How to Recognize and Fix Data Faults That Can Make or Break Analytics

- Data
- Interpretation
- Sampling

By Cary S. Shaw & Associates, LLC
For further information contact (203)505-3180
caryshaw@optonline.net

Get the Data Quality Right

- Business depends on data.
- Whether you are:
  - Requestor
  - Developer
  - Producer
  - Interpreter
  - Analytics Consultant
  - Analytics Recipient

We will cover:

- Qualify
  - Sniff, Define, Frequency Check
- Interpret
  - Lurk, Miss, Population
- Sample
  - Methods, Application
- Summary
  - Tools, Lessons

How many products are in company's data base?

- Marketing client sent request to IT
- Detail result provided in Excel.
- Marketing client relayed file: Answer: 1,000.

How many products are in company's data base?

- Marketing client sent request to IT
- Detail result provided in Excel.
- Marketing client relayed file: Answer: 1,000.

- What happened: SQL Navigator timed out.

- Real answer: 3,646.

Spreadsheet data received in this form.

Business Name	Street	City	State	Zip	Phone
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx	xxxxx	(xxx)xxx-xxxx
xxxxxxxxxx	xxxxxxxxxx	xxxxx	xx		

Assume data is in error. Search for errant values.

- Frequency tables
- Outliers - Confidence intervals
- Outliers - One variable graphed/estimated against another
- Referential Integrity (cross table connections within an integrated database)
- Sniff test

- Don’t dismiss data if it is different; only if it is wrong.

How unique are our customer phone numbers?

- Are we really reaching the person / business we intend?
- Are we alienating by multiple calls?
- Are we missing multiple clients?
- Practical question: How frequently does the same phone number appear on our customer contact file?

Exploring Dividend Yields

- Found company with 7.2 % dividend yield
- What's going on?

Could a ‘9’ or ‘0’ mean something else?

How do we want to select business vs. home prospects?

- H = Home
- B = Business

- And how do we want to convert this to a numeric factor?

Select B (business) or Reject H (home)? Different impact.

Business, Home	Frequency
(blank)	98,947
B	5,032
b	159
Business	49
Busn	1
H	24,891
h	475
Home	63
Hume	1
Hme	11

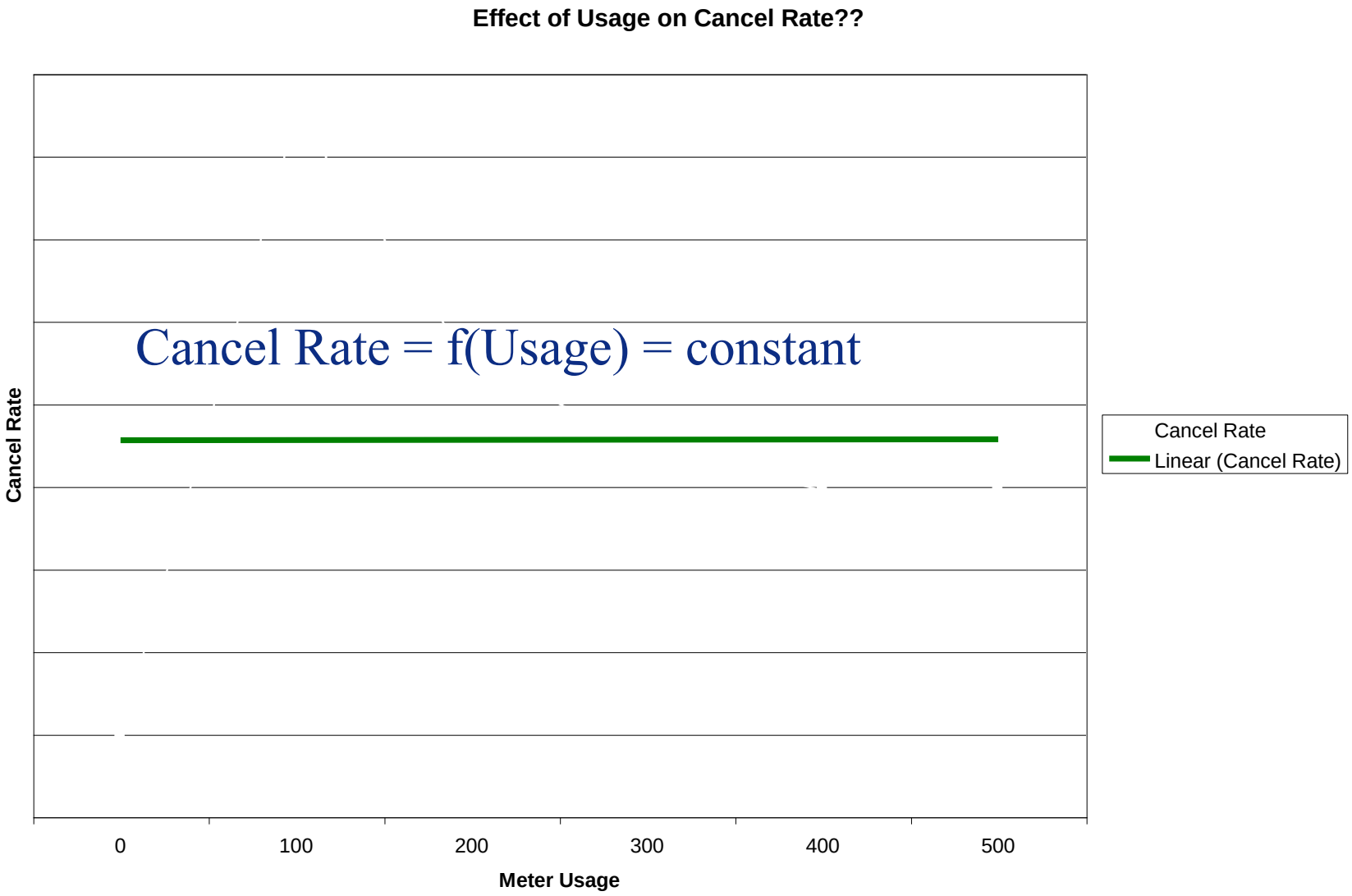
- Other values have impact
- Is field up-to-date?
- Is client (blank) different from proprietary (blank)?
- Is (blank) or misspelling a proxy for other characteristics?
- How about Duns Industry sector?

