EuroSPI'99
Workshop on Data Analysis

Popular Pitfalls of Data Analysis

Tore Dybå, M.Sc.

Research Scientist, SINTEF Telecom & Informatics
Research Fellow, Norwegian University of Science and Technology
E-mail: Tore.Dyba@informatics.sintef.no

The Problem with Statistics

There are three kinds of lies: Lies, Damned Lies, and Statistics.
-Benjamin Disraeli

Statistics requires the ability to consider things from a probabilistic perspective, employing concepts such as: "confidence", "reliability", and "significance".

It is not a mathematical method for finding "the right answer".

We consider three broad classes of statistical pitfalls:

- **Sources of bias, which affect the external validity of our results**
- **Errors in methodology, which can lead to inaccurate results**
- **Problems with interpretation, or how the results are (mis)applied**

Sources of Bias

Statistical methods assist us in making inferences about a large group based on observations of a smaller subset of that group.

To make legitimate conclusions about the population, two characteristics must be present within the sample:

- **Representative sampling**: While this might be feasible for manufacturing processes, it is much more problematic for software processes.
- **Statistical assumptions**: The validity of a statistical procedure depends on certain assumptions about the problem, and the robustness of the applied techniques to violations in these assumptions.

Errors in Methodology

There are numerous ways of misapplying statistical techniques to real world problems. This can lead to invalid or inaccurate results.

Three of the most common hazards are:

Statistical power: If there is little statistical power, you risk overlooking the effect you are attempting to discover.

Multiple comparisons: You need to consider the complexity and the number of factor combinations that need to be examined.

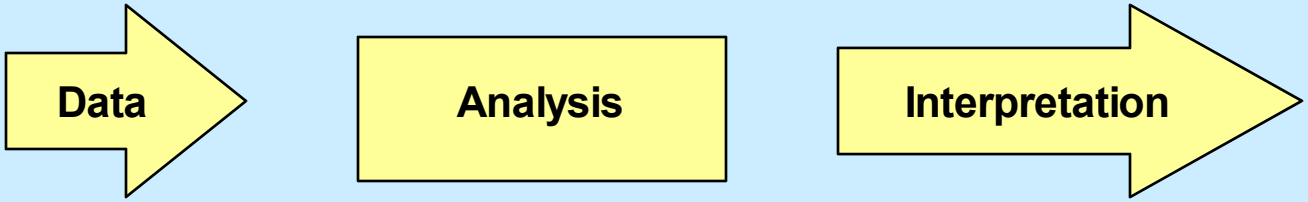
Measurement error: Most statistical models assume error free measurement. However, software processes are complex, making it difficult to measure precisely.

Problems with Interpretation (1)

Generally, the reason for analyzing process data is to draw inferences that can guide decisions and actions.

Drawing such inferences from data depends not only on using appropriate analytical methods and tools.

Equally important is the understanding of the underlying nature of the data and the appropriateness of assumptions about the conditions and environments in which the data were obtained.



Problems with Interpretation (2)

Confusion over significance: **“Significance” in the statistical sense is not the same as “significance” in the practical sense.**

Precision and accuracy: **Precision refers to how finely an estimate is specified, whereas accuracy refers to how close an estimate is to the true value.**

Causality: **Assessing causality is the reason of most statistical analysis. Evidence of causality involves:**

- **Concomitant variation.**
- **Time order of occurrence.**
- **Absence of other casual factors.**

How Do We Make Inferences?

We live in a world of self-generating beliefs which remain largely untested.

