Calibration

1

How to tell if a probability judgment is "right"

- Coherence:
  - Are a set of judgments internally consistent?
  - Do the judgments follow the principles of probability theory?
- Calibration:
  - Do probability judgments correspond to actual events in the world?
  - E.g., on the all the days I said 80% chance of rain, did it actually rain on 80% of those days?

2

Examples of incoherence

- Saying p(rain) = .80 and p(no rain) = 0.50
- Saying it's more likely that Linda is a feminist bankteller than that she's bankteller.

3

Calibration

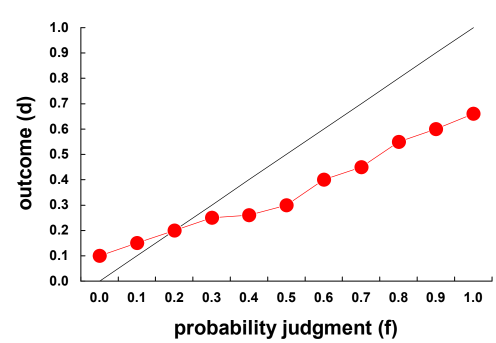
- Whether probability judgments correspond to actual outcomes
- Can't assess this for a single probability judgment
- But can do so for a group of probability judgments

4

Weather Forecasting Example

- Each day for 1,000 days the forecaster gives a probability of rain
- Record on each day whether it rained.
- Divide the days up into "bins" according to the prediction made
  - 0%, 10%, 20%, 30%, etc.
- For each bin, calculate the percentage of days on which it rained.

5



6

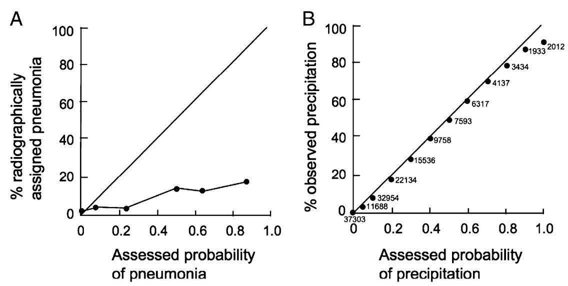


Fig. 5. Illustration of two extremes in expert calibration. (*A*) Assessment of probability of pneumonia (based on observed symptoms) in 1,531 first-time patients by nine physicians compared with radiographically assigned cases of pneumonia as reported by Christensen-Szalanski and Bushyhead (44). (*B*) Once-daily US Weather Service precipitation forecasts for 87 stations are compared with actual occurrence of precipitation (April 1977 to March 1979) as reported by Charba and Klein (43). The small numbers adjacent to each point report the number of forecasts.

7

Overconfidence

- Probability judgment is larger than the outcome index
  - E.g., when you say 90% chance of rain, it actually rains 65% of time.
- Calibration curve is below of the diagonal line

8

Probability Score

$$PS = (f - d)^2$$

- f = probability judgment
- d = outcome: 0 or 1
- Lower score is better (like golf)

9

Mean Probability Score

$$\overline{PS} = \frac{\sum (f - d)^2}{N}$$

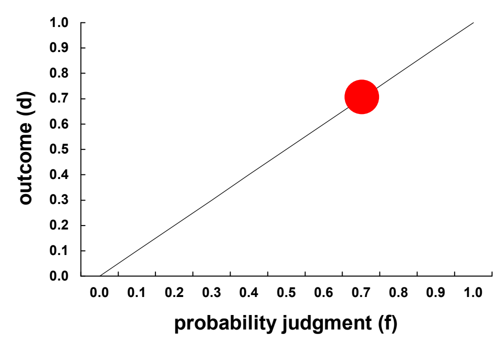
- Score of 0 = perfect calibration
- What's the maximum (worst) score you could get?

10

Base Rate Judge

- I know nothing about weather forecasting.
- I find out that it rains 70% of days in Pittsburgh.
- So, everyday, I predict 70% chance of rain.
- It does in fact rain on 70% of days

11



12

Base Rate Judge

- My calibration curve falls on the diagonal.
- But my judgments are uninformative.
- mean PS = $(.7)(1-.7)^2 + (.3)(0-.7)^2$
- =.7*.3*.3 + .3*.7*.7
- =.3*(.7*.3 + .7*.7)
- =.3*(.7*(.7 + .3))
- =.3*(.7) = 0.21