Riddles

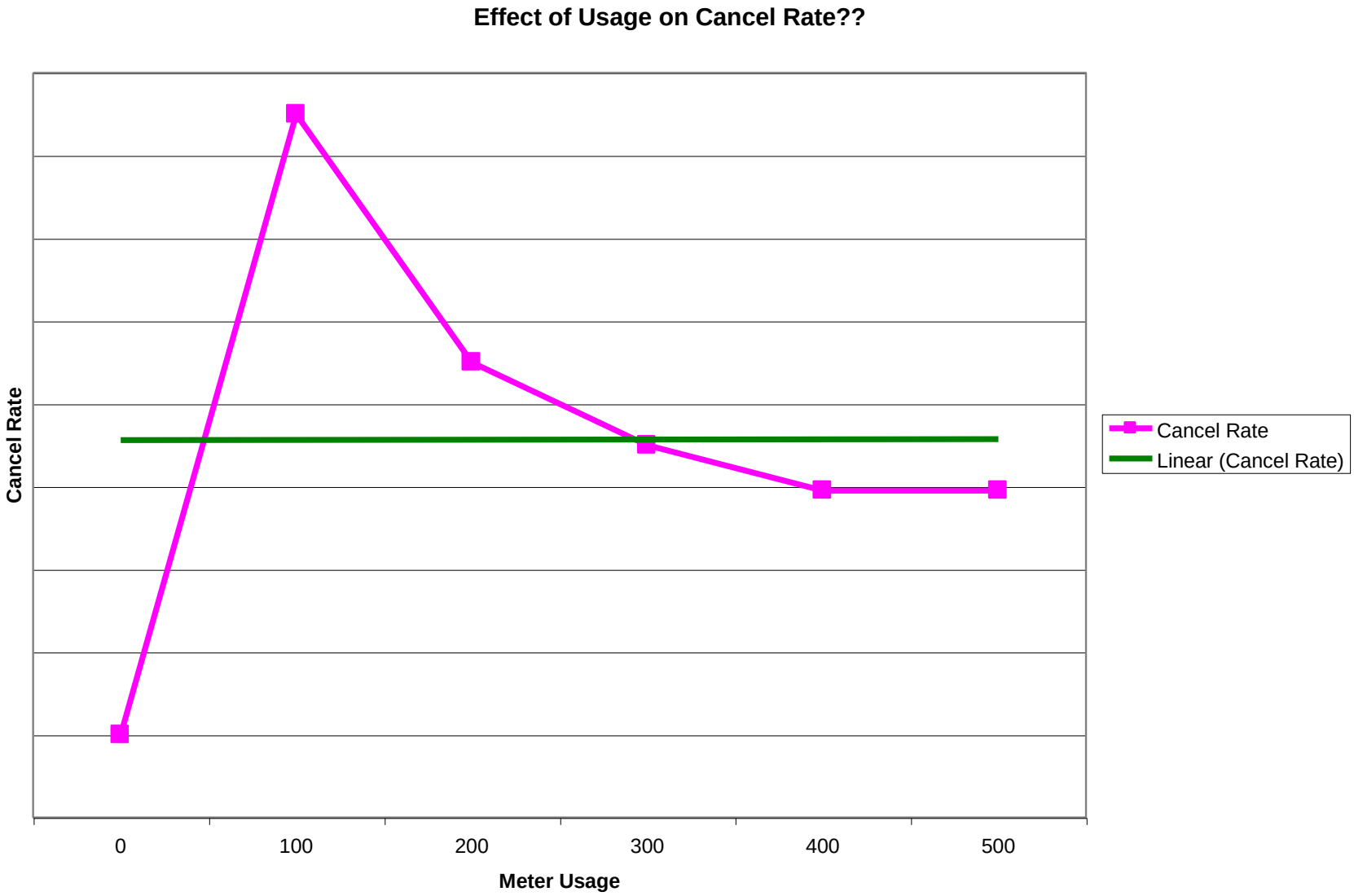
- Product Usage a predictor of retention
- Optimum student loan amount
- Effect of sales channel on retention
- Longevity dependency on location

- Fitting a line: effect of product usage on retention
  - Large database

Effect of Product Usage (Line fitted to data points)

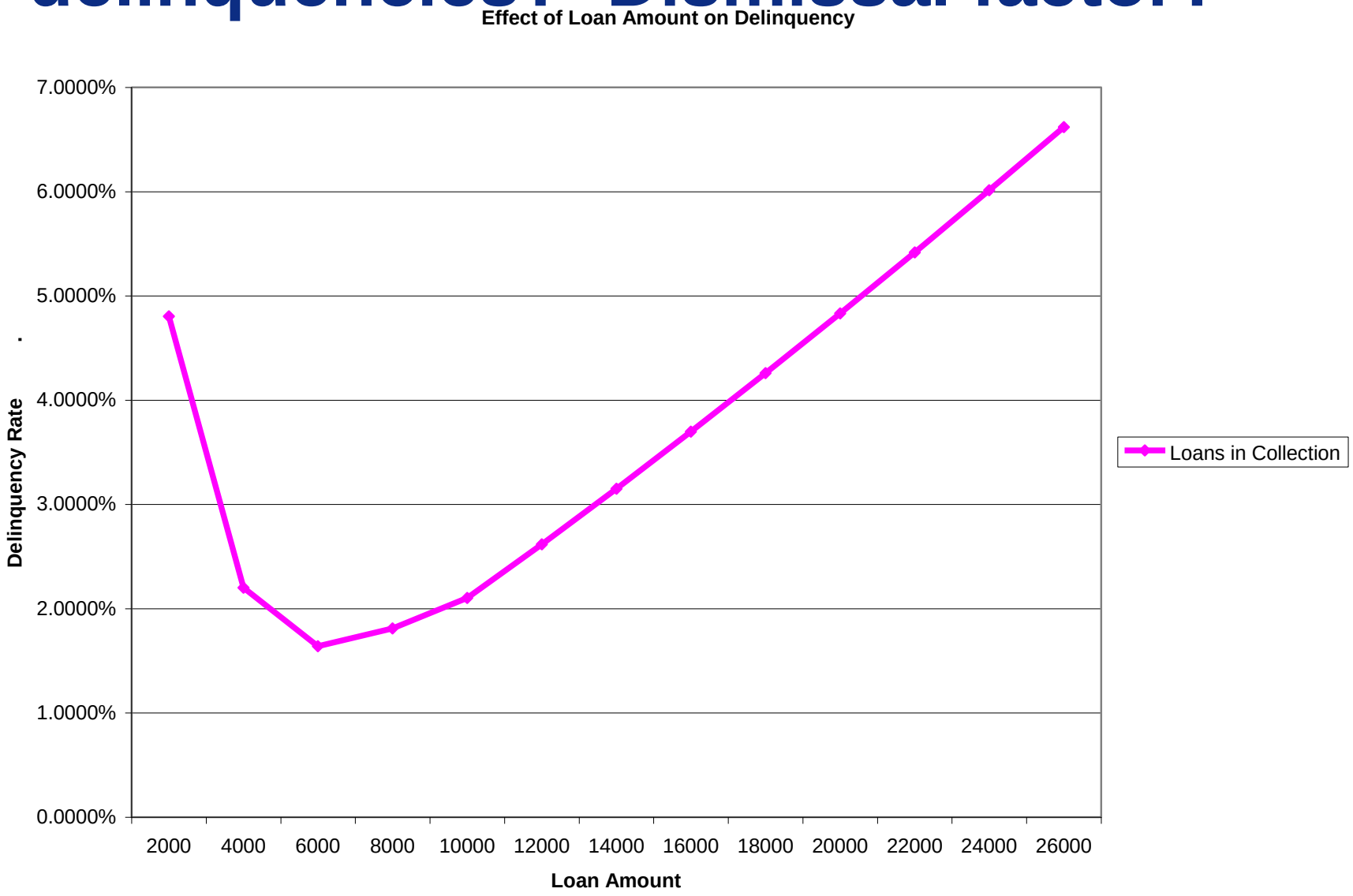


Is ‘zero’ really ‘missing’? Impact on bining and fitting.

