Data-Driven Safety Training

Introduction

Part III

Carlos Sun, Praveen Edara, Yaw Adu-Gyamfi

University of Missouri

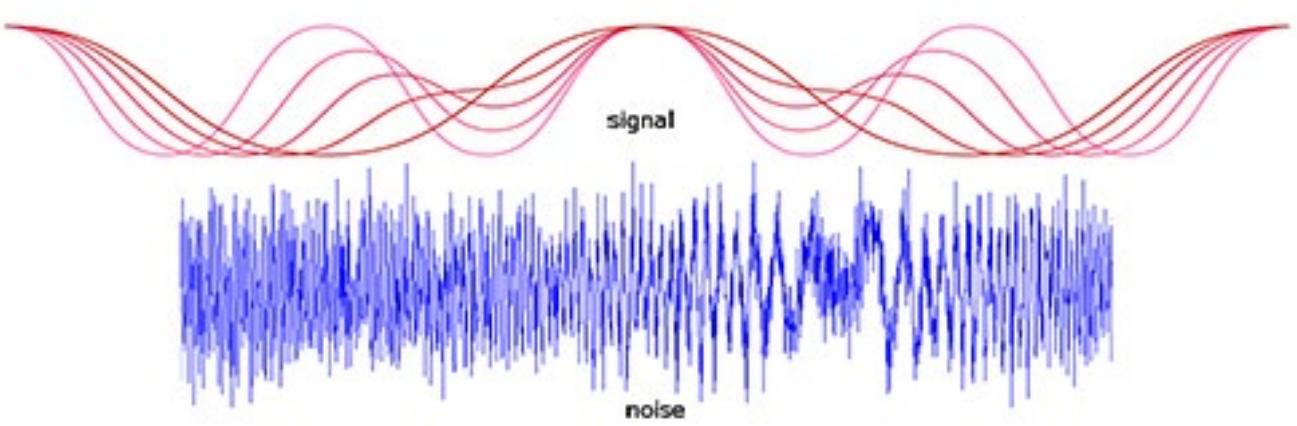
Missouri Center for Transportation Innovation

Outline

- Background/motivation
- Subjective vs. objective safety
- Complexity of traffic crashes & data
- **Regression to the mean bias**
- Review of statistics
- Use and application of data-driven safety methods

Regression to the Mean (RTM) Bias

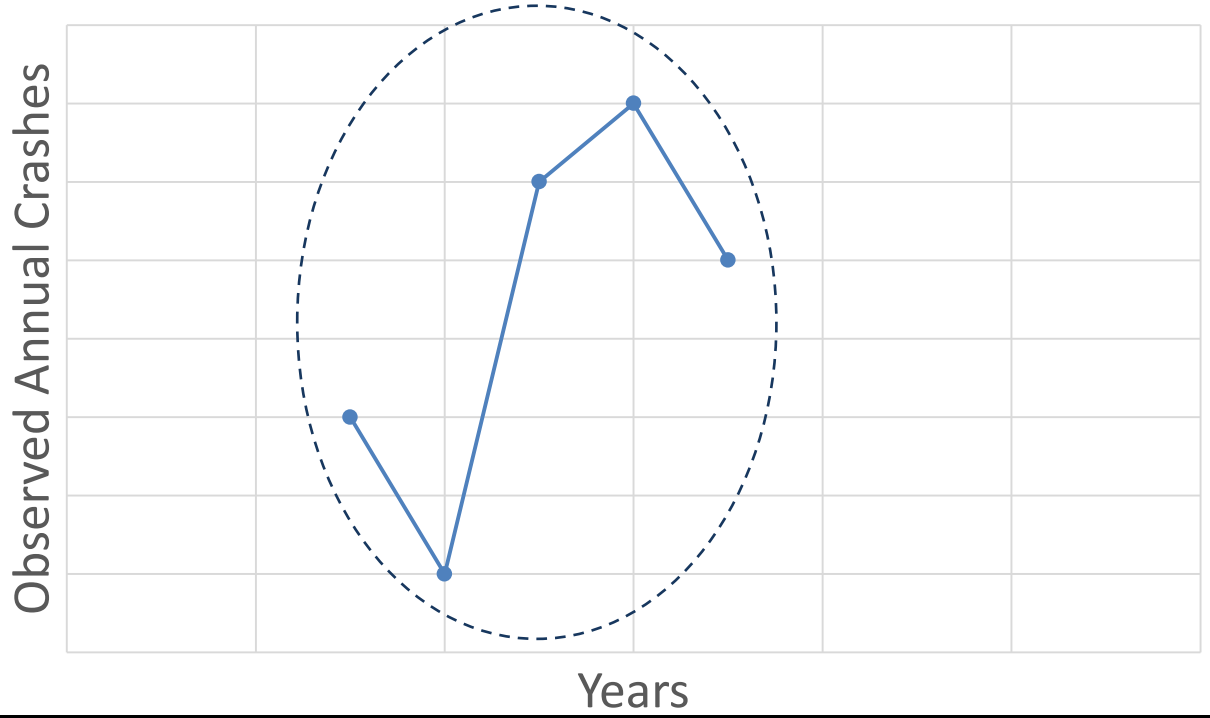
- Definition
  - making decisions based on limited data (noise) that is not representative of actual underlying trends



