MANAGING RISK DNV

Avoiding the tar pit of numbers

Challenges of Statistical Analysis with Software Engineering Data

Dave Card
David.card@dnv.com

Agenda

MANAGING RISK

- Background
- Data Issues
- Statistical Challenges
- Summary

Background

MANAGING RISK DNV

- Many forces encourage increased use of statistical analysis in management and improvement of software (and other) engineering processes, e.g., CMMI, Six Sigma
  - Process Performance Models
- Bad analysis leads to bad decisions – correctness is a CMMI appraisal issue

What is Statistical Analysis?

MANAGING RISK DNV

- Building mathematical models based on data
- Using those models to evaluate or estimate measurable phenomena
- Quantifying the uncertainty associated with models and decisions

A model underlies every statistical analysis technique

Data Issues

MANAGING RISK

- Alignment
- Quality
- Quantity
- Linkage

Alignment

MANAGING RISK DNV

- Data and analysis must be linked to something important to be useful, i.e., aligned with goals
- Many organizations only start to approach goals systematically at higher levels of maturity

Establish objectives before anything else.

Quality

MANAGING RISK DNV

- Most measurement involves humans
  - Misinterpretation is common
  - Motivation is important
- Quality of data is not really known until analysis is attempted

Good definitions, incremental implementation, and early feedback help.

Quantity

MANAGING RISK DNV

- Processes often not designed to be measured (or managed)
  - Large activities => infrequent data
  - Lean process designs make more data available
- Amount of data needed depends on analysis technique (degrees of freedom)
- Experiments (Six Sigma) versus Measurement Programs (CMMI)

Large amounts of data are not always necessary for meaningful analysis.

Linkage

MANAGING RISK DNV

- Individual databases are often constructed for specific purposes (e.g., peer reviews, estimation, problem reports, earned value) by different groups
- Many important problems require consideration of data from multiple sources
- Important potential linkages include: project, phase, artifact

Define a “master plan” encompassing all repositories.

Statistical Challenges

MANAGING RISK DNV

- Distributions, Types, and Scales
- Significance and Substance
- Interpreting Causality

No silver bullet – find the right technique for your data.