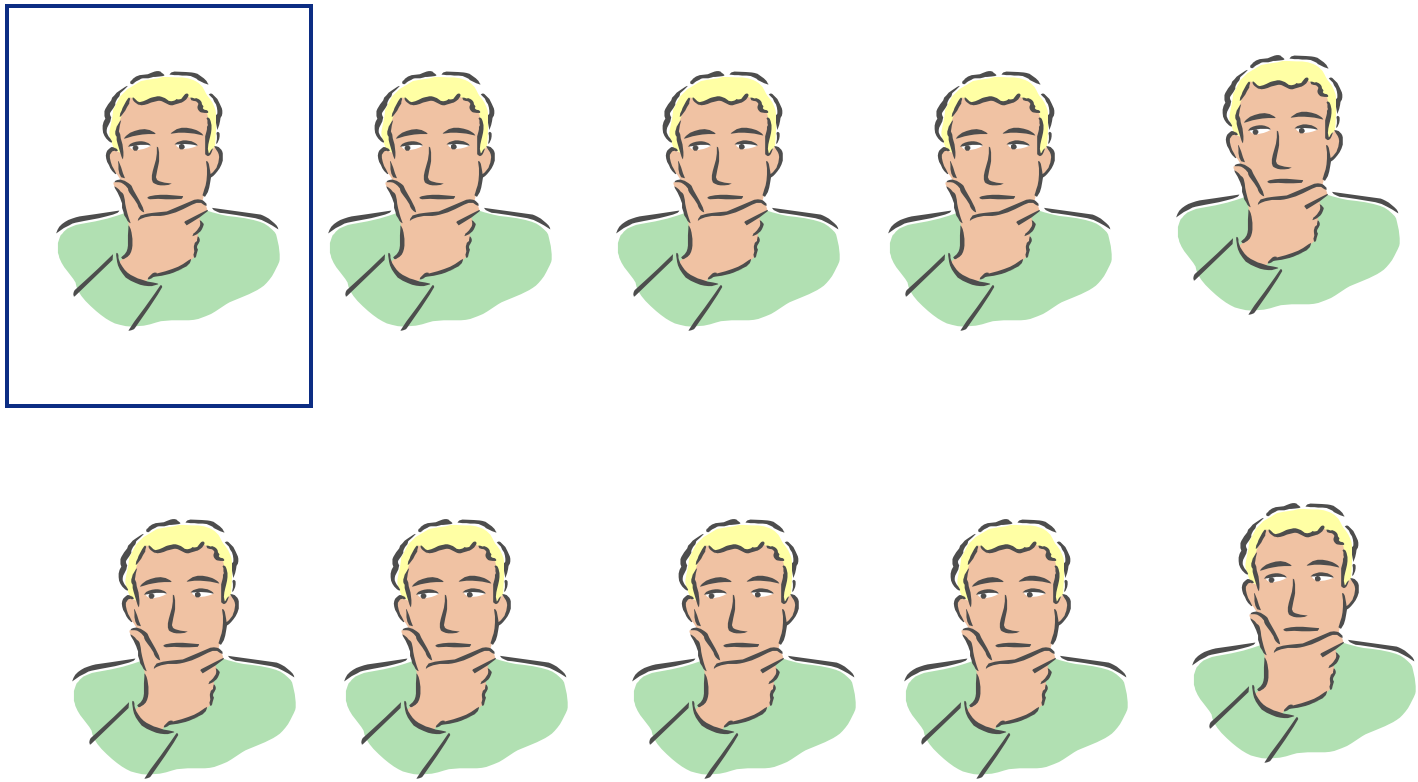


- Student loans and loan delinquency

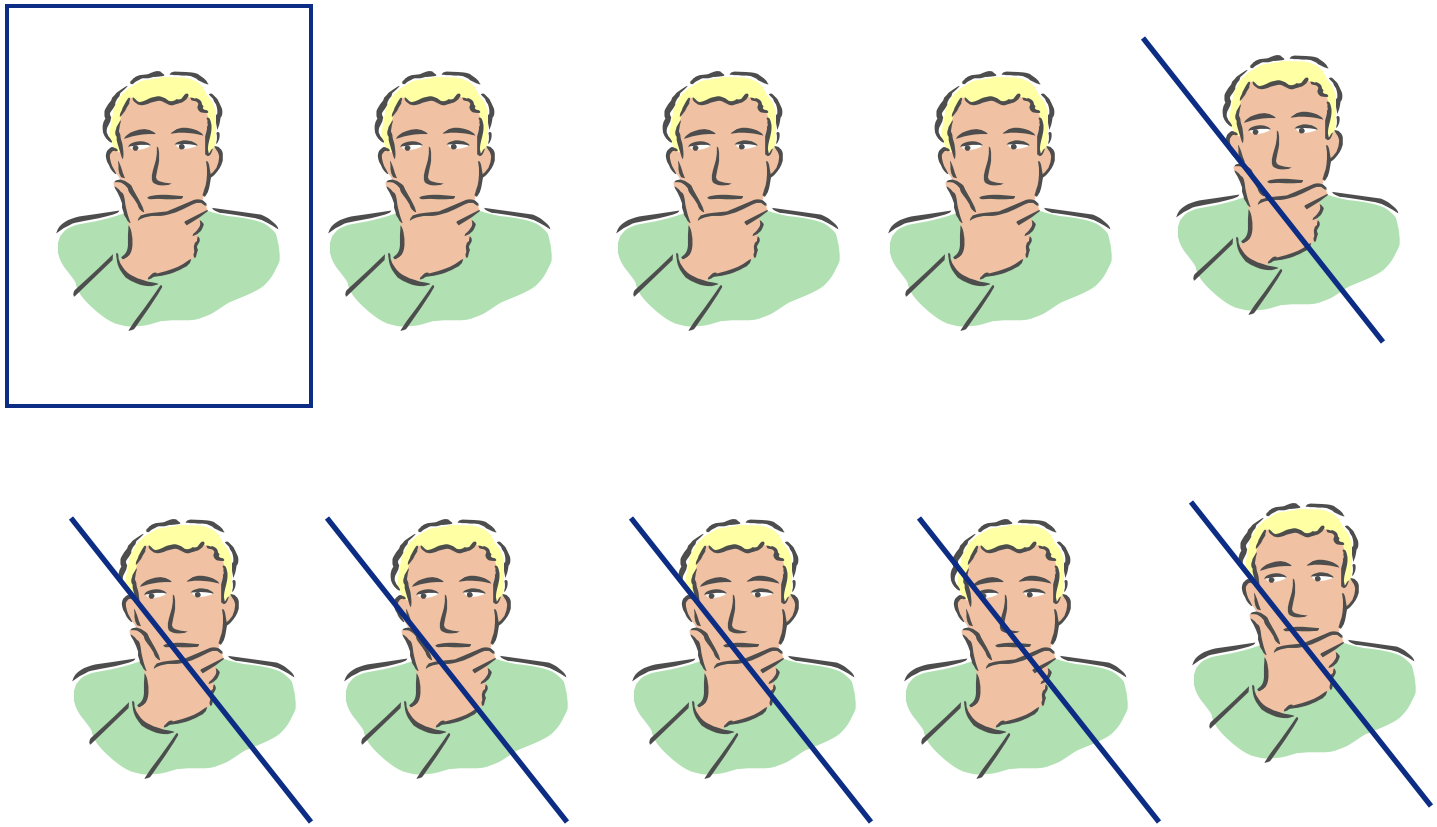
Do small loans result in high delinquencies? Dismissal factor?



