Raising Your Actuarial IQ

(Improving Information Quality)

CAS Data Management Educational
Materials Working Party
with Martin E. Ellingsworth

Actuarial IQ Introduction

- "IQ" stands for "**Information Quality**"
- Introduction to Data Quality and Data Management being written by the *CAS Data Management Educational Materials Working Party*
- Directed at actuarial analysts as much as actuarial data managers:
  - what every actuary should know about data quality and data management

---

Working Party Publications

- **Book reviews** of data management and data quality texts in the *Actuarial Review* starting with the August 2006 edition
- These reviews are combined and compared in "**Survey** of Data Management and Data Quality Texts," CAS *Forum,* Winter 2007, www.casact.org
- **This presentation** is based on our
- Upcoming **paper**: "**Actuarial IQ** (Information Quality)" to be published in the Winter 2008 edition of the CAS *Forum*

What is Data Quality?

- Quality data is data that is **appropriate** for its purpose.
- Quality is a **relative** not absolute concept.
  - Data for an annual rate study may not be appropriate for a class relativity analysis.
  - Promising predictor variables in Predictive Modeling may not have been coded or processed with that purpose in mind.

Introduction and Horror Stories

- Presented by Aleksey Popelyukhin

Data Flow

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

Information Quality involves **all** steps:
- Data Requirements
- Data Collection
- Transformations & Aggregations
- Actuarial Analysis
- Presentation of Results

To improve Final Step:
- Making Decisions

Principles on Data Quality: Perspectives

Step 0 Data Requirements	

ASB – ASOP 23 – *"Data Quality"*

CAS Management Data and Information Committee: *"White Paper on Data Quality"*

Richard T. Watson *"Data Management: Databases and Organization"*

ASOP No. 23

Due consideration to the following:

- **Appropriateness** for intended purpose …
- **Reasonableness** and **comprehensiveness** …
- Any known, material **limitations** …
- The cost and feasibility of obtaining **alternative data** …
- **The benefit** to be gained from an alternative data set …
- **Sampling methods** …

White Paper on Data Quality

Evaluating data quality consists of examining data for:

- **Validity**
- **Accuracy**
- **Reasonableness**
- **Completeness**

(Sidebar steps repeated in each slide: Step 0 Data Requirements, Step 1 Data Collection, Step 2 Transformations Aggregations, Step 3 Analysis, Step 4 Presentation of Results, Final Step Decisions)

Watson

18 Dimensions of Data Quality:

- Many overlap with previously mentioned principles.
- Others describe ways of storing data
  e.g. Representational consistency, Precision
- Others go beyond data characteristics to processing and management
  e.g. Stewardship, Sharing, Timeliness, Interpretation

Diagram steps (left column): Step 0 Data Requirements, Step 1 Data Collection, Step 2 Transformations Aggregations, Step 3 Analysis, Step 4 Presentation of Results, Final Step Decisions

Data Requirements

Redman: "Manage Information Chain"

- establish management *responsibilities*
- describe information *chart*
- understand customer *needs*
- establish *measurement* system
- establish *control* and check performance
- identify *improvement opportunities*
- make improvements

Diagram steps (left column): Step 0 Data Requirements, Step 1 Data Collection, Step 2 Transformations Aggregations, Step 3 Analysis, Step 4 Presentation of Results, Final Step Decisions

Data Requirements

Data Quality Measurement:

- Quantify traditional aspects of quality data such as **accuracy**, **consistency**, **uniqueness**, **timeliness** and **completeness** using a score assigned by an expert
- Measure the consequences of data quality problems
  - measure the **number of times** in a sample that data quality errors cause errors in analyses, and
  - the **severity** of those errors
- Use measurement to motivate improvement

Diagram steps (left column): Step 0 Data Requirements, Step 1 Data Collection, Step 2 Transformations Aggregations, Step 3 Analysis, Step 4 Presentation of Results, Final Step Decisions

Slide 1

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

Metadata

Big help in describing Data Requirements – **Metadata!**

- Data that Describes the Data
  - Key Data Management Tool
  - Reduces Risky Assumptions
    CWP means
    Closed with Payment?
    Closed without Payment?