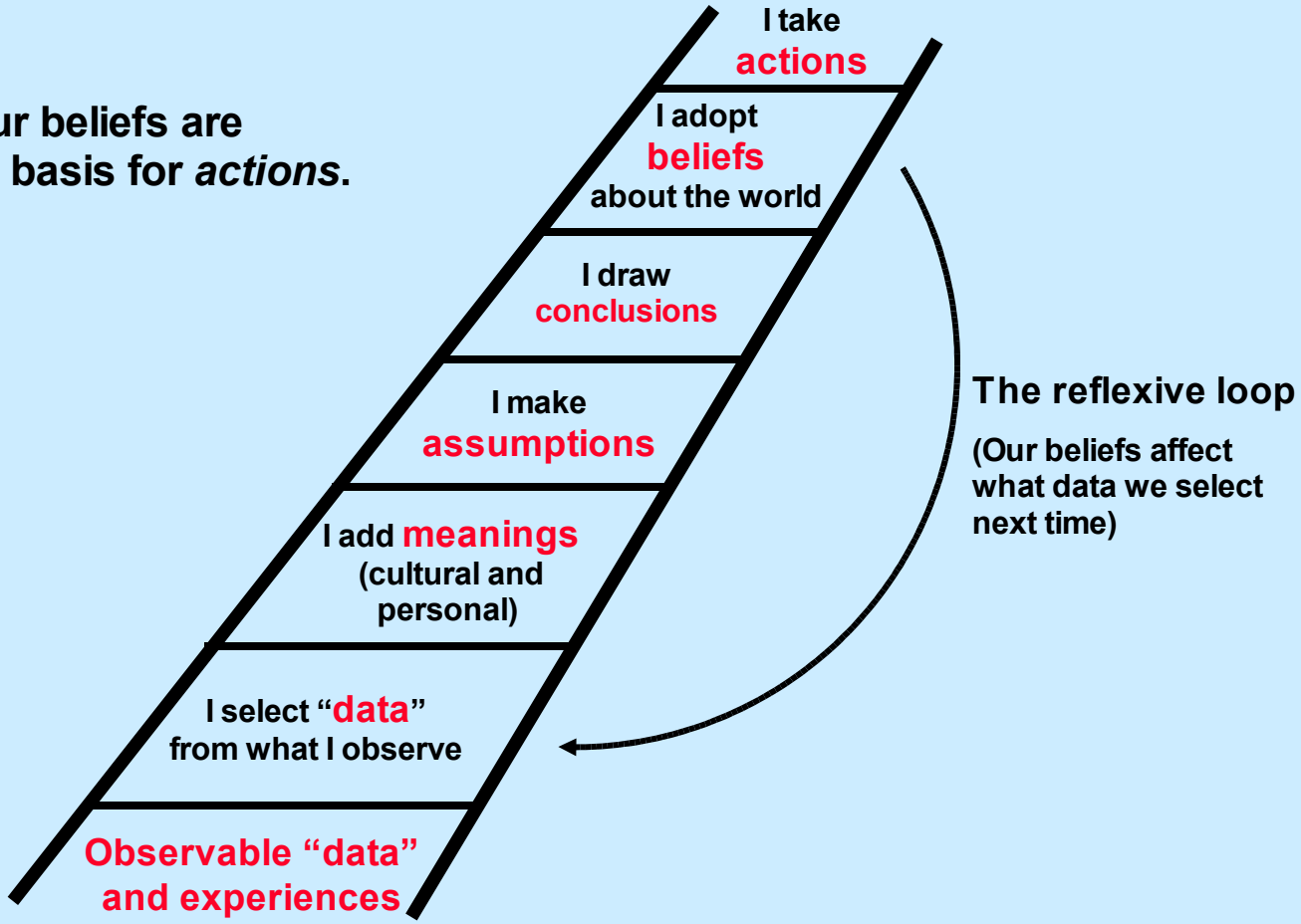
We adopt those beliefs because they are based on conclusions, which are inferred from what we observe, plus our past experiences.

Our ability to make "valid" inferences and achieve results is eroded by our feelings that:

- Our beliefs are *the* truth.
- The truth is obvious.
- Our beliefs are based on real data.
- The data we select are the real data.

The Ladder of Inference

Your beliefs are the basis for *actions*.