Quality Digest 2020

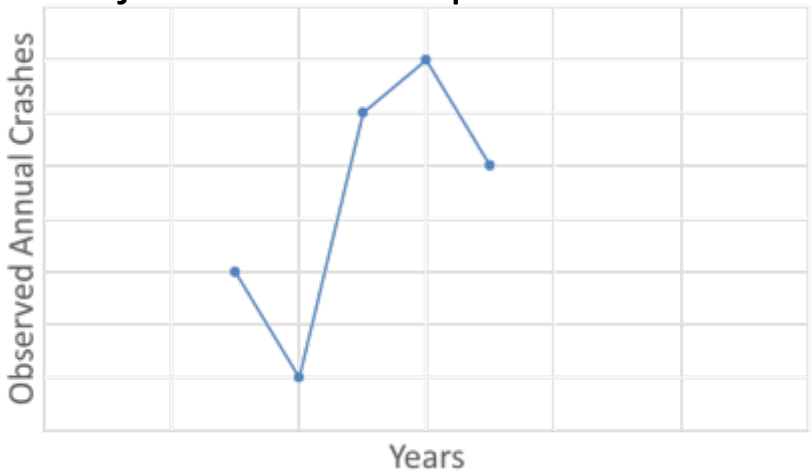
Regression to the Mean (RTM) Bias

- Example - tracking intersection crashes for 5 years



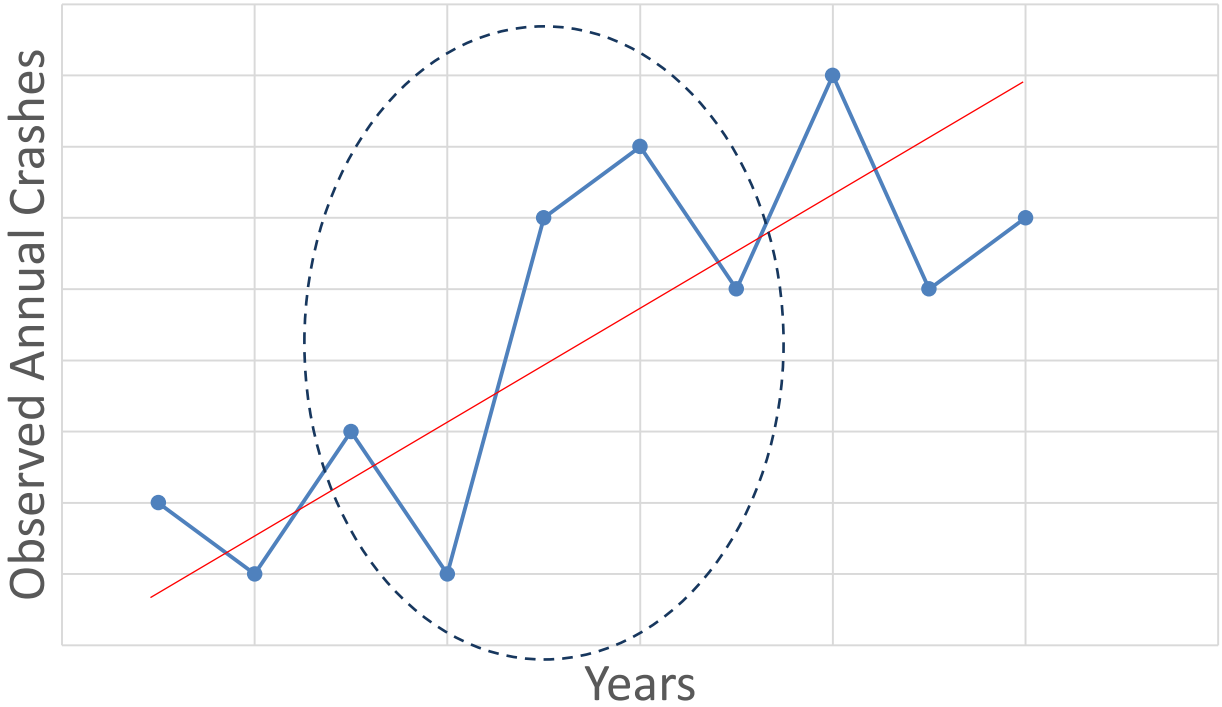
Regression to the Mean (RTM) Bias

- Example - tracking intersection crashes for 5 years
  - Given the increase in crashes for the past 3 years, should we?
    - Look into making intersection safety improvements?
    - Ignore since last 3 years was just a random spike?
    - Wait for more data?