Slide 2

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

Example – Marital Status

- What is in the Marital Status Variable?

Marital Status

	Frequency	Percent
1	5,053	14.3
	2,043	5.8
2	9,657	27.4
4	2	0
D	4	0
M	2,971	8.4
S	15,554	44.1
Total	35,284	100

Single? → 1

Married? → 2

Polygamist? → 4

Single / Separated? → S

Slide 3

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

Example: What is the Marital Status Variable?

Example of Metadata

Marital Status Value	Description
1	Married, data from source 1, straight move of field ms_code
2	Single, data from source 1, straight move of field ms_code
4	Divorced, data from source 1, straight move of field ms_code
D	Divorced, data from source 2, straight move of mstatus
M	Married, data from source 2, straight move of mstatus
S	Single, data from source 2, straight move of mstatus
Blank	Marital status is missing

Slide 1

| Step 0 **Data Requirements** |
| Step 1 **Data Collection** |
| Step 2 **Transformations Aggregations** |
| Step 3 **Analysis** |
| Step 4 **Presentation of Results** |
| Final Step **Decisions** |

What Is In It?

- Business Rules
- Data Processing Rules
- Report Compilation and Extraction Process
- Other

Slide 2

| Step 0 **Data Requirements** |
| Step 1 **Data Collection** |
| Step 2 **Transformations Aggregations** |
| Step 3 **Analysis** |
| Step 4 **Presentation of Results** |
| Final Step **Decisions** |

What Is In It?

- Business Rules
  - Data Elements
    - Definition of Field, e.g.,
      - How Claims are Defined
      - How Exposure is Calculated
    - Format of Field
      - mm/dd/yyyy
      - #,##0.00
    - Valid Values and Interdependencies
      - Alpha Only
      - Driver = Yes and Age > 15

Slide 3

| Step 0 **Data Requirements** |
| Step 1 **Data Collection** |
| Step 2 **Transformations Aggregations** |
| Step 3 **Analysis** |
| Step 4 **Presentation of Results** |
| Final Step **Decisions** |

What Is In It?

- Data Processing Rules
  - How Database is Populated
  - Sources of Data
  - Handling of Missing Data

What Is In It?

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Report Compilation and Extraction Process
    - How Data is Selected or Bypassed
    - Fiscal Period
    - Accounting Date for Transactions
    - Actuarial Evaluation Date
    - Calculations
    - Mappings

What Is In It?

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Other
    - Process Flow Documentation
    - Versioning

Why Actuaries Need Metadata?

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Better Analysis
- Avoid Being Mis-Informed about Data Variable and What It Represents
- Did Anything Change During the Experience Period?
    - Only if
        - Ask to receive this
        - Actually compare metadata lists / files

Example of Metadata

Step 0
Data Requirements
Step 1 — **Data Collection**
Step 2 — **Transformations Aggregations**
Step 3 — **Analysis**
Step 4 — **Presentation of Results**
Final Step — **Decisions**

- Statistical Plans in P/C Industry
  - General Reporting Requirements
  - Data Element Definitions
  - Standardize Data to the Extent Possible
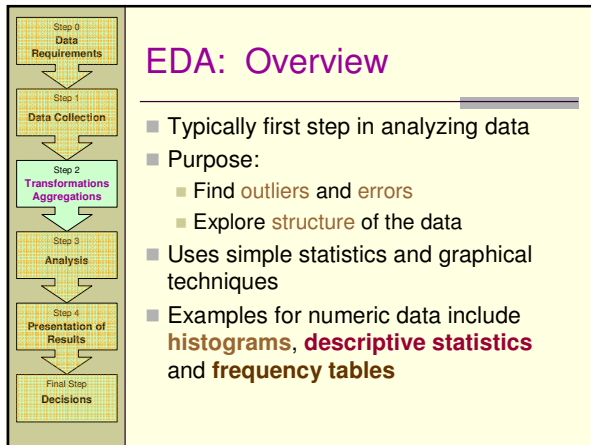
Data Collection