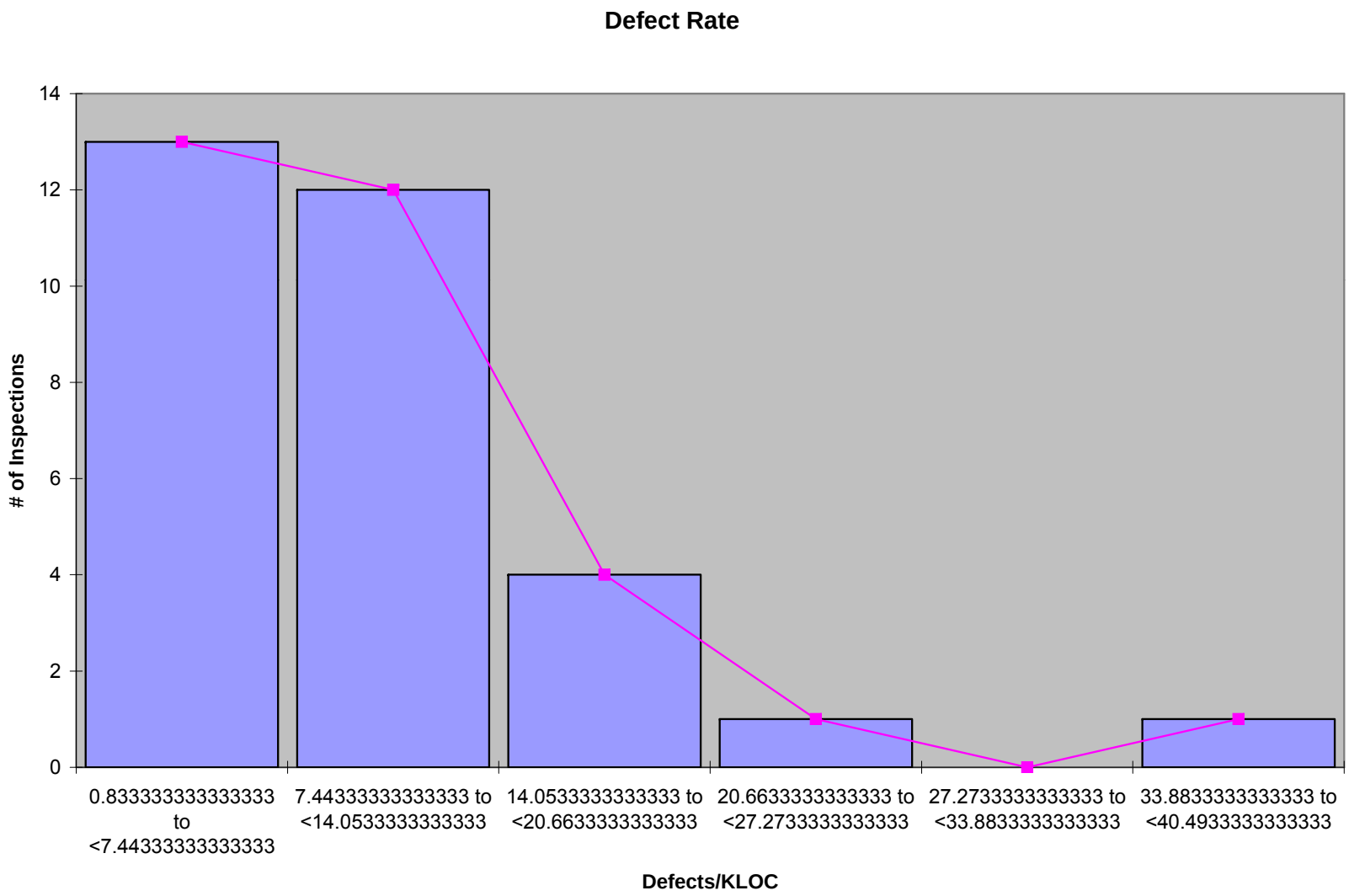
Distributions, Types, Scales

MANAGING RISK DNV

- Probability density function (continuous) or probability distribution (discrete)
- Nominal and ordinal data cannot be normally distributed
- Distributions of data from interval and ratio scales be observed (e.g., histogram) or inferred
  - Normal
  - Poisson
- Distribution of sample means may help (≠ law of large numbers)
- Avoid transformations

Look at the data before attempting statistical analysis.

Histogram Indicating Poisson Distribution



Examples of Statistical Techniques

MANAGING RISK DNV

	Second Measure		
First Measure	Nominal	Ordinal	Interval and Ratio
Nominal	Chi-square	Friedmann ANOVA	Normal ANOVA
Ordinal		Spearman Correlation	Spearman Correlation
Interval and Ratio			Regression Pearson Correlation

