Just How Capable is My Detection System, Really?

Victor W. Lowe, Jr.

Managing Knowledge

Context

- Detection Systems - measurement system plus decision rule- are becoming increasingly important
  - failed polygraph test leads FBI to call off warning of terrorist attacks on Las Vegas
  - K-9 detection systems
    - contraband
    - explosive
    - seizures
    - etc
  - C-130A airframe
- How do we / should we characterize the performance of a detection system

Examples

- Simple, yet dramatic
- Real data, readily available in the open literature
- Chosen to illustrate larger truths, which are alluded to but not spelled out in detail
- Suggest questions to ask about ***any*** detection
- **concepts presented generalize to to detection systems**
- Illustrate common data traps that ensnare the unwary

“Knowing there’s a trap is the first step in evading it”

Duke Leto Atreides

Dune, 1965

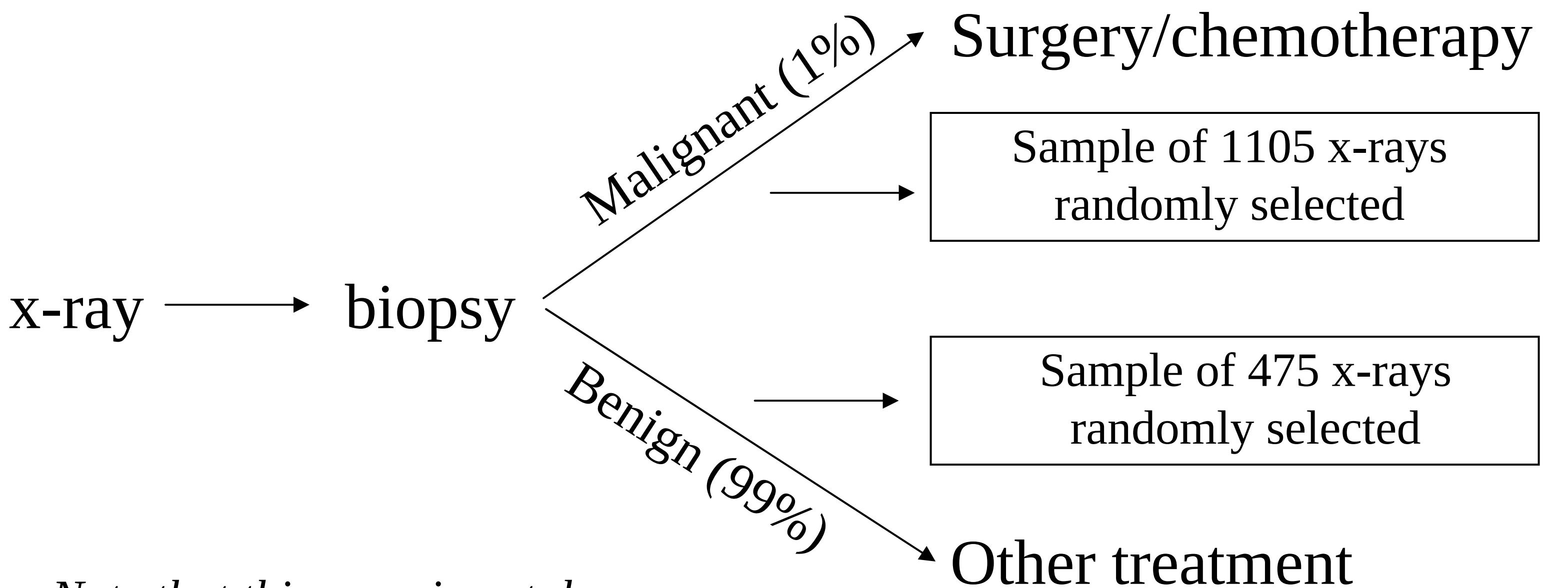
Frank Herbert

Real-world Example

- Data taken from ***Probabilistic-reasoning in clinical medicine: Problems and opportunities***, by David M. Eddy.
- Article appeared in ***Judgement under uncertainty: Heuristics and biases***, edited by David Kahneman, Paul Slovis, and Amos Tversky, Cambridge University Press, 1982
- Eddy uses data from Snyder, R. E. ***Mammography: Contributions and limitations***, published in Clinical Obstetrics and Gynecology, 1966, 9, 207-220.