Step 0 — **Data Requirements**
Step 1 — **Data Collection**
Step 2 — **Transformations Aggregations**
Step 3 — **Analysis**
Step 4 — **Presentation of Results**
Final Step — **Decisions**

- Data supplier management
  - Let suppliers know what you want
  - Provide feedback to suppliers
  - Balance the following
    - Known issues with supplier
    - Importance to the business
    - Supplier willingness to experiment together
    - Ease of meeting face to face

Transformations and Aggregations

Step 0 — **Data Requirements**
Step 1 — **Data Collection**
Step 2 — **Transformations Aggregations**
Step 3 — **Analysis**
Step 4 — **Presentation of Results**
Final Step — **Decisions**

- In this step data are put into standardized structures and then combined into larger, more centralized data sets
- "Actuarial IQ" introduces two ways to improve IQ in this step:
  - Exploratory Data Analysis (EDA)
  - Data Audits

EDA: Data Preprocessing

EDA: Overview

- Typically first step in analyzing data
- Purpose:
  - Find outliers and errors
  - Explore structure of the data
- Uses simple statistics and graphical techniques
- Examples for numeric data include **histograms**, **descriptive statistics** and **frequency tables**

EDA: Histograms

Slide 1
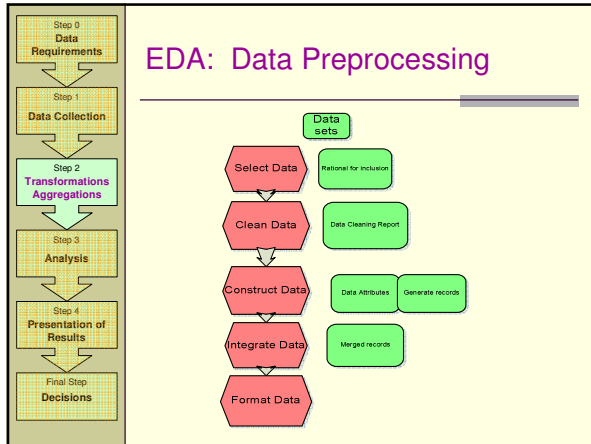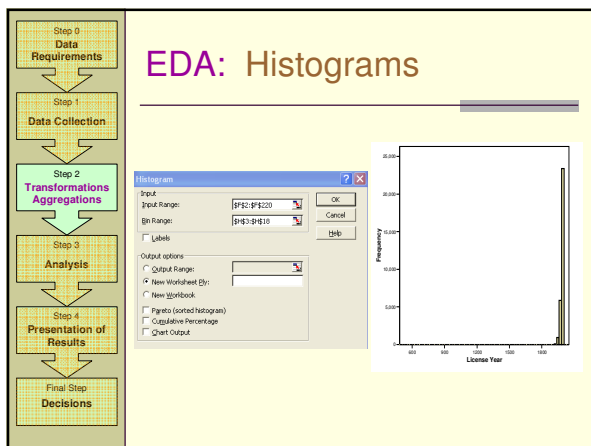
Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Descriptive Statistics

Statistic	Policyholder Age
Mean	36.9
Standard Error	0.1
Median	35.0
Mode	32.0
Standard Deviation	13.2
Sample Variance	174.4
Kurtosis	0.5
Skewness	0.7
Range	84
Minimum	16
Maximum	100
Sum	1114357
Count	30226
Largest(2)	100
Smallest(2)	16

Slide 2
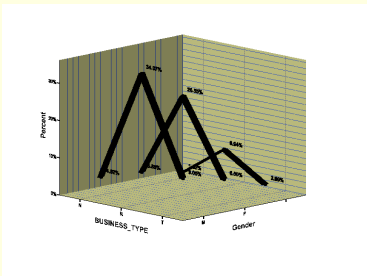
Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Categorical Data