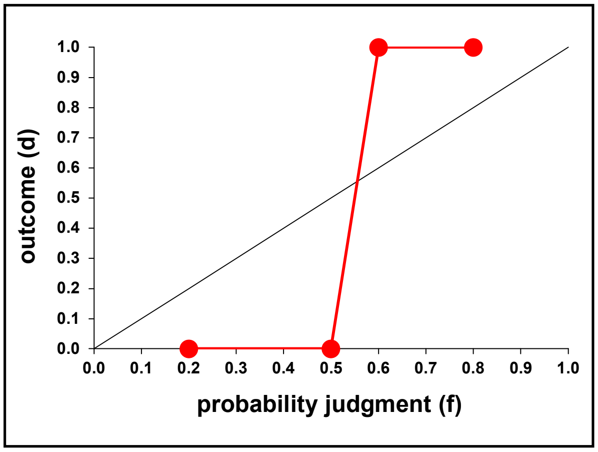
$$VI = \overline{d}(1-\overline{d})$$

13

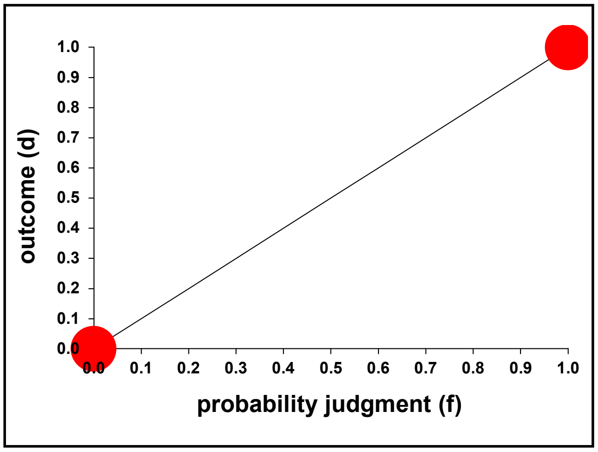
Discrimination

- Giving different probability judgments when the outcome occurs than when it does not occur.
- Data points near the top and bottom of the calibration graph = good discrimination
- Data points on the diagonal line = good calibration

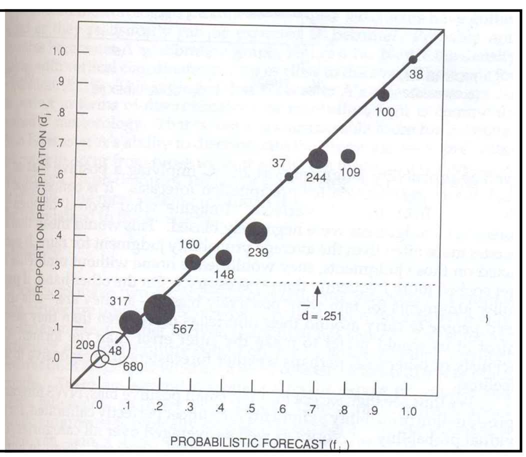
14



15



16



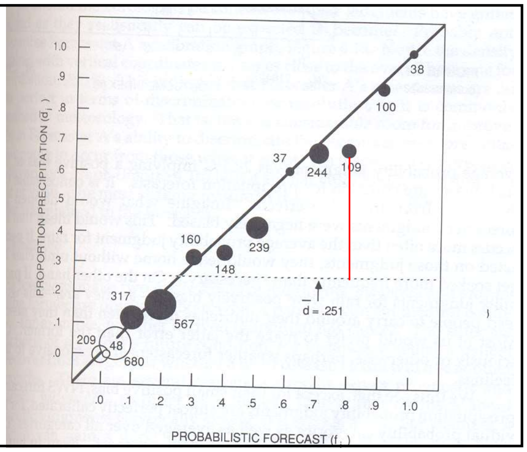
17

Discrimination Index

$$DI = \frac{\sum N_j(\overline{d}_j - \overline{d})^2}{N}$$

- For each judgment "bin", compare the percentage of outcomes to the overall baserate.
- High score is good – you want a big difference.
- Corresponds to distance between dot and baserate line

18



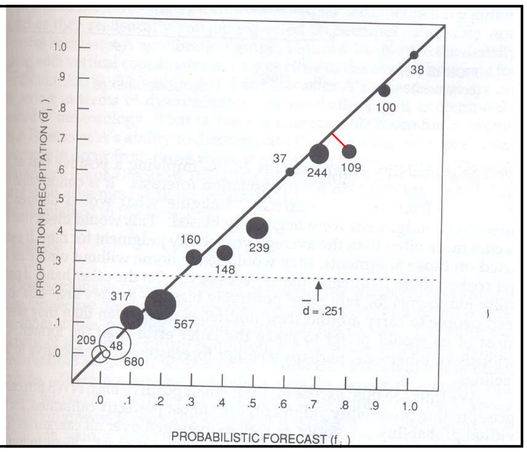
19

Calibration Index

$$CI = \frac{\sum N_j (f_j - \bar{d}_j)^2}{N}$$

- For each judgment "bin", compare the percentage of outcomes to predicted percentage of outcomes.
- Low score is good
- Corresponds to distance between dot and diagonal line

20



21

Decomposition

$$\overline{PS} = VI + CI - DI$$

Where VI (variability index) is:

$$VI = \bar{d}(1 - \bar{d})$$

22