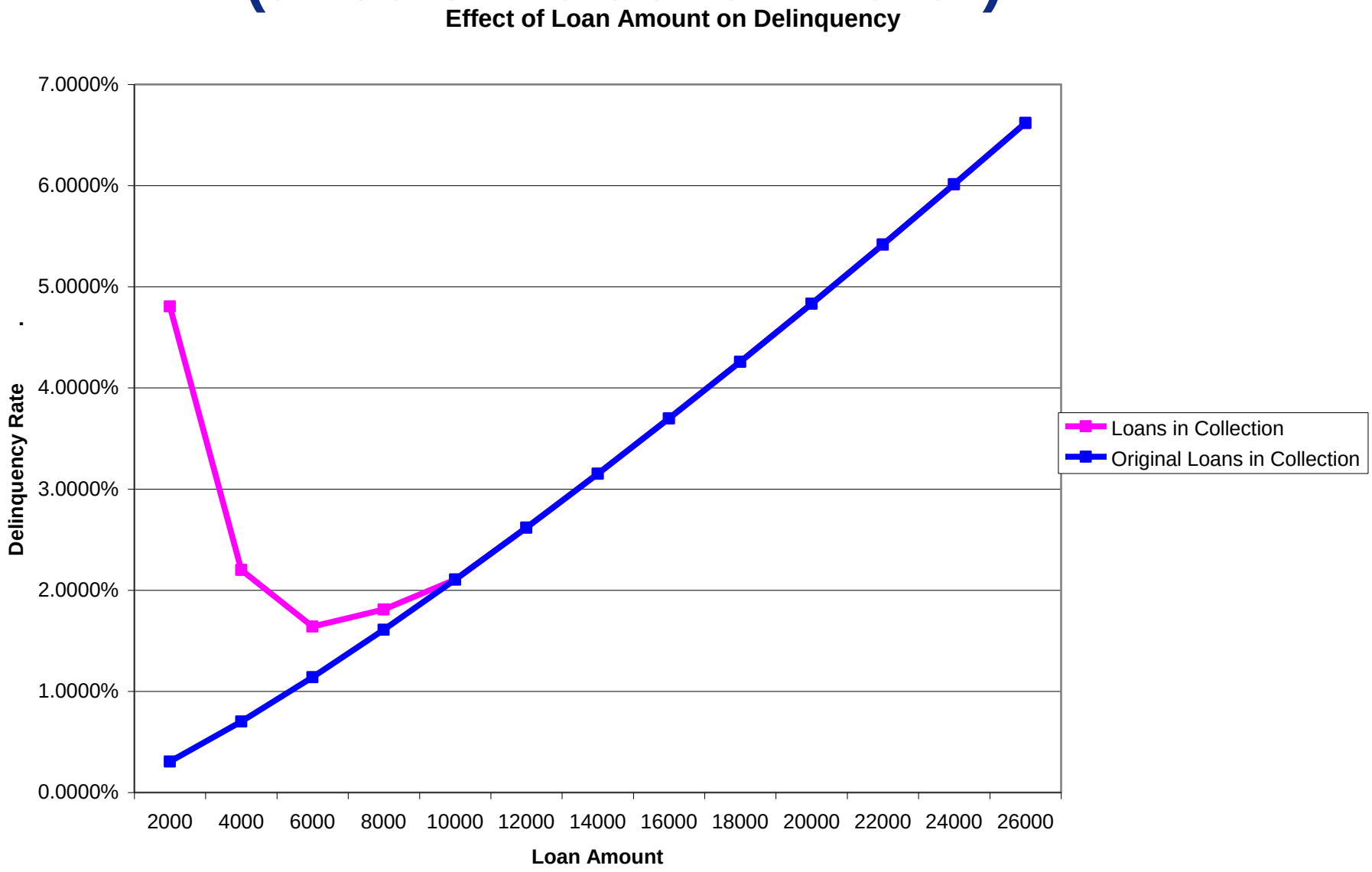
1 out of 10 in the original population



Looks like 1 out of 4 in the remaining population

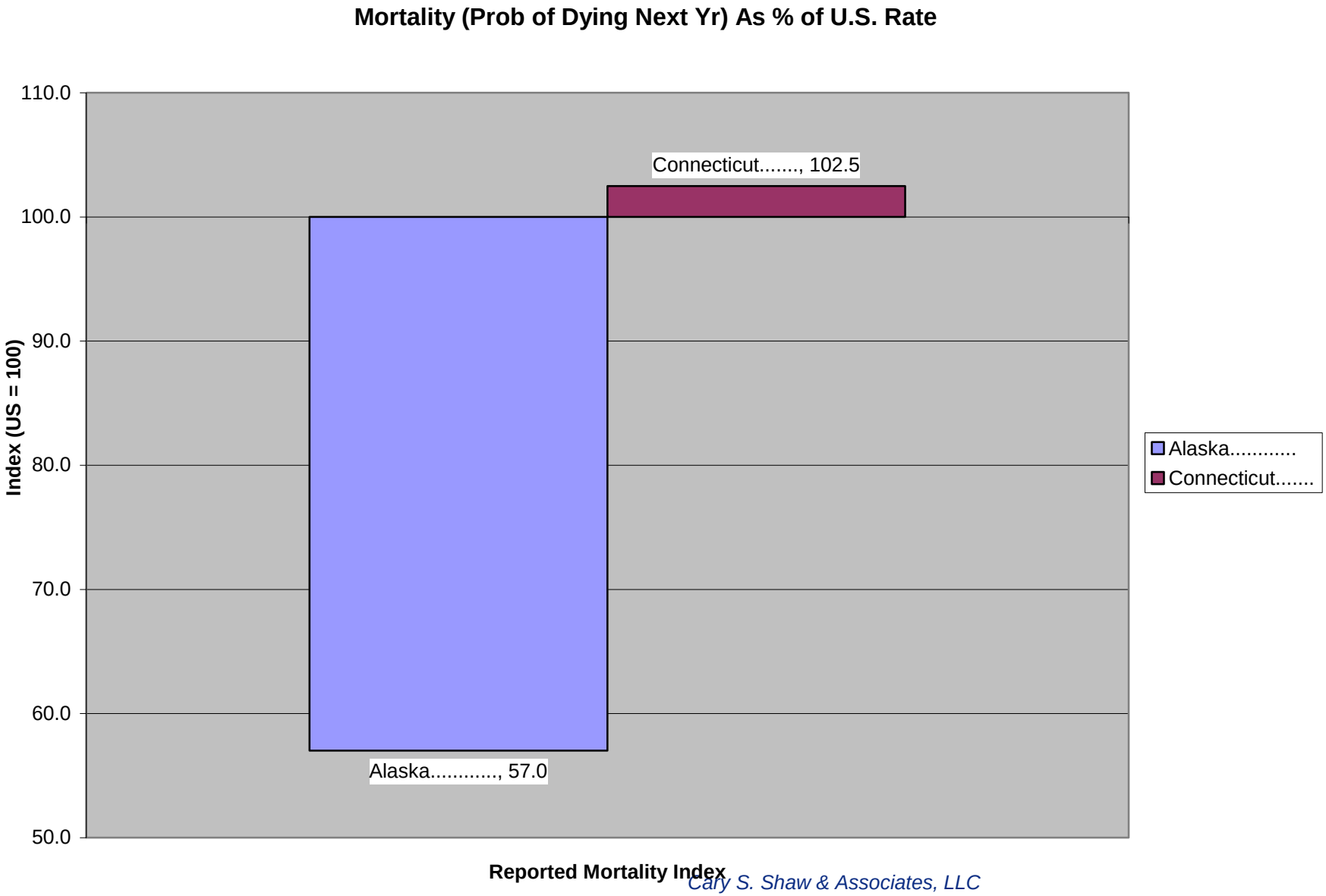


Watch the whole population of interest (affects the denominator)