Patients Suspected to Have Lesions Sent to University Clinic

(Snyder's data)



Note that this experimental design can be used with many other detection systems

X-Rays evaluated (by radiologists who did not know biopsy results) with the following results

False negative

Sensitivity of test

Biopsy Results	Radiologist Evaluation: malignant(+)	Radiologist Evaluation: benign(-)
malignant (M)	P(+ | M) = 0.792	P(– | M) = 0.208
benign (B)	P(+ | B) = 0.096	P(– | B) = 0.904

False positive

Note: P(M) = 0.01, P(B) = 0.99

Specificity of test

Note: ***data in table describes the performance of the measurement system.***

But the patient and clinician really want to know how good the ***detection system*** is ...

		Radiologist Evaluation	
		malignant(+)	benign(-)
Biopsy Results	malignant (M)	P(M | +) = ?	P(M | -) = ?
	benign (B)	P(B | +) = ?	P(B | -) = ?

Note: P(M) = 0.01, P(B) = 0.99

An *ideal* detection system would like like this

Radiologist Evaluation

Biopsy Results	malignant(+)	benign(-)
malignant (M)	P(M | +) = 1.0	P(M | -) = 0.0
benign (B)	P(B | +) = 0.0	P(B | -) = 1.0

Note: P(M) = 0.01, P(B) = 0.99

But the data we have looks like *this*, which Eddy asked physicians to evaluate

	malignant(+)	benign(-)
malignant (M)	P(+ | M) = 0.792	P(– | M) = 0.208
benign (B)	P(+ | B) = 0.096	P(– | B) = 0.904

Note: P(M) = 0.01, P(B) = 0.99

- Physician agreed that their clinical observations were consistent with Eddy's data: ~ 95% of the physicians estimated P(M | +) = ~ .75
- did not know P(M | +) 🕒 P(+ | M)

The mathematical way to determine how this test would perform in a clinical setting: Bayes Theorem

Note base rate (pointing to P(M) in the numerator)

$$P(\,M \mid +\,) = \frac{P(\,M\,)\,P(\,+ \mid M\,)}{P(\,M\,)\,P(\,+ \mid M\,) + P(\,B\,)\,P(\,+ \mid B\,)}$$

$$= \frac{(.01)(.792)}{(.01)(.792) + (.99)(.096)}$$

$$= 0.077$$

Eddy's study showed that ~ 95% of the physician estimated the number to be ~ .75

Data Trap 1

Working on the wrong problem!

- Not knowing the difference between the performance of the measurement system and the detection system
- Not knowing the difference between P(M | +) and P(+ | M)
- Measurement system is not the detection system

An intuitive view of the study:

"+" = malignant x-ray ;   "-" = benign x-ray

Numbers scaled to reflect base rate

100 white (malignant) balls in urn

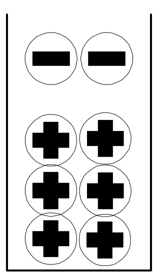
9900 grey (benign) balls in separate urn

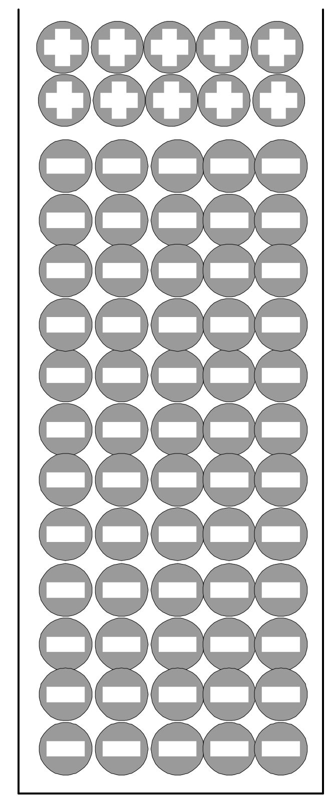
79 "+"'s
21 "-"'s

960 "+"'s
8940 "-"'s

P(+ | M) = 79 / 100

P(+ | B) = 960 / 9900





Thus, the study looked like this ...

