# Introduction to Loss Data Analytics

A short course authored by the Actuarial Community

29 Dec 2020

# Relevance of Analytics

# Relevance of Insurance

By almost any measure, insurance is a major economy activity.

- On a global level, insurance premiums comprised about 6.3% of the world gross domestic product (GDP) in 2013 (Source: International Insurance Fact Book: 2015).
  - Premiums accounted for 11.2% of GDP in Japan
  - Represented 7.5% of GDP in the United States.
- On a personal level:
  - Almost everyone owning a home has insurance to protect themselves in the event of a fire, hailstorm, or some other calamitous event.
  - Almost every country requires insurance for those driving a car.

# Analytics and Loss Data

- Insurance is big business
  - Because of the size, it is not surprising that these firms employ analytics in the same manner as other large corporations.
  - These areas include (i) sales and marketing, (ii) compensation analysis, (iii) productivity analysis, and (iv) financial forecasting. For example, in sales and marketing
    - Predict customer behavior/needs (target appropriate customers)
    - Anticipate customer reactions to promotions/rate changes
    - Manage acquisition costs (online sales, agent compensation)
- One could introduce analytics from many perspectives; we focus on *loss data*, also known as *insurance claims* or *insurance amounts*

# What is Analytics?

- Insurance is a data-driven industry – analytics is a key to deriving information from data.
  - But what is analytics? Some alternative descriptors:
    - *business intelligence* may focus on processes of collecting data, often through databases and data warehouses
    - *business analytics* utilizes tools and methods for statistical analyses of data
    - *data science* can encompass broader applications in many scientific domains
  - Analytics – the process of using data to make decisions.
    - This process involves gathering data, understanding models of uncertainty, making general inferences, and communicating results.

# Insurance Processes

- How does data arise from an insurer?
- In a "micro" oriented view, we can think specifically about what happens to a contract at various stages of its existence.

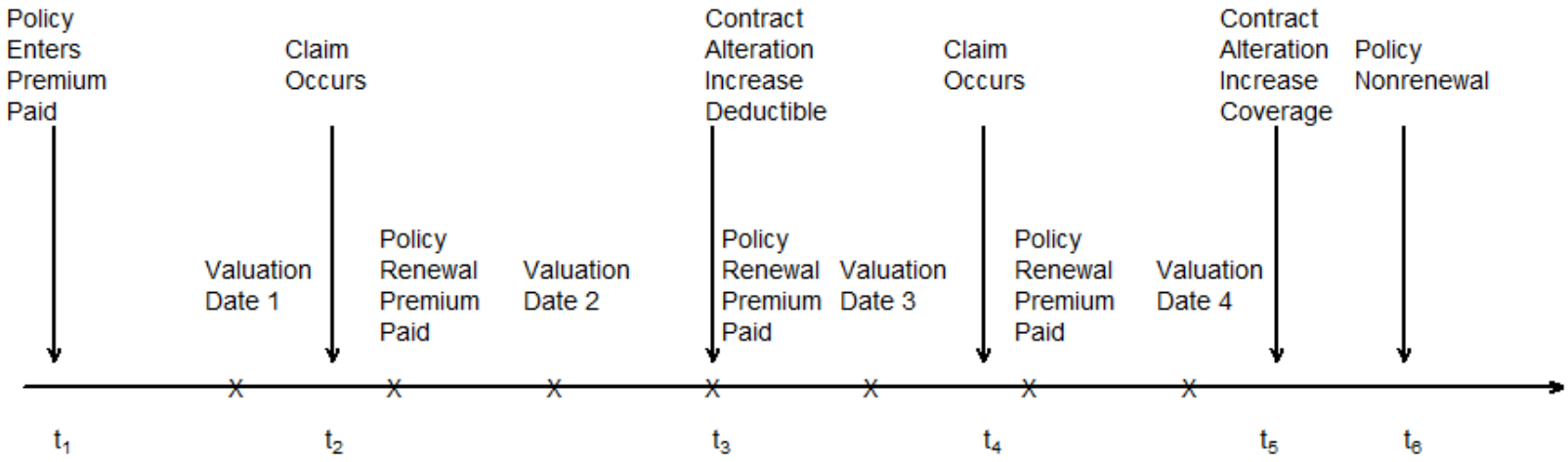



Figure 1: Timeline of a Typical Insurance Policy. Arrows mark the occurrences of random events.