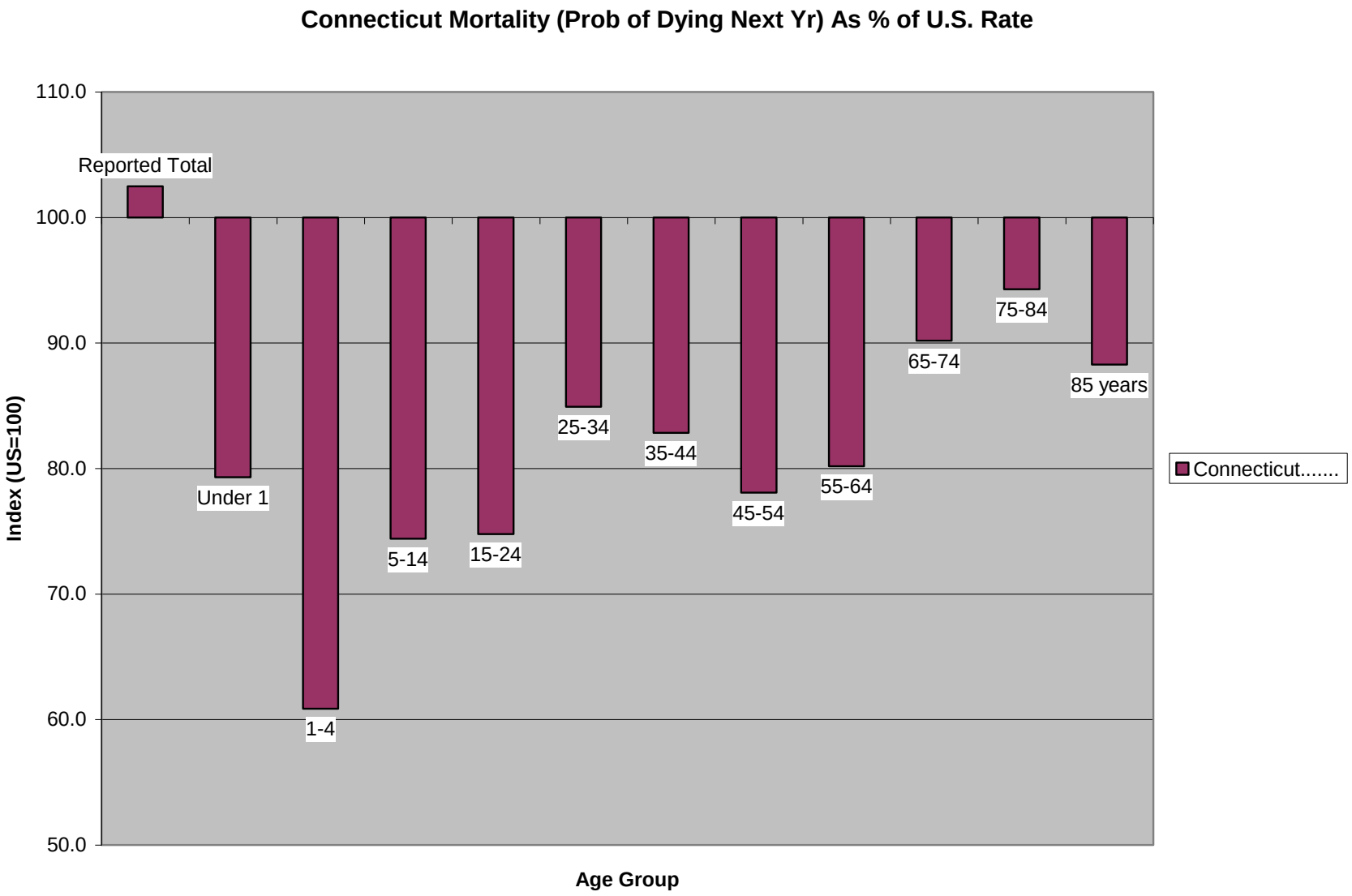


- Where you live can affect your life span.
  - Connecticut ?

Where is Mortality Lowest?