Top row is modeled by urn with white balls,

bottom row by urn with grey balls

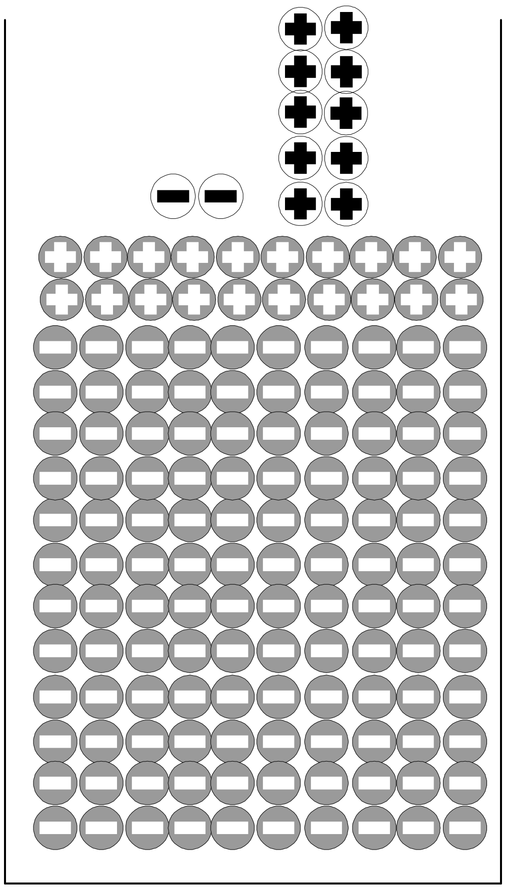
Radiologist Evaluation

Biopsy Results		malignant(+)	benign(-)
	malignant (M)	P(+ | M) = 79 / 100	P(– | M) = 21 / 100
	benign (B)	P(+ | B) = 960 / 9900	P(– | B) = 8940 / 9900

Note: P(M) = 100 / 10,000; P(B) = 9900 / 10,000

Patient Interested in P(M|+), Not P(+|M)

1. All 10,000 balls are in one urn
2. One ball is chosen at random
3. The ball has a "+" on it; what is the probability that it is grey?



$$P(\ M \mid +\ ) = \frac{\text{Number of white balls with +}}{\text{Total number of balls with +}}$$

$$= \frac{79}{79 + 960}$$

$$= 0.076 \quad (\ = 0.77 \text{ with round off }\ )$$

Managing Knowledge

The "Complete" table of interest derived from Snyder's data

Biopsy Results		Radiologist Evaluation: malignant(+)	Radiologist Evaluation: benign(-)
	malignant (M)	P(M | +) = .077	P(M | -) = .0023
	benign (B)	P(B | +) = .923	P(B | -) = .9977

Note: P(M) = 0.01, P(B) = 0.99

Suppose a better measurement system was available

Note improvement over original test

Radiologist Evaluation

Biopsy Results	malignant(+)	benign(-)
malignant (M)	P(+ | M) = 0.99	P(– | M) = 0.01
benign (B)	P(+ | B) = 0.01	P(– | B) = 0.99

Note: P(M) = 0.01, P(B) = 0.99

In clinical setting for same population of patients, the “better” measurement system would perform thusly ...

$$P(M \mid +) = \frac{P(M)\,P(+ \mid M)}{P(M)\,P(+ \mid M) + P(B)\,P(+ \mid B)}$$

Note base rate is unchanged

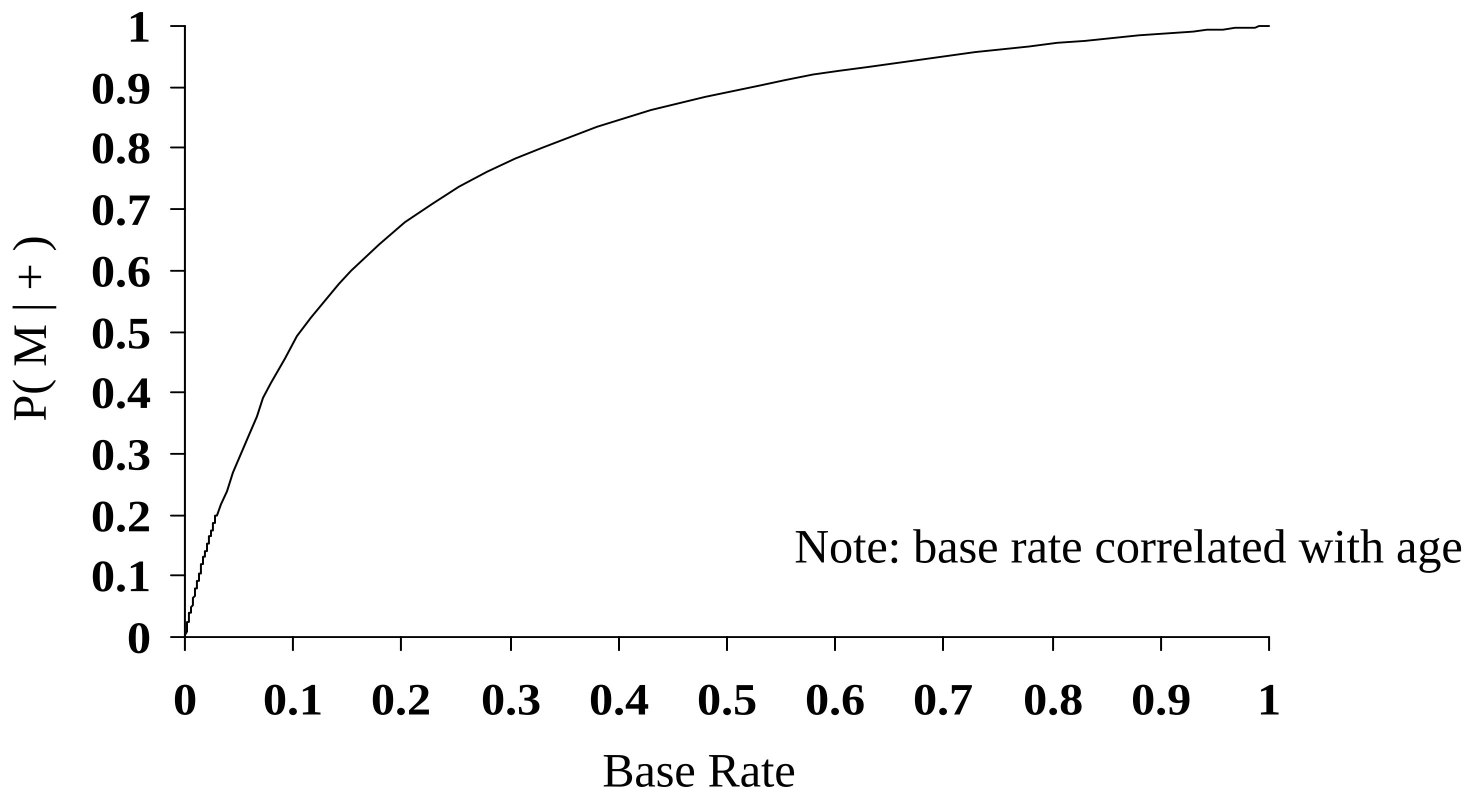
$$= \frac{(.01)(.99)}{(.01)(.99) + (.99)(.01)}$$