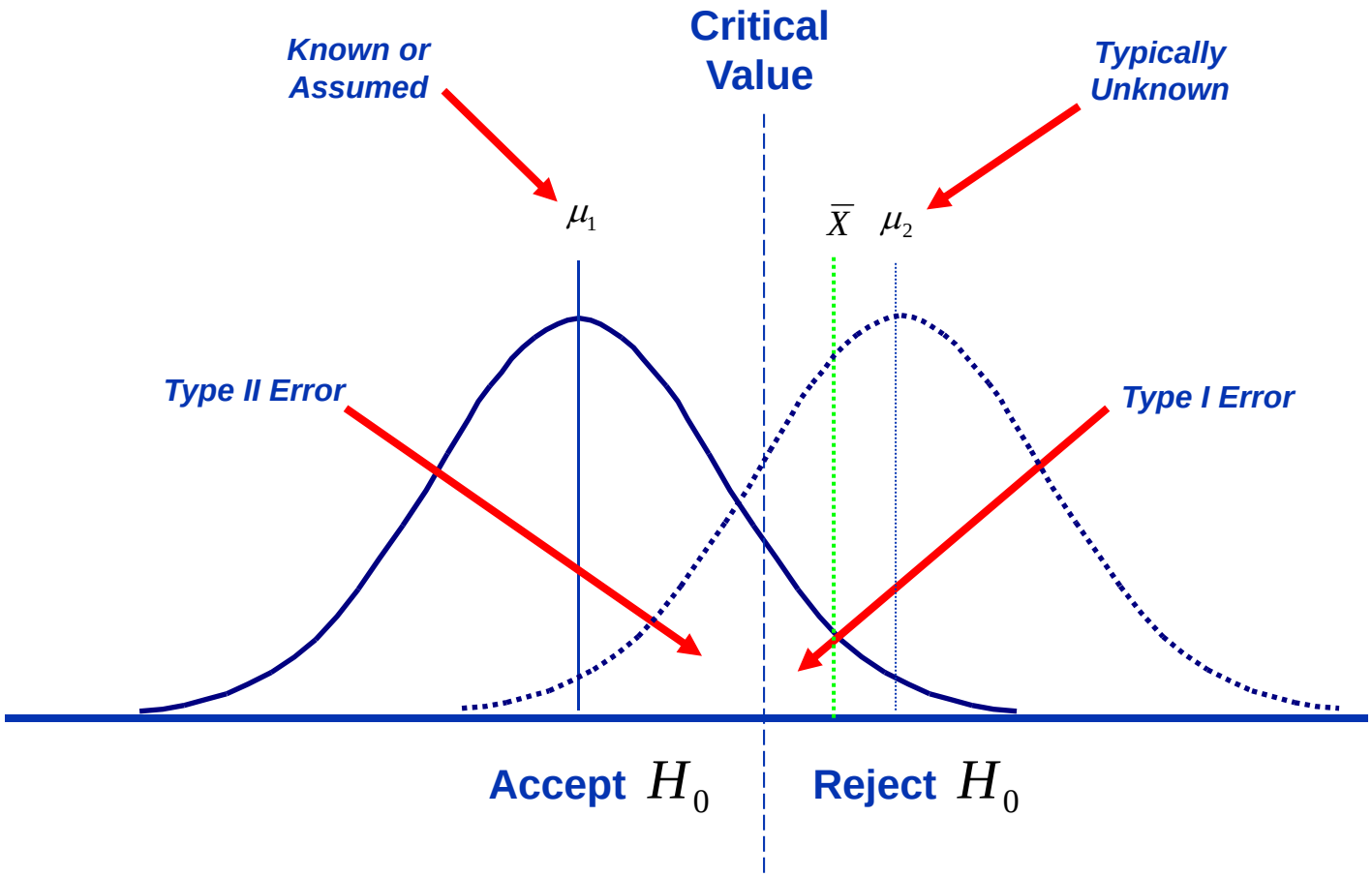
Substance and Significance

MANAGING RISK DNV

- Not all significant relationships (p) are substantive
- Consider magnitude of difference or use r^2 to assess importance of regression relationships

Look for important relationships first, then try to show them as significant.

Statistical Decision-Making



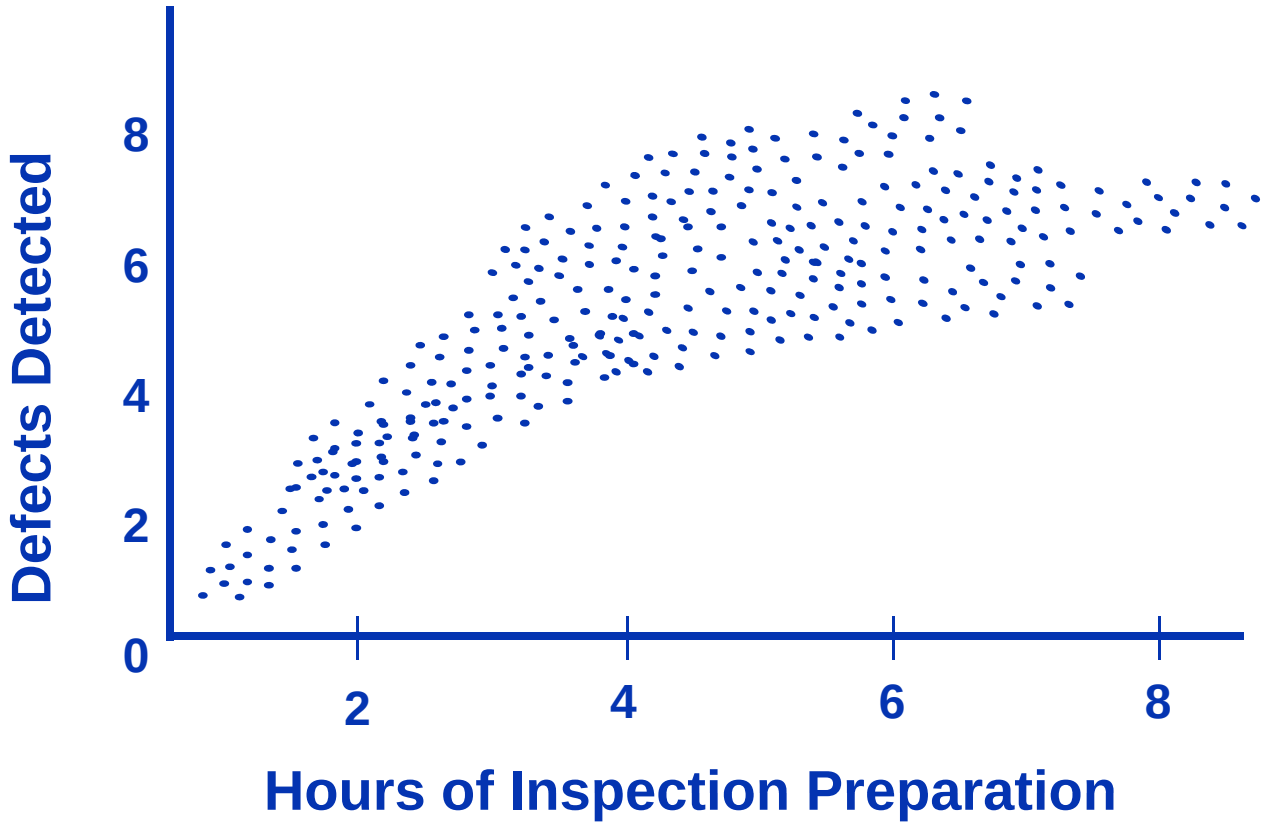
Conditions for Causality

MANAGING RISK DNV

- Association
- Cause strictly precedes effect
- Mechanism for interaction

Many people has an instinctive understanding of causal analysis, which is wrong.

A Causal Relationship?



Summary

- Statistical analysis is about modeling
- Nature of data determines the appropriate analysis technique
- Data collection must be planned with intended analysis in mind
- Use a systematic approach to causal analysis
- Learn and adapt