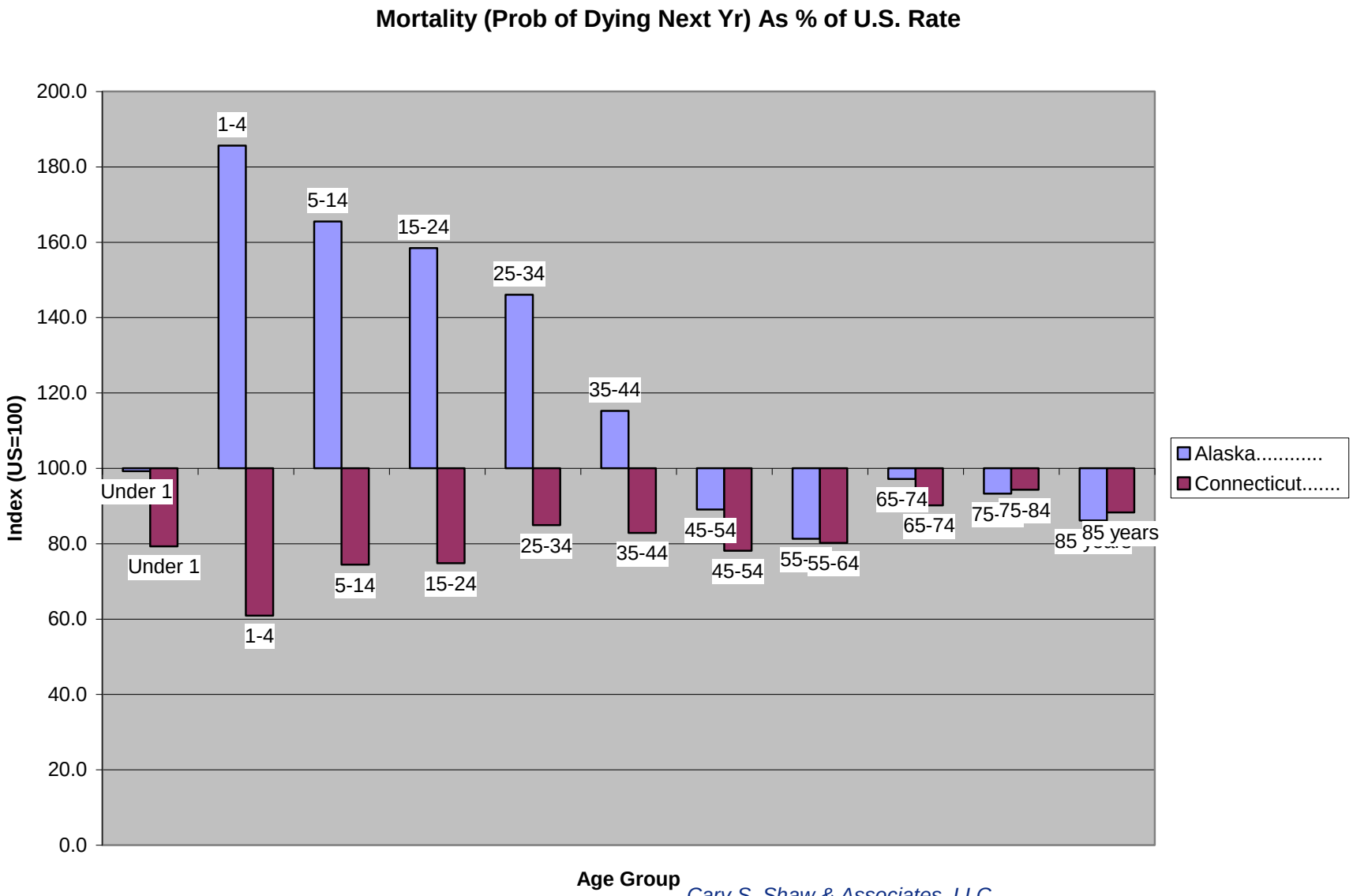


Is Connecticut Mortality Above Avg?