$$= 0.50$$

← Better, but still not great

When would <u>this</u> detection system be useful?

(i.e. for what base rate would it be useful?)



How good does a measurement system have to be for a given base rate?

(what's Z values do I need?)

Note:symmetry not required, but it does simplify things nicely

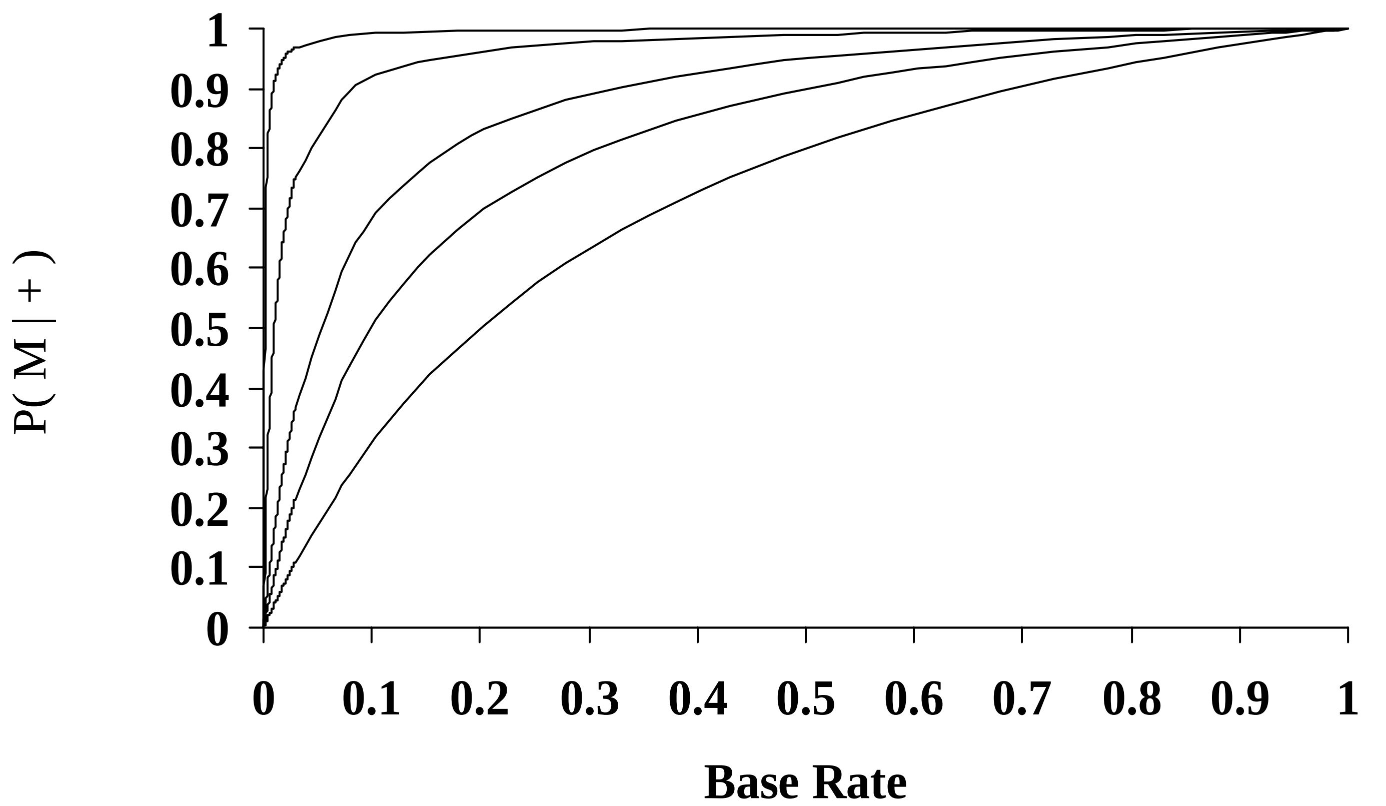
Radiologist Evaluation

Biopsy Results	malignant(+)	benign(-)
malignant (M)	P(+ | M) = 1 - Z	P(– | M) = Z
benign (B)	P(+ | B) = Z	P(– | B) = 1 - Z

