Richard W. Hamming

Learning to Learn

The Art of Doing Science and Engineering

Session 27: Unreliable Data

Bad data

The data you get is not as reliable as advertised

Simulation data is normally assumed to be accurate, but we cannot depend on data the way we expect

First, some examples of bad physical data…

Life Testing

Life testing – how reliable is equipment

Example: testing reliability of vacuum tubes

- Wanted 20 year life – expected failure cost = $1M
- Tubes available for testing 18 months before use

Why believe testing equipment is as reliable as what you are testing?

Life Testing (cont')

Tricks in Life Testing

- Raise temp 17 degrees C and item ages twice as fast
- Increase voltage and stress resistance breakdown
- Increase frequency and expose failures earlier

Hard to adequately test highly reliable equipment in a short time period

- Too anxious to get latest technology into the field
- Not enough time to test it right, but plenty of time to fix

Accurate Measurements

Never assume people, equipment can't make measurement mistakes

Take apart test equipment and recalibrate it

- Never trust the manufacturer on advertised accuracy

Just because a machine records data doesn't make it correct

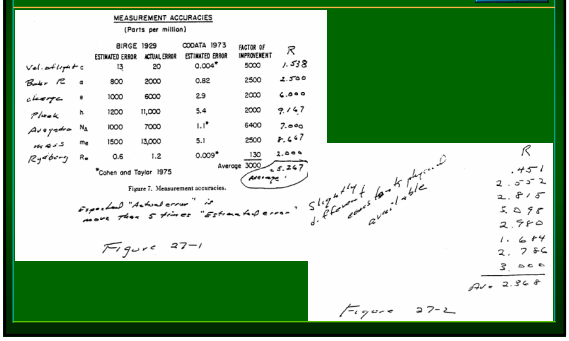
Never process data before checking for errors first

- it wastes more time to redo work

Test Studies

Never trust a pilot program (i.e. a test study) to accurately predict the success of the main study

Basic Scientific Data



MEASUREMENT ACCURACIES
(Parts per million)

		BIRGE 1929 ESTIMATED ERROR	BIRGE 1929 ACTUAL ERROR	CODATA 1973 ESTIMATED ERROR	FACTOR OF IMPROVEMENT	R
Velocity of light	c	13	20	0.004*	5000	1.538
Bohr radius	a	800	2000	0.82	2500	2.500
Electronic charge	e	1000	6000	2.9	2000	6.000
Planck constant	h	1200	11,000	5.4	2000	9.167
Avogadro number	N_A	1000	7000	1.1*	6400	7.000
Electronic mass	m_e	1500	13,000	5.1	2500	8.667
Rydberg constant	R_∞	0.6	1.2	0.003*	130	2.000
				Average	3000	6.267

*Cohen and Taylor 1975

Figure 7. Measurement accuracies.

Figure 27-1

Figure 27-2

Basic Scientific Data

Hamming examines increased accuracy of National Bureau of Standards physical constants

- Assuming new constants are correct, averaged 5 sigma for previous estimated error

Any number with a probable error is typically much smaller than it should be

- Scientists fine tune equipment to minimize variance, not error

90% of the time the next independent measurement will fall outside the previous 90% confidence limits

Fitting Data to a Model

There can be errors in both the model and the data

- Initially, the model is a guess and the data is bad
- As the data & measurements improve, the model is rarely reexamined for error
- Thus, as measurement accuracy increases, the error in the model increases

Estimates combined with reliable measurements are taken at the reliability of the best measurement

- "A chain is only as strong as its weakest link"

Economic Definitions

An economic definition today is not necessarily what it was yesterday.

- e.g. gold flow, income, inventory figures, poverty rate

Hard to draw conclusions from a time series when the definition of what is being measured constantly changes.

Often a changing definition is used to satisfy the needs of a desired policy (e.g. poverty rate) or outcome (e.g. unstated inventory to avoid taxes).

Economic Predictors and Indices

… change very slowly & lag behind reality

- Poorly represents current state of affairs
- Example: manufacturing vs. processing information

Worthy argument: it's better to have an index that is comparable and wrong, versus gaining the right index that is incomparable

Economists

Refuse to admit a data problem exists

Economists & governments typically skew figures to enact a particular policy

- (e.g. economic aid, gold holdings)

Surveys

A very small (1% or .1%) and accurate survey is more accurate than a 100% survey with the same amount of money.

Verbal surveys are very unreliable (biased based on wording of questions, order of questions, person who asks questions)

- Written surveys just as bad

Some questions just don't get answered honestly

- e.g. infidelity rates

Social statistics

Average person doesn't exist

- e.g. one breast, one testicle

Can only average homogenous data

- social data is rarely homogeneous

Management use of data

Surveys are answered quicker by those who think they will rate well than those who don't

- Presence of management will change data, surveys, observation, etc.
- Subordinates will skew data for the boss so as not to break bad news.

Keep all this in mind as you move through the ranks, and realize that the data that reaches you is often unreliable