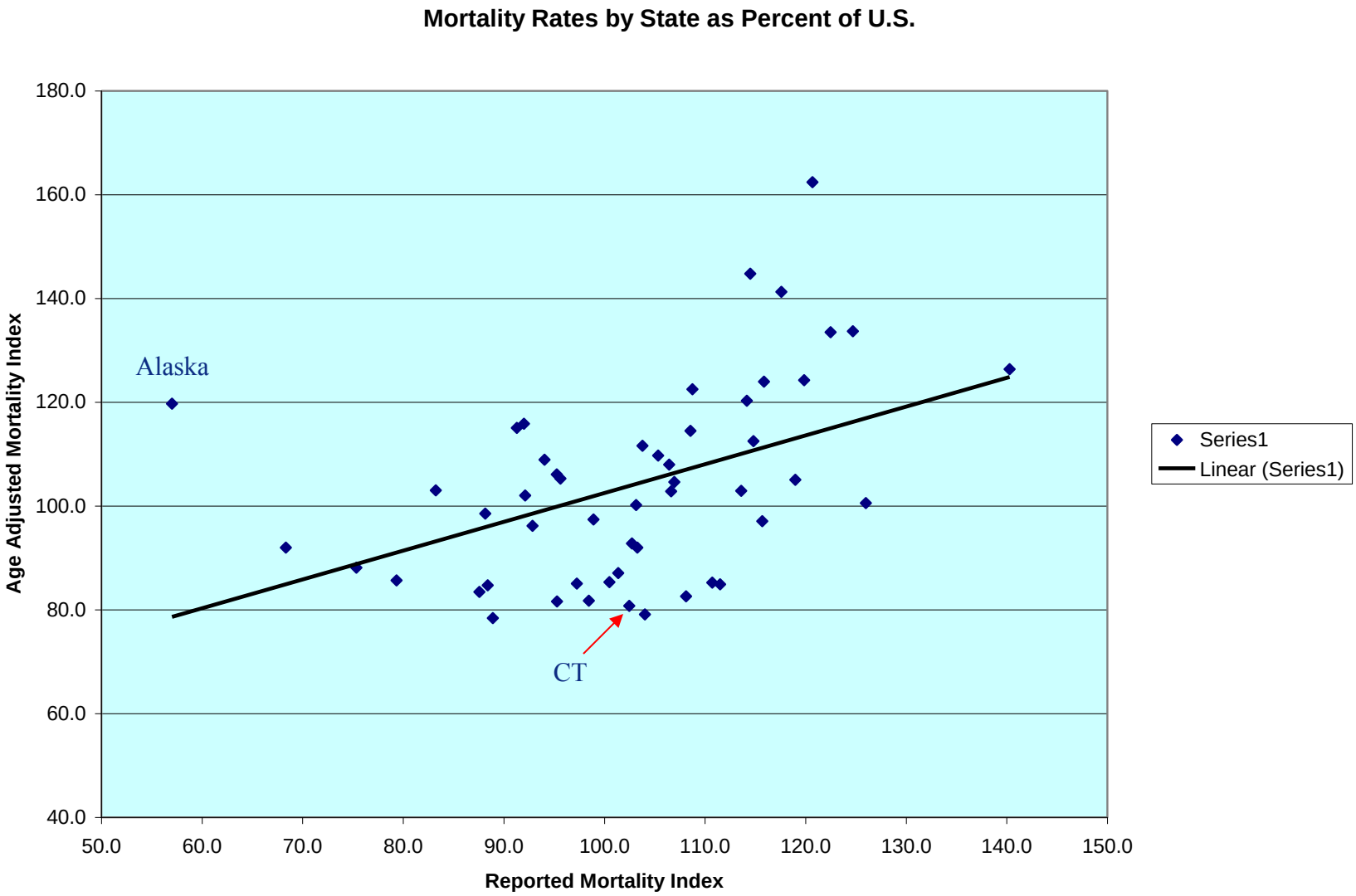


Cary S. Shaw & Associates, LLC

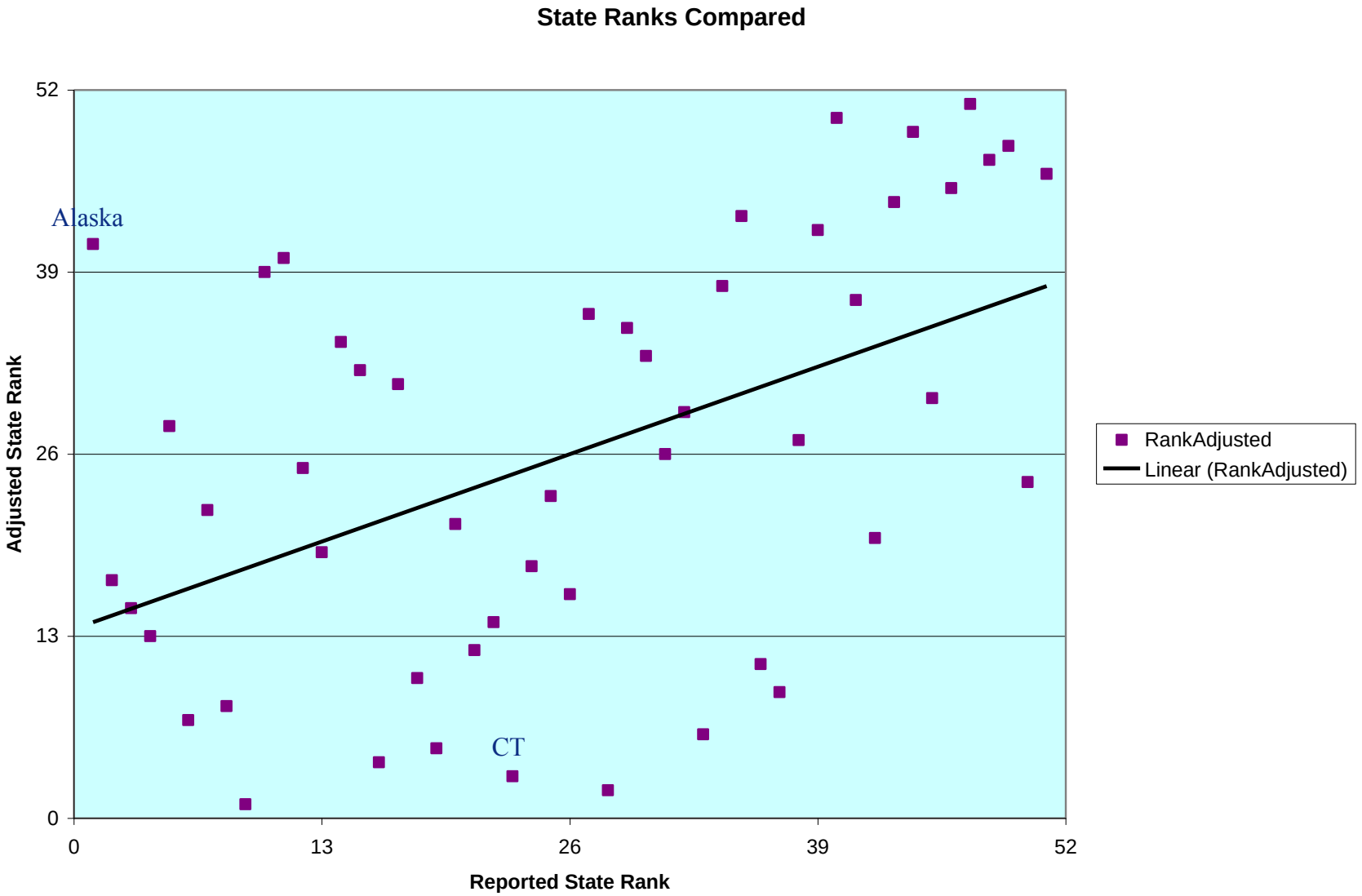
Compare with Alaska by Age



Compare Observed with Age-Adjusted



Compare Rankings



- For full table identifying mortality rate in each state, contact Cary S. Shaw & Associates, LLC

Riddles

- Product Usage a predictor of retention
- Optimum student loan amount
- Effect of sales channel on retention
- Longevity dependency on location

Lessons

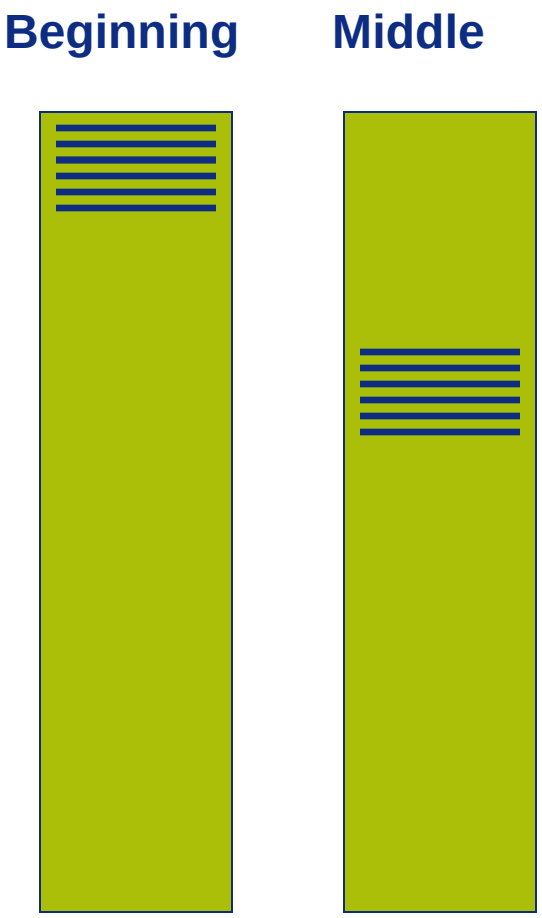
- Separate cause and effect by chronology.
- Find the initial populations under study.
- Distinguish between missing and zero values.
- Search for significant lurking variables.
- Creatively test your own ideas.

Why Sample?

- Examine large databases
- Rich on detail, economical on resources
- Check own analytics
- Cross check IT
- Cross check vendors

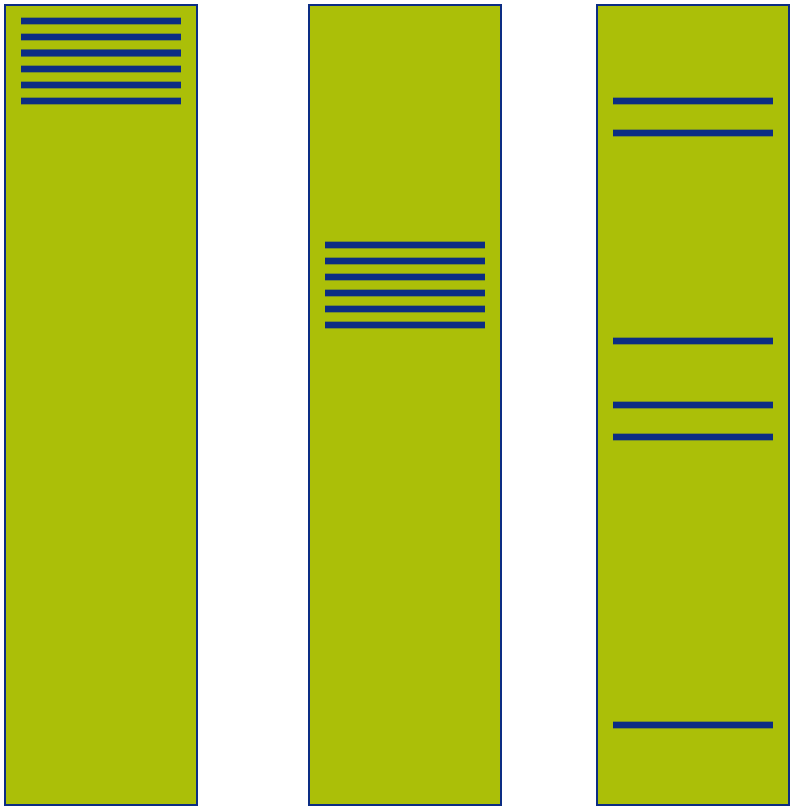
- May enable GREATER accuracy because
  - Enables correction of representative points
  - Enables understanding of representative points

Which sample is most representative?