The reflexive loop

(Our beliefs affect what data we select next time)

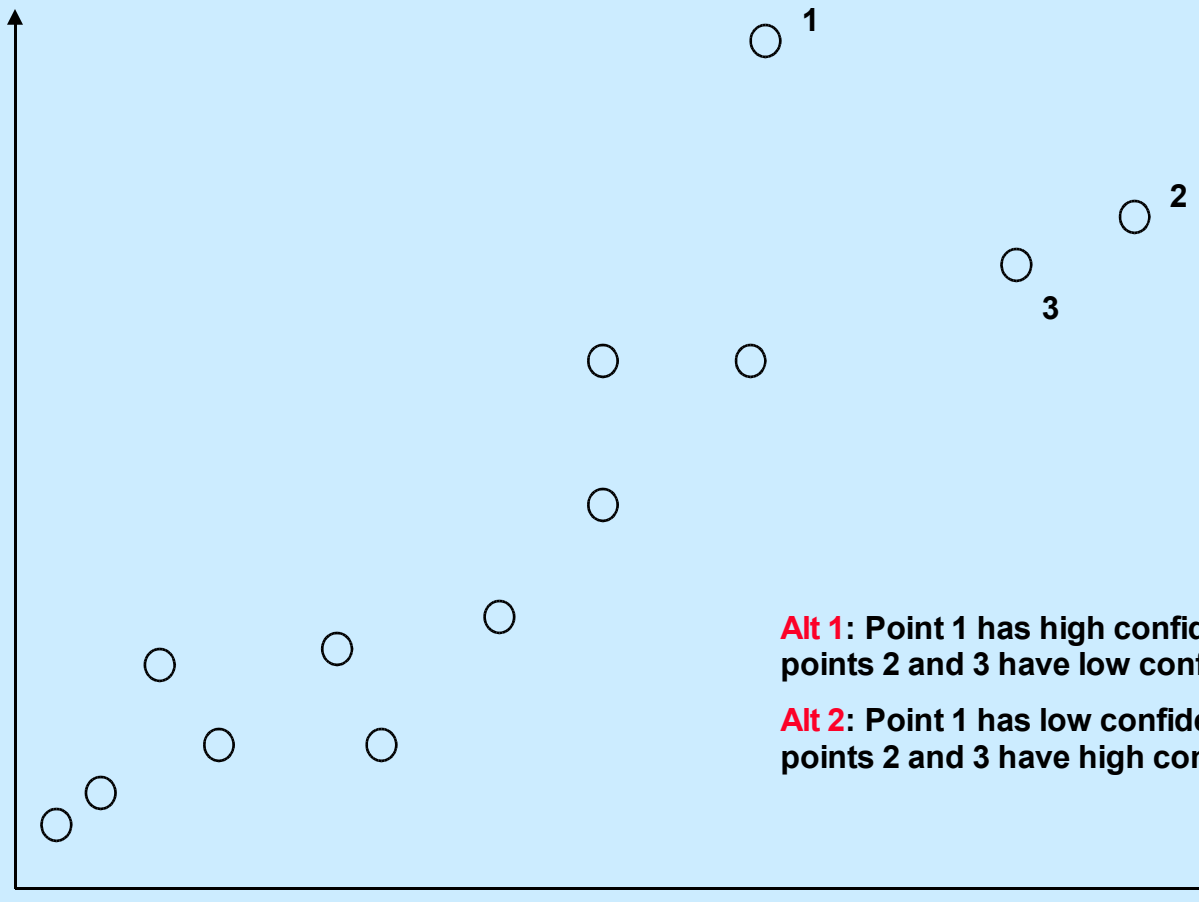
Choosing a Model

Choosing a model - that is, how we believe the data to be related - will influence all later analyses.

The decision to categorize one or more observations as “wrong” - or as “outliers” - is a relationship between data and model which cannot be decided upon without some form of confidence (belief) about the data and the model.

Expert opinion can be used as a method of triangulation to get subjective data to further guide the decision on choosing the “right” model.