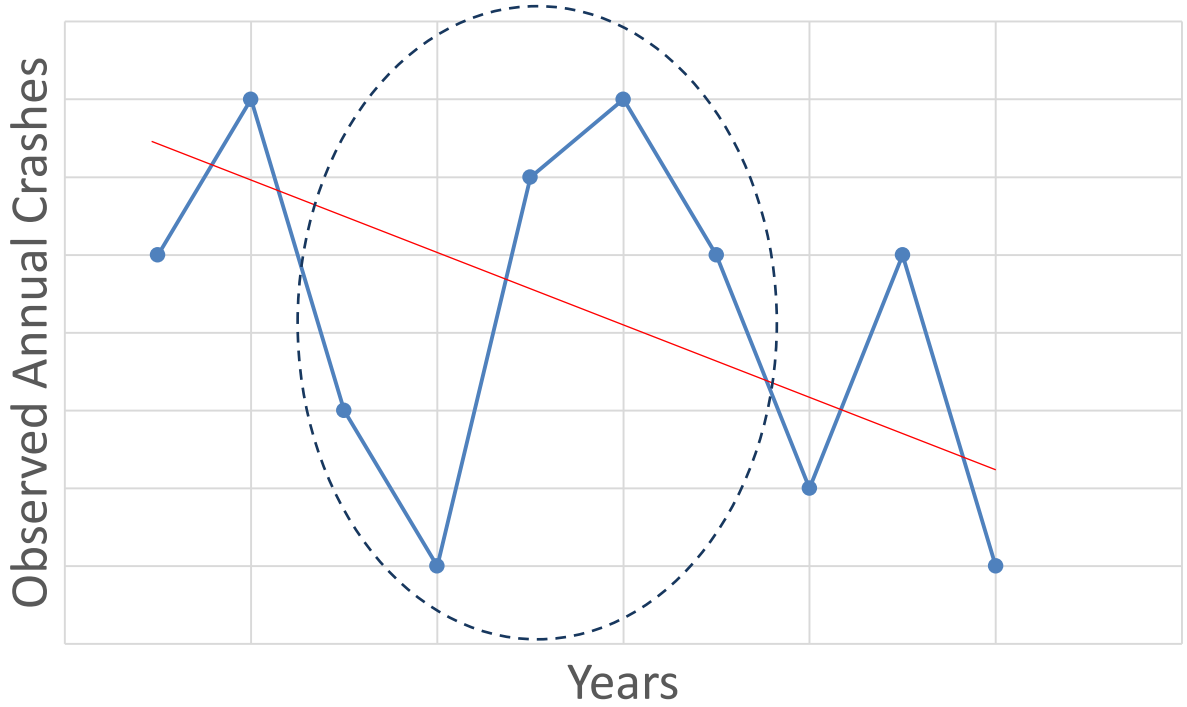
Regression to the Mean (RTM) Bias

- What if the long term trend was …



Regression to the Mean (RTM) Bias

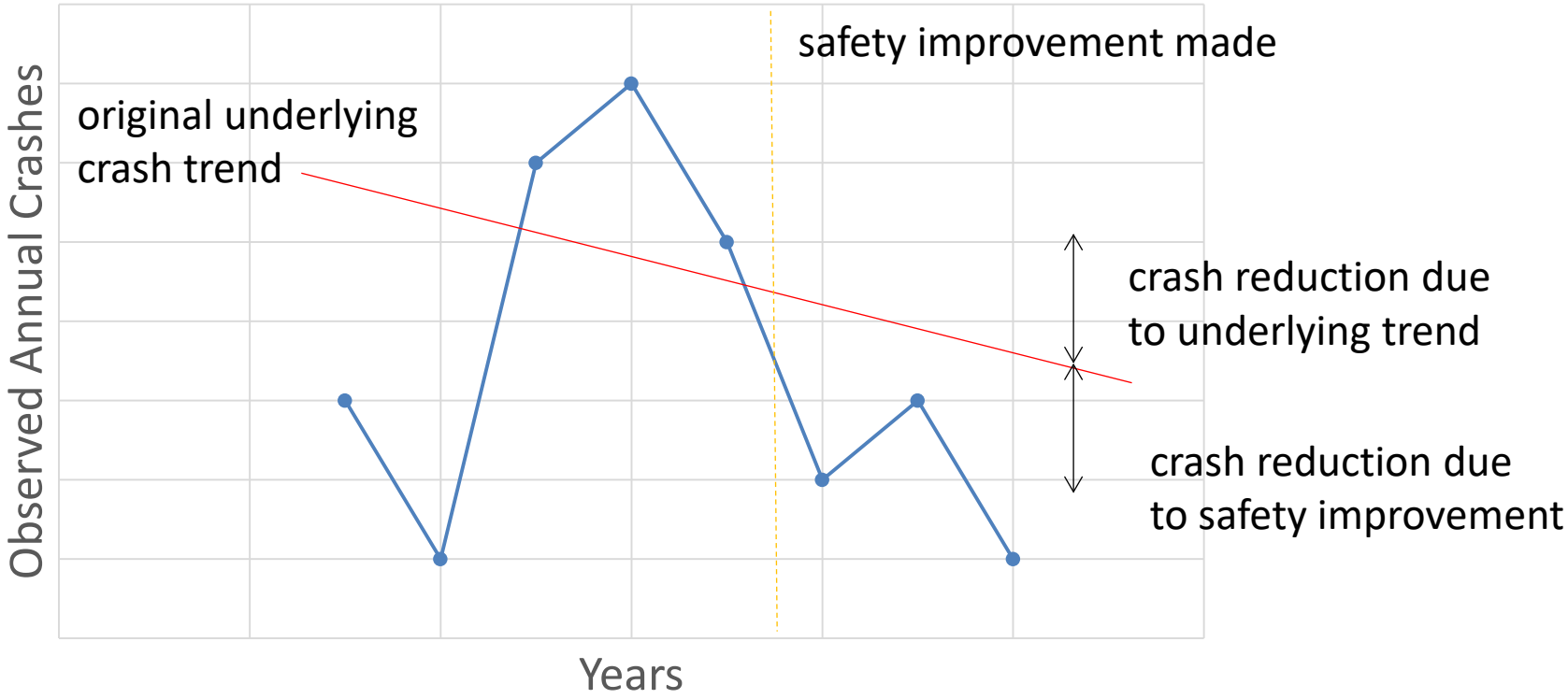
- What if the long term trend instead was …



Regression to the Mean (RTM) Bias

- Data-driven safety seeks to discover the underlying trend
- Avoid making decision using small sample sizes
- Example – naïve before/after study after implementing a safety countermeasure
  - 3 years before = 9 crashes/year
  - 3 years after = 3 crashes/year
  - was the countermeasure effective?
  - numbers do not lie, or do they?

Accounting for RTM Bias



Accounting for RTM Bias

- How do data-driven safety methods
  - counter small sample size problems?
  - mitigate the RTM bias?
- By using more data, how?
  - take advantage of national crash databases
  - take advantage of national safety research
  - applying Empirical Bayes method: predicted + observed
- Example – rural 4-lane interstates
  - use data from multiple states with similar facilities

Empirical Bayes Adjustment

- Observed crash frequency comes directly from the relevant site
  - advantage – from the relevant site at issue
  - disadvantage – only one site, small sample
- Predicted crash frequency comes from a large national database + Missouri calibration database
  - advantage – data from multiple similar sites, larger sample
  - disadvantage – uncaptured differences between relevant site and similar sites

Empirical Bayes Adjustment

- Use both observed and predicted
- Compute a reliability weight between 0 and 1
  - to trust observed or predicted more
  - w=1 only rely on predicted
  - w=0 only rely on observed
- overdispersion is a measure of the reliability of the predicted value, used to compute w

Accounting for RTM Bias

- But things are a bit different in Missouri
  - we are not as flat as KS
  - our drivers are more polite than those aggressive East Coast drivers
  - we get snow unlike many cities in the west coast

Accounting for RTM Bias

- But things are a bit different in Missouri
- We calibrate our prediction models to our State's conditions and produce our own severity distributions

Accounting for RTM Bias

- Rural multilane divided highway example
- 2.18 miles of US 50 W in Centertown, Cole County
  - use national data to predict safety as 9.83 total crashes per 3 years
  - apply MO calibration factor of 0.74, then predicted safety is changed to 7.27 crashes/3 years instead
  - combine predicted crash with observed crash number of 9 crashes/3 years using Empirical Bayes
    - 7.81 crashes/ 3 years