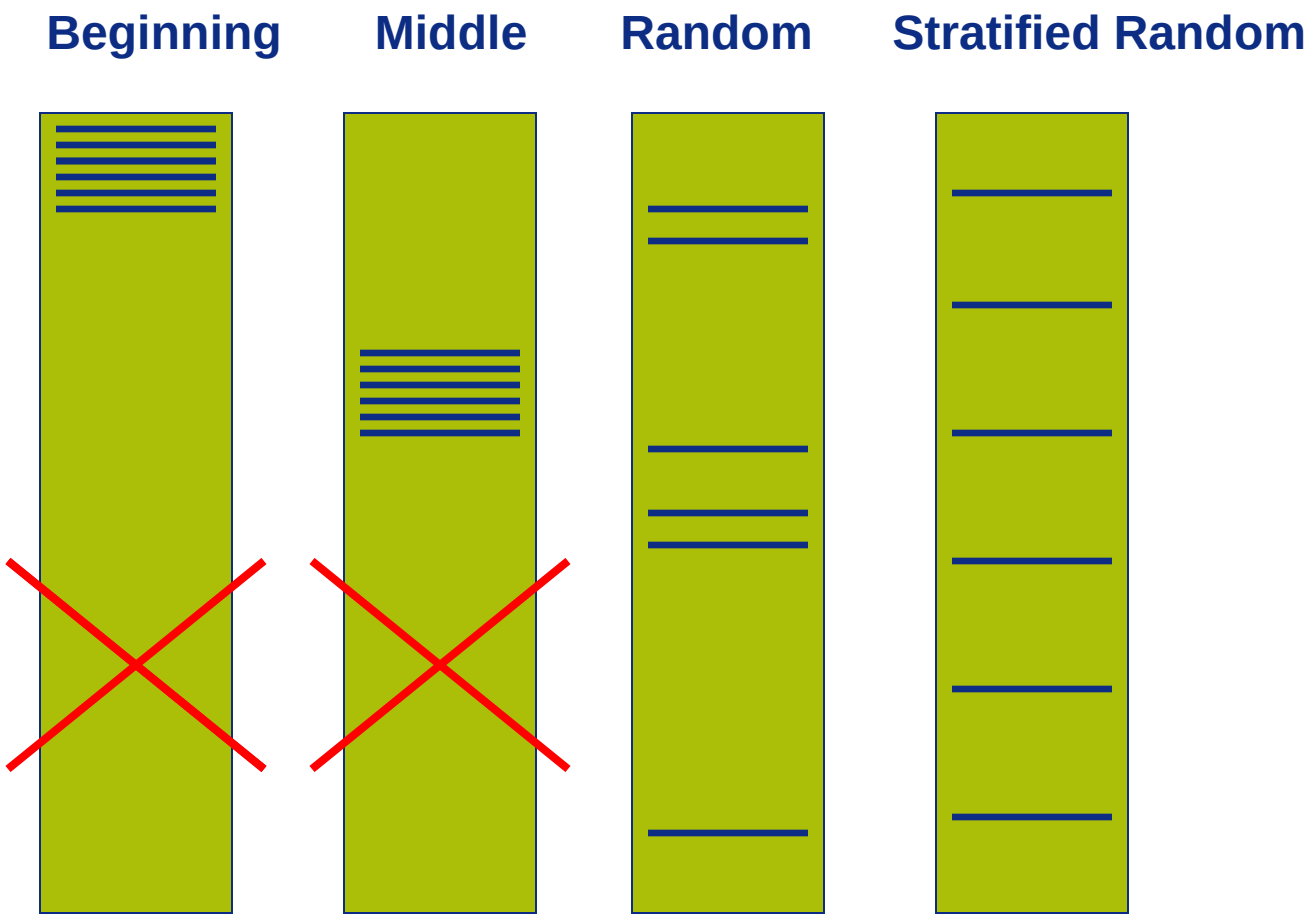


Which sample is most representative?

Beginning Middle Random



Which sample is most representative?



Stratified Random is making sure the sample is representative.

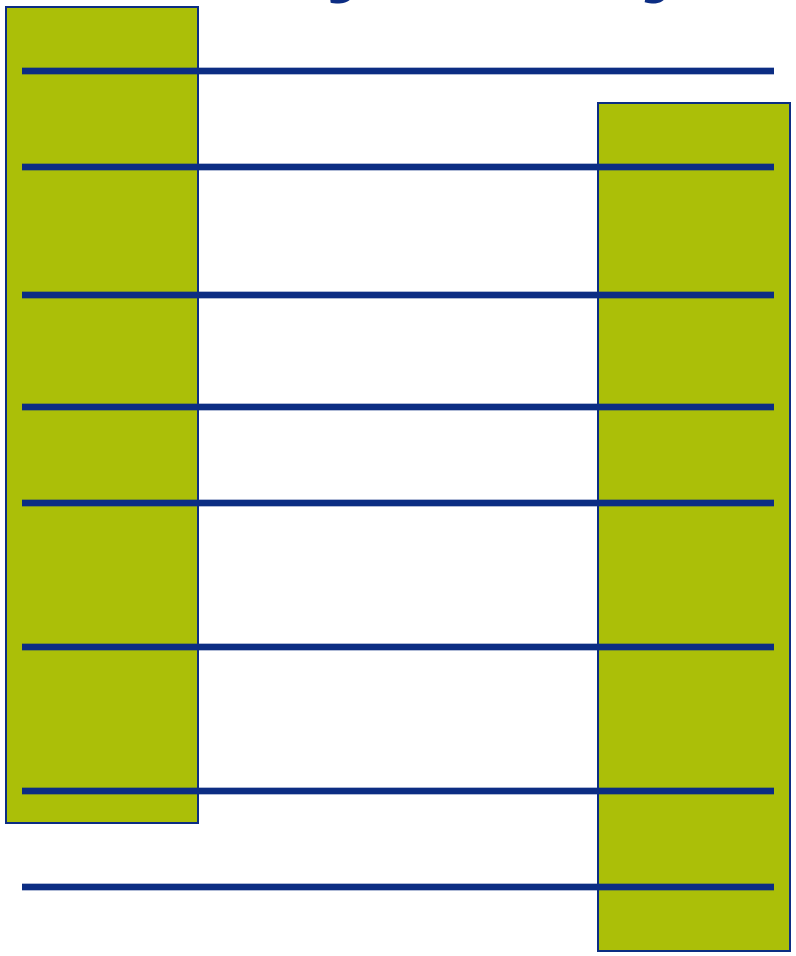
- Method A:
  - Divide data into strata (categories, deciles, etc.)
  - Pick target number from each strata
- Method B:
  - Sort file by factors known to be important (such as income, age, even ID number)
  - Pick every n th record, starting with record n/2

What happens when you match samples?
How big is the resulting sample?

- Year 1 Year 2 Year 3
- 10% 10% 10% = 0.1%

- Customer Product Sales$ Supplies A/R
- 10% 10% . 10% 10% 10%
- = 0.001%

There's a way to take samples that match even as the data is changing (don't lose by inner joins)



Examples of Record IDs

- Social Security Number
- Credit Card Number
- Account ID

- Develop algorithm to transform Record ID into Random Number.
- Then select from each file based on Random Number range.
- Samples from different files will then match.

Example: Duns File Update

- How did NAICS code change?

- Jan NAICS = April NAICS
- Jan NAICS <> April NAICS
- No January Record
- No April Record

Example: Duns File Update

- How did NAICS code change?

Jan NAICS = April NAICS	1006
Jan NAICS <> April NAICS	0
No January Record	0
No April Record	0

Example: Duns File Update

- How did NAICS code change?

Jan NAICS = April NAICS	829
Jan NAICS <> April NAICS	107
No January Record	64
No April Record	70

We covered:

- Qualify
  - Sniff, Define, Frequency Check
- Interpret
  - Lurk, Miss, Population
- Sample
  - Methods, Application
- Summary
  - Tools, Lessons

Lessons

- Perform the sniff test.
- Check number of records.
- Do frequency counts (highest, lowest).
- Look for missing values (distinguish from zeros and duplicates).
- Investigate lurking interactive variables.
- Analyze what happened to original populations, not final populations.
- Examine representative samples (stratified preferred).
- Assume errors and look for them. (Apply G.I.V.O.).

- FIN -

- Avoid the pitfalls of found data and reach for the Stars

- For further information contact:
- Cary S. Shaw & Associates, LLC
  (203)505-3180
  caryshaw@optonline.net