Slide 3

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

EDA: Data Cubes

- Usually frequency tables
  - Example: search for missing gender values

Gender	Frequency	Percent
	5,054	14.3
F	13,032	36.9
M	17,198	48.7
Total	35,284	100

EDA: Data Cubes

Step 0 — Data Requirements
Step 1 — Data Collection
Step 2 — Transformations Aggregations
Step 3 — Analysis
Step 4 — Presentation of Results
Final Step — Decisions

- Example: identify inconsistent coding of marital status

Marital Status

	Frequency	Percent
	5,053	14.3
1	2,043	5.8
2	9,657	27.4
4	2	0
D	4	0
M	2,971	8.4
S	15,554	44.1
Total	35,284	100

Missing Data

Multiple codes for same status

Underutilized data elements?

EDA: Missing Data

Step 0 — Data Requirements
Step 1 — Data Collection
Step 2 — Transformations Aggregations
Step 3 — Analysis
Step 4 — Presentation of Results
Final Step — Decisions

		BUSINESS TYPE	Gender	Age	License Year
N	Valid	35,284	35,284	30,242	30,250
	Missing	0	0	5,042	5,034
Percentiles	25			27.00	1,986.00
	50			35.00	1,996.00
	75			45.00	2,000.00

EDA: Summary

Step 0 — Data Requirements
Step 1 — Data Collection
Step 2 — Transformations Aggregations
Step 3 — Analysis
Step 4 — Presentation of Results
Final Step — Decisions

- Before data is analyzed,
  - Gathered
  - Cleaned
  - Integrated
- EDA Techniques used
  - to explore the data
  - to detect missing values,
  - to identify invalid values and
  - to highlight outliers
- Use histograms, descriptive statistics and frequency tables

　
Data Audits

Step 0 — Data Requirements
Step 1 — Data Collection
Step 2 — Transformations Aggregations
Step 3 — Analysis
Step 4 — Presentation of Results
Final Step — Decisions

- ASOP No. 23 does not require actuaries to audit, but good to understand
- **Main Idea:** compare the data intended for use to its original source, e.g., policy applications or notices of loss
- **Top-Down:** check that totals from one source match the totals from a reliable source (????)
- **Bottom-Up:** follow a sample of input records through all the processing to the final report

Analysis Quality

Step 0 — Data Requirements
Step 1 — Data Collection
Step 2 — Transformations Aggregations
Step 3 — Analysis
Step 4 — Presentation of Results
Final Step — Decisions

Data   **Models**   **Results**

On its way to results data can be:
- **Rejected**
  - wrong Format
- **Underutilized**
  - wrong Model
- **Distorted**
  - wrong model Parameterization

Analysis is a crucial component in the overall process quality

Model Quality

Step 0 — Data Requirements
Step 1 — Data Collection
Step 2 — Transformations Aggregations
Step 3 — Analysis
Step 4 — Presentation of Results
Final Step — Decisions

- Model Design quality
- Implementation quality
- Testing and Documentation

Model Quality

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Model Design quality
  - Model Selection and Validation
  - Parameters Estimation
  - Verification

  Did I use the right model ?
  Did I use the model right ?

  - Model Performance

Model Quality

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Model Performance

Models predict observable events.

Outcomes can be compared to predictions leading to...

- Model's Improvements
- Model's Recalibration
- Model's Rejection

leading to... higher process quality.

Model Quality

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Model Design quality
- Implementation quality
- Testing and Documentation

Model Quality

Step 0 — Data Requirements
Step 1 — Data Collection
Step 2 — Transformations Aggregations
Step 3 — Analysis
Step 4 — Presentation of Results
Final Step — Decisions

- Implementation quality
  - Programming languages: C++, VBA, SQL
    many books on good design patterns
  - Formulae in a Spreadsheet - also programming
    no books on good design patterns
  - Need good software design to simplify:
    - Usage
    - Testing
    - Modifications / Improvements
    - **Recovery** ← *(side benefit)*

Model Quality

Results do not belong to the template either

- Implementation quality
  - Separation of data and algorithms

Data does not belong in template.