# Relevance of Analytics

In this section, we learned how to:

- Summarize the importance of insurance to consumers and the economy
- Describe analytics
- Identify data generating events associated with the timeline of a typical insurance contract

# Insurance Company Operations

# Insurance Company Operations

*Insurer's Viewpoint:*

✓Need ways of bringing money in, paying it out, managing costs, and making sure that we have enough money to meet obligations.
✓Insurers aggregate detailed insurance processes into larger *operational* units.

- Initiating Insurance
  - Offer right price for the right risk
  - Avoid adverse selection
- Renewing Insurance
  - Retain profitable customers longer
  - Update prices using experience
- Claims and Product Management
- Reserving
- Capital Allocation and Solvency

# Insurance Company Operations

- Initiating Insurance
- Renewing Insurance
- Claims and Product Management
  - Detect and manage claims fraud
  - Manage claims costs
  - Understand excess layers for reinsurance and retention
- Reserving
  - Predict future obligations
  - Quantify the uncertainty of the estimates
  - Match projections of obligations to income streams
- Capital Allocation and Solvency
  - Decide appropriate level of necessary capital
  - Manage external stakeholders' expectations; regulators, rating agencies, reputation

# Operations – Initiating Insurance

- Setting the price of an insurance good can be a perplexing problem.
  - In manufacturing, the cost of a good is (relatively) known
  - In other areas of financial services, market prices are available
  - In many lines of insurance, start with an expected cost, add "margins" to account for the product's riskiness, expenses incurred in servicing the product, and a profit/surplus allowance for the insurance company.

# Operations – Initiating Insurance

- Setting the price of an insurance good can be a perplexing problem.
  - In manufacturing, the cost of a good is (relatively) known
  - In other areas of financial services, market prices are available
  - In many lines of insurance, start with an expected cost, add "margins" to account for the product's riskiness, expenses incurred in servicing the product, and a profit/surplus allowance for the insurance company.
- Especially in automobile and homeowners insurance, analytics sharpens the market by making the calculation of the good's expectation more precise.
  - Multivariate pricing strategies now routinely involve generalized linear model (GLM) techniques

# Big Data

- Traditionally, insurers used only information reported by policyholders on application forms, combined with selected external sources. As examples:
  - police reports for automobile insurance and
  - medical exam results for life insurance.
- Now, there is interest in collecting more information about policyholders
  - An early example was the use of credit scores by Progressive Insurance for automobile insurance.
  - From an analyst's viewpoint, these additional sources have proven to be significant from hypothesis predictive and economic viewpoints.
- Ethically permissible? - these debates are important.

# Insurance Company Operations

In this section, we learned how to:

- Describe five major operational areas of insurance companies.
- Identify the role of data and analytics opportunities within the pricing area.