Note: P(M) = 0.01, P(B) = 0.99

Comparison of detection systems

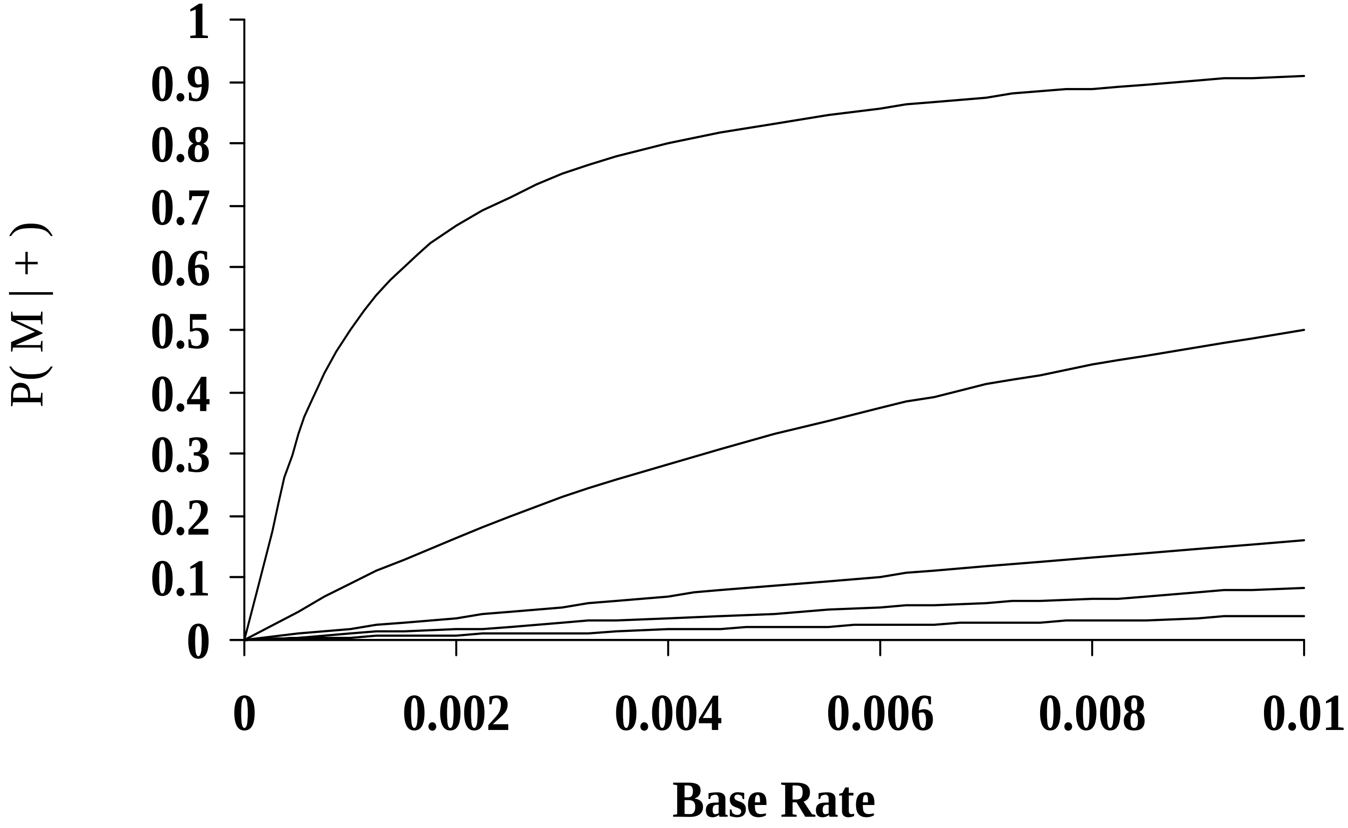
Z = 0.001, 0.01, 0.05, 0.1, 0.2



Managing Knowledge

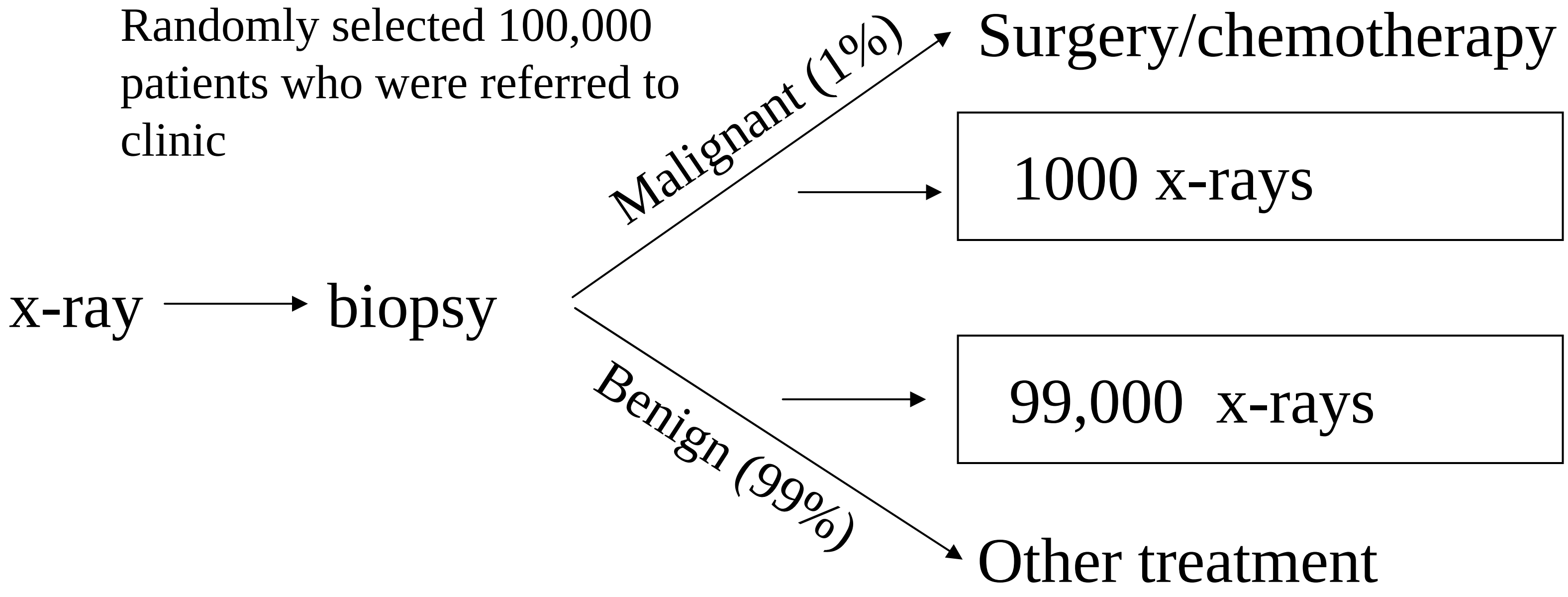
Comparison over a smaller region

Z = 0.001, 0.01, 0.05, 0.1, 0.2



Suppose data from the same population had been collected using a **new** experimental design

Randomly selected 100,000 patients who were referred to clinic



As before, X-Rays evaluated by radiologists who did not know biopsy results

Radiologist Evaluation

Biopsy Results	malignant(+)	benign(-)
malignant (M)	P(+ & M) = 792 / 100,000	P(– & M) = 208 / 100,000
benign (B)	P(+ & B) = 9504 / 100,000	P(– & B) = 89496 / 100,000