Model Quality

Each Step on its own tab

- Implementation quality
  - Layering

Model Quality

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Model Design quality
- Implementation quality
- Testing and Documentation

Model Quality

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Testing and Documentation
  - Validation

  black-box treatment: comparing results with correct ones...
  - Verification

  inside-the-box treatment: checking formulae...

1. Should be **integral** part of development
2. Should be performed **by outsiders**
3. Should be **well-documented**

Model Quality

Database "Documenter"

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

- Testing and Documentation
  - Self-documenting features
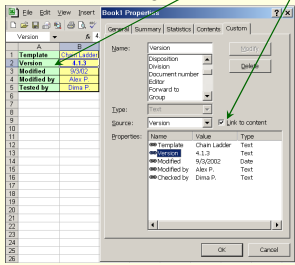
Model Quality

One can link
Document Properties to
Spreadsheet Cells

- Testing and Documentation
  - Version management
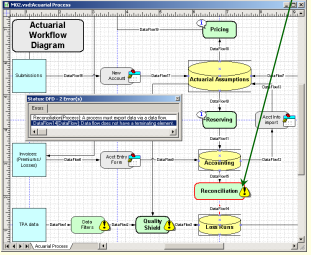
Model Quality

"Smart diagrams"
can be automated

- Testing and Documentation
  - Documenting Workflow

Working Example

Presented by Martin Ellingsworth

PWC 2004 Study

"The key is to understand the impact data is having on your business and do something about it."

"Data quality is at the core – if you improve your data you will directly impact your overall business results."

Global Data Management Survey 2004, PriceWaterhouseCoopers

Conclusion

- Data Quality is a core issue affecting the quality and usefulness of the actuarial work product
- Data Quality is not just about how data is coded: phrase 'information quality' is coined to emphasize the impact of processes on the quality of final product

---

Conclusion

- Ways to improve actuarial IQ
  - Applying Data Quality principles
  - Defining and using Metadata
  - Measuring data quality to track progress and awareness of quality audit
  - Utilizing Exploratory Data Analysis to identify outliers and explore the structure of a dataset
  - Testing the quality of actuarial models
  - Clarifying actuarial presentations and reports
  - Employing Actuarial Data Management best practices

Step 0
Data Requirements

Step 1
Data Collection

Step 2
Transformations Aggregations

Step 3
Analysis

Step 4
Presentation of Results

Final Step
Decisions

Conclusion

- Expansions of actuarial frame of reference
  - Data is a corporate asset that needs to be managed and actuaries can play a role
  - Data needs to be appropriate for all of its intended uses
  - Expansion of data quality principles to support these broader perspectives

Acknowledgement

The working party would like to thank the **Insurance Data Management Association** (www.idma.org) for their help in:

- Developing a shortlist of texts that would be relevant to actuaries, and
- Reviewing our papers

Author, Author…

This presentation is a publication of CAS **Data Management and Information Educational Materials Working Party:**

- Keith P. Allen
- Robert Neil Campbell, *Chairperson*
- Louise A. Francis
- David Dennis Hudson
- Gary W. Knoble
- Rudy A. Palenik
- Aleksey Popelyukhin Ph.D.
- Virginia R. Prevosto
- Lijuan Zhang

CAS Data Management Educational Materials Working Party Publications

- **Book reviews** of data management and data quality texts in the *Actuarial Review* starting with the August 2006 edition
- These reviews are combined and compared in "**Survey** of Data Management and Data Quality Texts," CAS *Forum,* Winter 2007, www.casact.org
- **This presentation** is based on our
- Upcoming **paper**: "**Actuarial IQ** (Information Quality)" to be published in the Winter 2008 edition of the CAS *Forum*

_____

_____

_____

_____

_____

_____

_____

_____