# Case Study: Wisconsin Property Fund

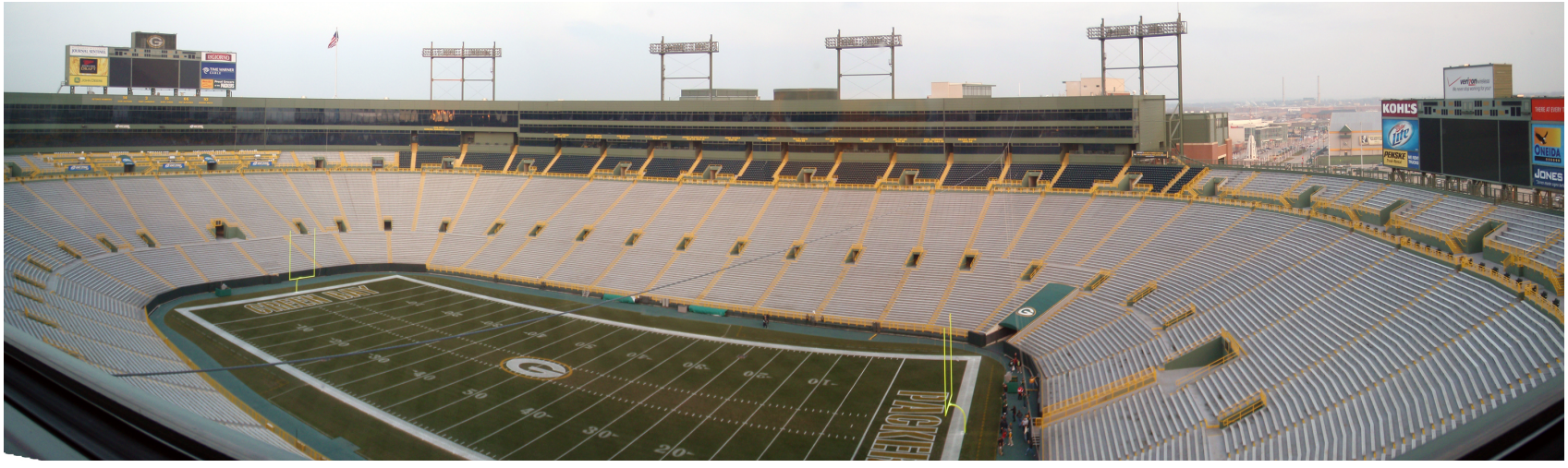
# Wisconsin Property Fund

- The Wisconsin Office of the Insurance Commissioner administers the Local Government Property Insurance Fund (LGPIF).
- Property coverage has been available since 1911.
- The fund insures property such as government buildings, schools, libraries, and motor vehicles.
- Local government entities include counties, cities, towns, villages, school districts, and library boards
  - The fund has over 1,000 such entities.

# LGPIF Policyholder A


- Example – Madison Metropolitan School District
  - it has 98 buildings, 18 major pieces of equipment (mowers, etc.), and 630 properties in the open (benches, playsets, goals, etc.);
  - the property coverage alone is $640 millions.
  - this is Crestwood Elementary School, one of the 98 buildings.

# LGPIF Policyholder B




- The largest contract – the City of Green Bay
  - contains 118 sites,
  - one of which is Lambeau Field – a stadium in which a professional football team, the Green Bay Packers, plays
  - Property coverage is approximately $2.4 billions
  - LGPIF has a separate terrorism reinsurance coverage for this property.

# Property Fund

- The fund receives approximately $25 million in premiums each year and provides insurance coverage for about $75 billion.
- The fund offers three major groups of insurance coverage: building and contents, construction equipment, and motor vehicles.
  - For building and contents, the fund covers all property losses except those resulting from flood, earthquake, wear and tear, extremes in temperature, mold, war, nuclear reactions, and embezzlement or theft by an employee.

# Claims Frequency - R Code

```
Insample <- read.csv("Insample.csv", header=T,
     na.strings=c("."), stringsAsFactors=FALSE)
Insample2010 <- subset(Insample, Year==2010)
table(Insample2010$Freq)
```

# Claims Frequency (2010)

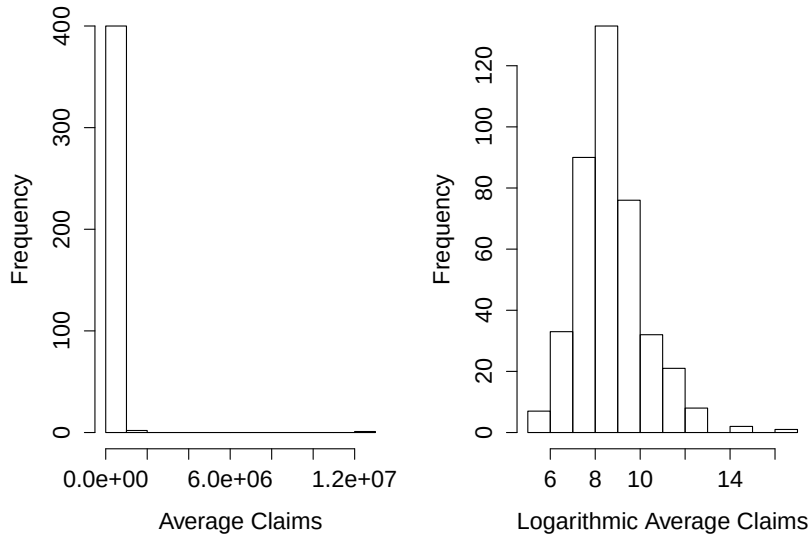
| Type | Number of Claims | | | | | |
|---|---|---|---|---|---|---|
| Number | 0 | 1 | 2 | 3 | 4 | 5 |
| Count | 707 | 209 | 86 | 40 | 18 | 12 |
| Proportion | 0.637 | 0.188 | 0.077 | 0.036 | 0.016 | 0.011 |
| | | | | | | |
| Number | 6 | 7 | 8 | 9 or more | Sum | |
| Count | 9 | 4 | 6 | 19 | 1,110 | |
| Proportion | 0.008 | 0.004 | 0.005 | 0.017 | 1.000 | |

