively, are independent, if for all values x_1 and x_2, we have

$$P(X_1 \leq x_1, X_2 \leq x_2) = F_1(x_1) F_2(x_2).$$

The function $F_{X_1,X_2}(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$ is called the distribution function for a pair of random variables.

Let $F_i(x)$ be the cdf of random variable X_i. The variables X_1, ..., X_k are independent if

$$P(X_1 \leq x_1, ..., X_k \leq x_k) = \prod_{i=1}^{k} F_i(x_i).$$

Expectations and Variances

Some rules you can use, for constants a and b,

$$\begin{aligned} E\left[aX+b\right] &= aE\left[X\right]+b, \\ E\left[\sum_{i=1}^{n} a_i X_i + b\right] &= \sum_{i=1}^{n} a_i E\left[X_i\right] + b, \\ V\left[aX+b\right] &= a^2 V\left[X\right], \\ V\left[\sum_{i=1}^{n} a_i X_i + b\right] &= \sum_{i=1}^{n} a_i^2 V\left[X_i\right]. \end{aligned}$$

If $E\left[X_i\right]=m$ and $V\left[X_i\right]=\sigma^2$ for all i, then

$$E\left[\bar{X}\right]=m, \qquad V\left(\bar{X}\right)=\frac{\sigma^2}{n}.$$

Standard Deviation

- The standard deviation is $D[X] = \sqrt{V[X]}$.
- For X with positive expectation, the coefficient of variation is defined as

$$R[X] = \frac{D[X]}{E[X]}.$$

  - The influence of units in which X is measured is removed from $R[X]$.

Maximum Likelihood

- Suppose that we have n independent observations x_1, ..., x_n from some distribution with probability density $f(x;\theta)$ or probability mass $p(x;\theta)$. Our goal is to estimate θ.
- Maximum likelihood (ML) focuses on the likelihood function $L(\theta)$, defined as

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} f(x_i;\theta) & \text{continuous random variable} \\ \prod_{i=1}^{n} p(x_i;\theta) & \text{discrete random variable} \end{cases}$$

- The value of θ that maximizes $L(\theta)$ is called the ML estimate. The estimator is called the ML estimator. We often write MLE.

Examples of MLE

1. Suppose that

$$f(x;\theta) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right), \quad x > 0.$$

   Find the MLE of θ.

2. Suppose that

$$p(x;\theta) = \frac{\theta^x}{x!}\exp(-\theta); \quad x = 0, 1, 2, ...$$

   Find the MLE of θ.

Summary of MLE

Distribution	MLE
$\text{Po}(\theta)$	$\bar{X}$
$\text{Bin}(n, \theta)$	X/n
$\text{Exp}(\theta)$	$\bar{X}$
$N(m, \sigma^2)$	$(\bar{X}, S_n^2)$

Bias

Let Θ^* be an estimator of θ. The bias of an estimator is

$$E\left[\Theta^* - \theta\right]$$

If such expectation is 0, the estimator is unbiased. Otherwise, it is biased.

Unbiased Estimator

Are the estimators in the previous examples unbiased for θ?

Efficiency

Unbiased estimators are not unique. To estimate the population mean, $\bar{X}$ is unbiased, and

$$\frac{X_1 + 2X_2 + 3X_3}{6}$$

is also unbiased.

- Among unbiased estimators, we want the variance of the estimator is as small as possible.
- For two unbiased estimators Θ_1^* and Θ_2^*, Θ_1^* is more efficient than Θ_2^* if $V\left(\Theta_1^*\right) < V\left(\Theta_2^*\right)$.

Efficiency

Suppose that data are independent and identically distributed. Which one of $\bar{X}$ and $\frac{X_1+2X_2+3X_3}{6}$ are more efficient if the sample size is $n > 3$?

Central Limit Theorem

Theorem (Theorem 4.4, Central Limit Theorem)

Let $X = (X_1, X_2, ..., X_n)$ be iid random variables. Assume that the expected value $E(X_i) = m$ and variance $V[X_i] = \sigma^2$ are finite. Then,

$$\bar{X} \in AsN\left(m, \sigma^2/n\right).$$

That is, $\bar{X}$ is approximately $N\left(m, \sigma^2/n\right)$.

CLT tells us that for large n,

$$\frac{\bar{X} - m}{\sigma/\sqrt{n}} \text{ is approximately } N(0, 1).$$

Confidence Interval

A $1-\alpha$ confidence interval satisfies

$$\mathrm{P}\left(e_L \leq \theta - \Theta^* \leq e_U\right) = 1-\alpha.$$

Typically, e_L and e_U are chosen such that

$$\mathrm{P}\left(\varepsilon \leq e_L\right) = \mathrm{P}\left(\varepsilon \geq e_U\right) = \frac{\alpha}{2}.$$

In other words,

$$1-\alpha = P\left(\Theta^* + e_L \leq \theta \leq \Theta^* + e_U\right).$$

Examples of Confidence Interval

1. Let X_1, X_2, ..., X_n be iid from $N\left(\mu, \sigma^2\right)$ where σ^2 is known. Then the confidence interval for μ is

$$\left[\bar{x} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

2. Let X_1, X_2, ..., X_n be iid from $N\left(\mu, \sigma^2\right)$ where σ^2 is unknown. Then the confidence interval for μ is

$$\left[\bar{x} - t_{\alpha/2}\left(n-1\right)\frac{s}{\sqrt{n}}, \quad \bar{x} + t_{\alpha/2}\left(n-1\right)\frac{s}{\sqrt{n}}\right].$$