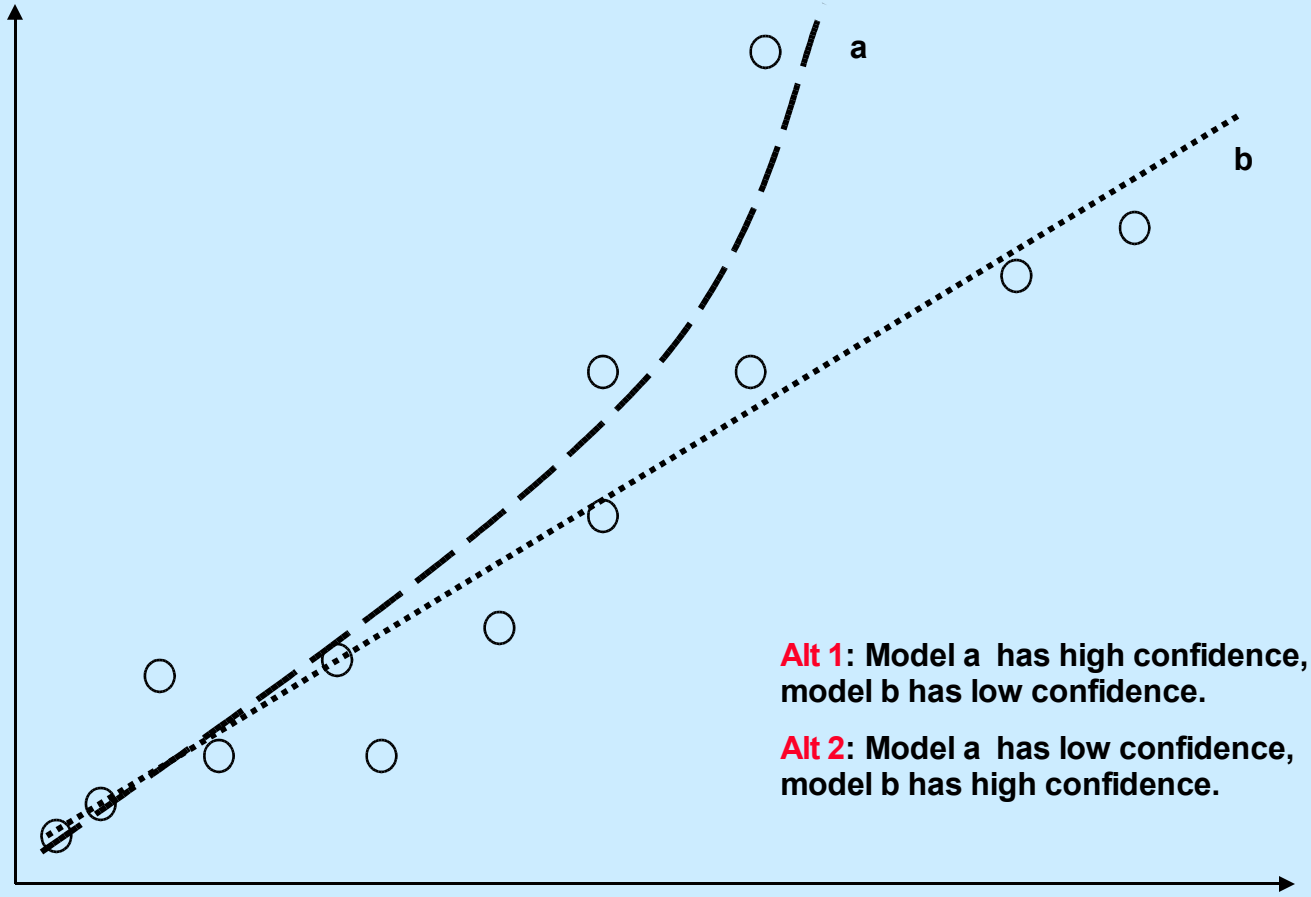
Is the Model Correct? (1)



Alt 1: Point 1 has high confidence, points 2 and 3 have low confidence.

Alt 2: Point 1 has low confidence, points 2 and 3 have high confidence.

Is the Model Correct? (2)



Alt 1: Model a has high confidence, model b has low confidence.

Alt 2: Model a has low confidence, model b has high confidence.

Summary

Errors using inadequate data are much less than those using no data at all.
-Charles Babbage

Be sure your sample is representative of the population in which you're interested.

Be sure you understand the assumptions of your statistical procedures, and be sure they are satisfied.

Be sure you have the right amount of statistical power.

Be sure to use the best measurement tools available.

Beware of multiple comparisons.

Keep clear in your mind what you're trying to discover - look at magnitudes rather than p-values.

Don't confuse precision with accuracy.

Be sure you understand the conditions for causal inference.

Be sure your models are accurate and reflect the data variation clearly.