- The table shows 1,110 policyholders who have 1,377 claims.
- Almost two-thirds (0.637) of the policyholders did not have any claims, 18.8% had one claim and remaining 17.5% (=1 - 0.637 - 0.188) had more than one claim.
- The policyholder with the highest number recorded 239 claims.
- The average number of claims for this sample was 1.24 (=1377/1110).

# Severity Distribution (2010)

- 403 (=1110-707) policyholders had at least one claim
- The following summarizes the distribution of the average claims of those policyholders with claims.
- To illustrate, 209 policyholders had only one claim. Here, the claim amount equals the average claim.

| Minimum | First Quartile | Median | Mean | Third Quartile | Maximum |
|---|---|---|---|---|---|
| 167 | 2,226 | 4,951 | 56,330 | 11,900 | 12,920,000 |



# Claims Severity - R Code

```
Insample <- read.csv("Insample.csv", header=T,
     na.strings=c("."), stringsAsFactors=FALSE)
Insample2010 <- subset(Insample, Year==2010)
InsamplePos2010 <- subset(Insample2010, yAvg>0)
# Table
summary(InsamplePos2010$yAvg)
length(InsamplePos2010$yAvg)
# Figures
par(mfrow=c(1, 2))
hist(InsamplePos2010$yAvg,
     main="", xlab="Average Claims")
hist(log(InsamplePos2010$yAvg),
     main="", xlab="Logarithmic Average Claims")
```

# Claim Outcomes and Coverage by Year

- Average frequency is more stable than severity over the years
- Coverage is stable and increasing
- Number of policyholders is stable but declining

| Year | Average Frequency | Average Severity | Average Coverage | Number of Policyholders |
|---|---|---|---|---|
| 2006 | 0.951 | 9,695 | 32,498,186 | 1,154 |
| 2007 | 1.167 | 6,544 | 35,275,949 | 1,138 |
| 2008 | 0.974 | 5,311 | 37,267,485 | 1,125 |
| 2009 | 1.219 | 4,572 | 40,355,382 | 1,112 |
| 2010 | 1.241 | 20,452 | 41,242,070 | 1,110 |

# Analysis by Year - R Code

R documentation for the doBy package available at https://www.rdocumentation.org/packages/doBy/versions/1.5/topics/summaryBy

```
library(doBy)
FUN1 <- function(x) { c(m = mean(x), num =length(x)) }
T1A <- summaryBy(Freq ~ Year,  data=Insample, FUN = FUN1)
T1B <- summaryBy(yAvg ~ Year,  data=Insample, FUN = FUN1)
T1C <- summaryBy(BCcov ~ Year, data=Insample, FUN = FUN1)
Table1In <- cbind(T1A[1],T1A[2],T1B[2],T1C[2],T1A[3])
names(Table1In) <- c("Year", "Average Frequency",
                     "Average Severity", "Average",
                     "Number of Policyholders")
Table1In
```