Data modeled differently to reflect the new experimental design

Note: P(M) = 1000/100,000 = 0.01

P(B) = 99,000/100,000 = 0.99

The correct way to analyze data from the new experimental design:

$$P(\ M \mid +\ ) = \frac{P(\ +\ \&\ M\ )}{P(\ +\ )}$$

Note base rate not required; it is already "baked" into the data

$$= \frac{792\ /\ 100{,}000}{792\ /\ 100{,}000 + 9504\ /\ 100{,}000}$$

$$= \quad 0.077$$

This result should be comforting; the same test applied to the same population yields the same results, but data analysis must be compatible with the way the sample was selected

The frightening part of the story

- Suppose the first analyst, thinking that the *real* data had been obtained by the second experimental design, analyzed it accordingly ...

$$P(M \mid +) = \frac{P(+ \,\&\, M)}{P(+)}$$

$$= \frac{0.792}{0.792 + 0.096} = \mathbf{.892}$$

A pretty good reason to elect surgery

Data Trap 2

Right problem, wrong data

- Not knowing the difference between P(+ & M) and P(+ | M)

Let’s emphasize that again!

- life-or-death decision making situation
  - properly analyzed, the data say
    - $P(M | +) = .077$, very weak reason to elect surgery
  - improperly analyzed, the data appear to say
    - $P(M | +) = .892$, very strong reason to elect surgery
- Suppose the data analyst does not know how the data to be analyzed was collected. What statistical inferences can be legitimately drawn from the data?
  - Fundamental question for data miners

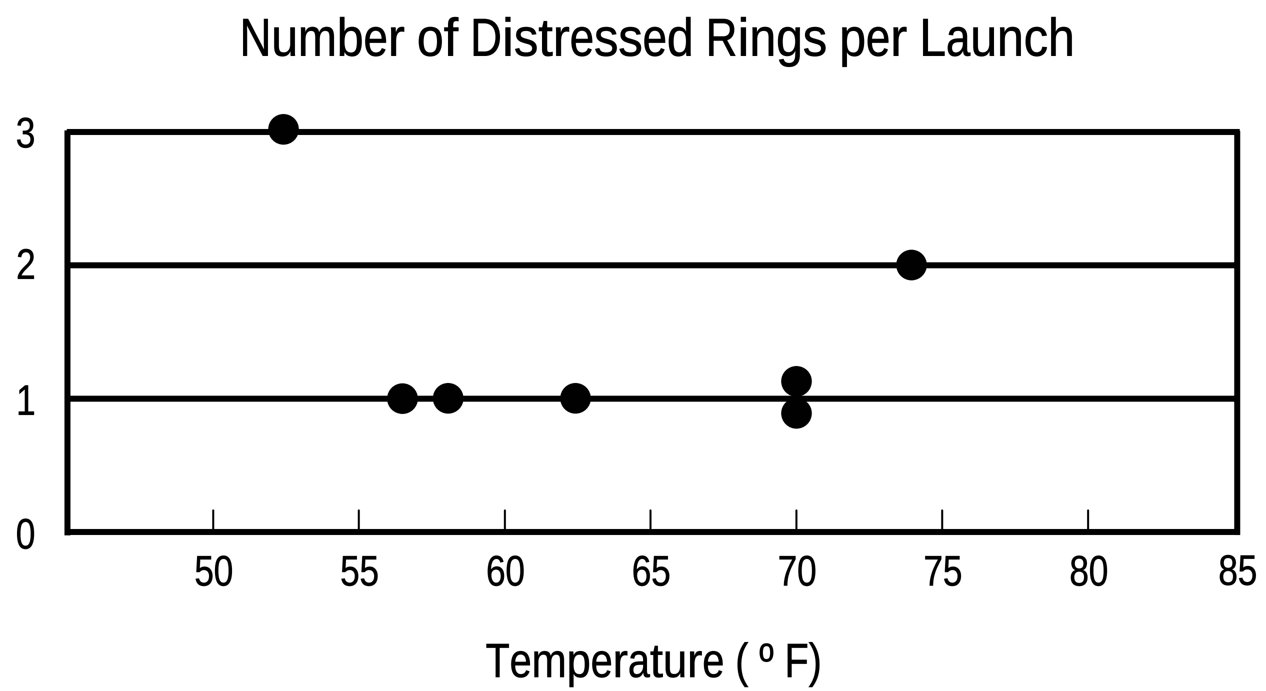
Points to Ponder

- Simple statistical techniques no help here
  - appropriate statistical techniques often much more mathematically sophisticated than the ones illustrated here
  - may not be possible to apply statistical techniques without some technical knowledge
- to be able to analyze the data, you ***must*** know how it was gathered (which urn it was sampled from)
  - many people, including professionals, don't understand that, they don't seek to obtain the appropriate data