# Claim Frequency and Severity, Deductibles, and Coverage

- The two outcomes variables are frequency and severity. Each has many zeros (more than half)
- Two other variables are deductible and coverage
- For each of the four distributions, Mean $>$ Median, suggesting skewed distributions

| | Minimum | Median | Mean | Maximum |
|---|---|---|---|---|
| Claim Frequency | 0 | 0 | 1.109 | 263 |
| Claim Severity | 0 | 0 | 9,292 | 12,922,218 |
| Deductible | 500 | 1,000 | 3,365 | 100,000 |
| Coverage (000's) | 8.937 | 11,354 | 37,281 | 2,444,797 |

# Claim Frequency and Severity, Deductibles, and Coverage - R Code

```
FUN2 <- function(x) {c(minx=min(x), medx=median(x),m=mean(x),maxx=max(x))
Insample$BCcov.1000 <- Insample$BCcov/1000
t1 <- as.numeric(summaryBy(Freq ~ 1,        data=Insample, FUN = FUN2 ))
t2 <- as.numeric(summaryBy(yAvg ~ 1,        data=Insample, FUN = FUN2 ))
t3 <- as.numeric(summaryBy(Deduct ~ 1,      data=Insample, FUN = FUN2 ))
t4 <- as.numeric(summaryBy(BCcov.1000 ~ 1, data=Insample, FUN = FUN2 ))
Table2 <- rbind(t1,t2,t3,t4)
colnames(Table2) <- c("Minimum", "Median","Average", "Maximum")
rownames(Table2) <- c("Claim Frequency","Claim Severity",
                      "Deductible","Coverage (000's)")
Table2
```

# Cost of Insurance

- Because coverage cannot be denied, underwriting not a major issue
- How much to charge?
- Based on 2010 data, might use 33,026.

$$= \frac{\text{total fund claims}}{\text{number of policyholders}} = \frac{\text{36.66 million USD}}{1110}$$

- However, very different answer based on 2009 data (9,934).

# Description of Rating Variables

- May wish to vary rates according to these characteristics

| Variable | Description |
|---|---|
| `EntityType` | Categorical variable that is one of six types: (Village, City, County, Misc, School, or Town) |
| `LnCoverage` | Total building and content coverage, in logarithmic millions of dollars |
| `LnDeduct` | Deductible, in logarithmic dollars |
| `AlarmCredit` | Categorical variable that is one of four types: (0%, 5%, 10%, or 15%), for automatic smoke alarms in main rooms |
| `NoClaimCredit` | Binary variable to indicate no claims in the past two years |
| `Fire5` | Binary variable to indicate the fire class is below 5 (The range of fire class is $0 \sim 10$) |

# Claims by Entity Type, Fire Class, and No Claim Credit

- There is substantial variation in the frequency and severity by entity type.
- As anticipated, lower frequency and severity when the policyholder had no claims in the past two years, (`NoClaimCredit=1`).
- Higher frequency and severity for the `Fire5` (=1) variable.
  - **Counter-intuitive**: one would expect lower claim amounts for those policyholders in areas with better public protection (when the protection code is five or less).

| Variable | Number of Policies | Claim Frequency | Average Severity |
|---|---|---|---|
| `EntityType` | | | |
| Village | 1,341 | 0.452 | 10,645 |
| City | 793 | 1.941 | 16,924 |
| County | 328 | 4.899 | 15,453 |
| Misc | 609 | 0.186 | 43,036 |
| School | 1,597 | 1.434 | 64,346 |
| Town | 971 | 0.103 | 19,831 |
| `Fire5=0` | 2,508 | 0.502 | 13,935 |
| `Fire5=1` | 3,131 | 1.596 | 41,421 |
| `NoClaimCredit=0` | 3,786 | 1.501 | 31,365 |
| `NoClaimCredit=1` | 1,853 | 0.310 | 30,499 |
| Total | 5,639 | 1.109 | 31,206 |

# Claim Frequency and Severity, Deductibles, and Coverage - R Code

```
ByVarSumm<-function(datasub){
  tempA <- as.numeric(summaryBy(Freq ~ 1 , data = datasub,
           FUN = function(x) { c( numx=length(x), mx = mean(x))} ))
  datasub1 <-  subset(datasub, yAvg>0)
  tempB <- as.numeric(summaryBy(yAvg ~ 1, data = datasub1,
           FUN = function(x) { c(mxx = mean(x)) } ))
  tempC <- c(tempA,tempB)
  return(tempC)
  }
t1 <- ByVarSumm(subset(Insample, TypeVillage == 1))
t2 <- ByVarSumm(subset(Insample, TypeCity == 1))
t3 <- ByVarSumm(subset(Insample, TypeCounty == 1))
t4 <- ByVarSumm(subset(Insample, TypeMisc == 1))
t5 <- ByVarSumm(subset(Insample, TypeSchool == 1))
t6 <- ByVarSumm(subset(Insample, TypeTown == 1))
t7 <- ByVarSumm(subset(Insample, Fire5 == 0))
t8 <- ByVarSumm(subset(Insample, Fire5 == 1))
t9 <- ByVarSumm(subset(Insample, Insample$NoClaimCredit == 0))
t10 <- ByVarSumm(subset(Insample, Insample$NoClaimCredit == 1))
t11 <- ByVarSumm(Insample)

Tablea <- rbind(t1,t2,t3,t4,t5,t6,t7,t8,t9,t10,t11)
Table4 <- round(Tablea, digits = 3)
colnames(Table4) <- c("Number of Policies", "Claim Frequency", "Average Severity")
rownames(Table4) <- c("Village","City","County","Misc","School",
          "Town","Fire5--No","Fire5--Yes","NoClaimCredit--No",
        "NoClaimCredit--Yes","Total")
Table4
```

# Claims by Entity Type and Alarm Credit Category

- **Counter-intuitive** results for Alarm Credit. Would expect lower frequency/severity for 15% alarm credits.

| Entity Type | No Alarm Credit Claim Frequency | No Alarm Credit Avg. Severity | No Alarm Credit Num. Policies | Alarm Credit 5% Claim Frequency | Alarm Credit 5% Avg. Severity | Alarm Credit 5% Num. Policies |
|---|---|---|---|---|---|---|
| Village | 0.326 | 11,078 | 829 | 0.278 | 8,086 | 54 |
| City | 0.893 | 7,576 | 244 | 2.077 | 4,150 | 13 |
| County | 2.140 | 16,013 | 50 | - | - | 1 |
| Misc | 0.117 | 15,122 | 386 | 0.278 | 13,064 | 18 |
| School | 0.422 | 25,523 | 294 | 0.410 | 14,575 | 122 |
| Town | 0.083 | 25,257 | 808 | 0.194 | 3,937 | 31 |
| Total | 0.318 | 15,118 | 2,611 | 0.431 | 10,762 | 239 |

| Entity Type | Alarm Credit 10% Claim Frequency | Alarm Credit 10% Avg. Severity | Alarm Credit 10% Num. Policies | Alarm Credit 15% Claim Frequency | Alarm Credit 15% Avg. Severity | Alarm Credit 15% Num. Policies |
|---|---|---|---|---|---|---|
| Village | 0.500 | 8,792 | 50 | 0.725 | 10,544 | 408 |
| City | 1.258 | 8,625 | 31 | 2.485 | 20,470 | 505 |
| County | 2.125 | 11,688 | 8 | 5.513 | 15,476 | 269 |
| Misc | 0.077 | 3,923 | 26 | 0.341 | 87,021 | 179 |
| School | 0.488 | 11,597 | 168 | 2.008 | 85,140 | 1,013 |
| Town | 0.091 | 2,338 | 44 | 0.261 | 9,490 | 88 |
| Total | 0.517 | 10,194 | 327 | 2.093 | 41,458 | 2,462 |

# Initiating Insurance

- How much to charge?
  - Based on 2010 data, might use 33,026.
  - However, very different answer based on 2009 data (9,934).
- Single premium for all policyholders does not seem fair.
- Outcomes seem to vary by entity type
- Charge more for those with greater amounts of coverage
- What about alarm credits???

# REVIEW

In this section, we learned how to:

- Describe how insurance events can produce data of interest to analysts.
- Produce relevant summary statistics for each variable.
- Describe how these summary statistics can be used to develop the cost of insurance.