More Points to Ponder

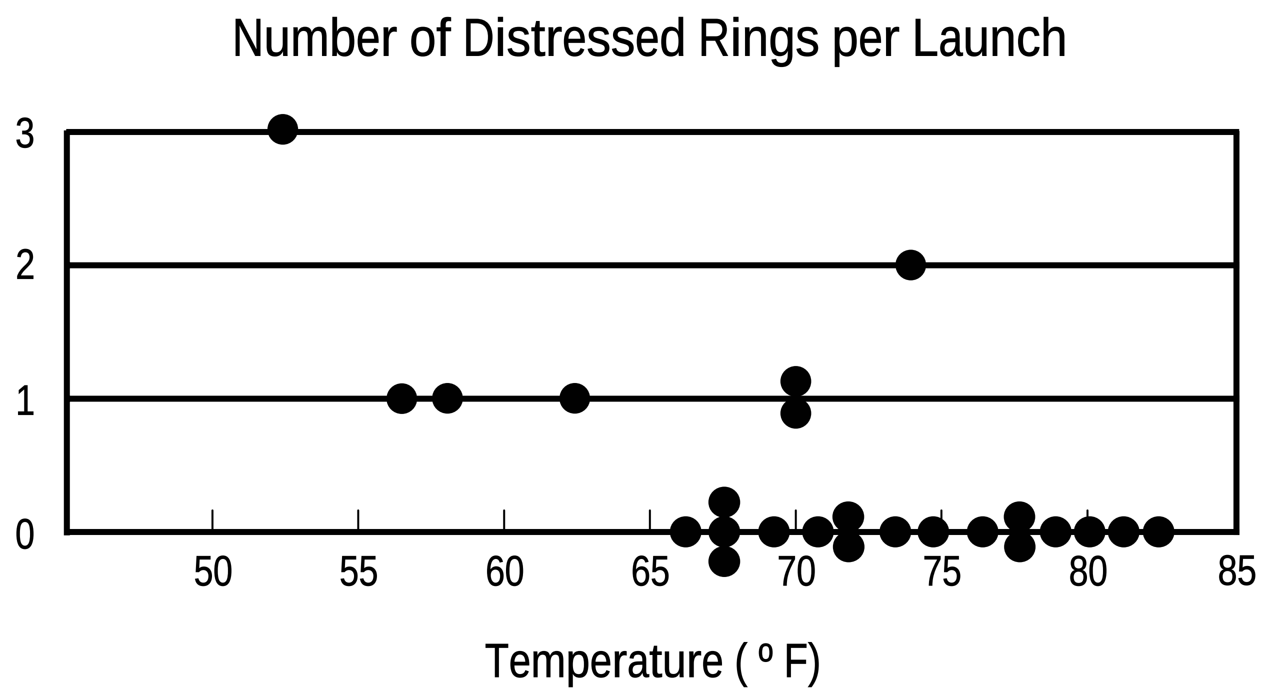
- performance of detection system (P(M | +), P(B | -)) depends on base rate; performance of measurement system doesn't
- detection system specifications cannot be well specified without anticipating base rate
  - engineering
  - purchasing
  - social determinations
    - good vs bad employees; detecting child abuse, etc.
- base rate, and therefore the usefulness of the measurement system, may change over time

Actual Data Used for Decision to Launch the Challenger



Managing Knowledge

Actual Data That Could / Should Have Been Used for the Decision



Summary Points

- "There is no substitute for Knowledge"
  - W. Edwards Deming
- Conditional probabilities want to be your friends. Be nice to them. Don't ignore them. They can help you.
- Techniques used to analyze data ***MUST*** be compatible with the data collection procedure
- Bad statistical thinking can kill

“Knowing there’s a trap is the first step in evading it”

Duke Leto Atreides

Dune, 1965

Frank Herbert

Backup Information

		+	-	
Population Model	M	Count 79	Count 21	100
	B	Count 960	Count 8940	9900
		1039	8961	10000

Eddy's Data Collection	M	P(+ | M) 0.79	P(- | M) 0.21	1
	B	P(+ | B) 0.0969697	P(- | B) 0.9030303	1

Natural Data Method	M	P(+ & M) 0.0079	P(- & M) 0.0021	0.01
	B	P(+ & B) 0.096	P(- & B) 0.894	0.99
		0.1039	0.8961	1

Desired Table	M	P(M | +) 0.07603465	P(M | -) 0.00234349
	B	P(B | +) 0.92396535	P(B | -) 0.99765651
		1	1

	Formula For Eddys Data Collection	Formula For Natural Data
Eddy's Data Collection Method	0.076034649	0.890673044
Natural Data Collection Method	0.000830539	0.076034649

Managing Knowledge