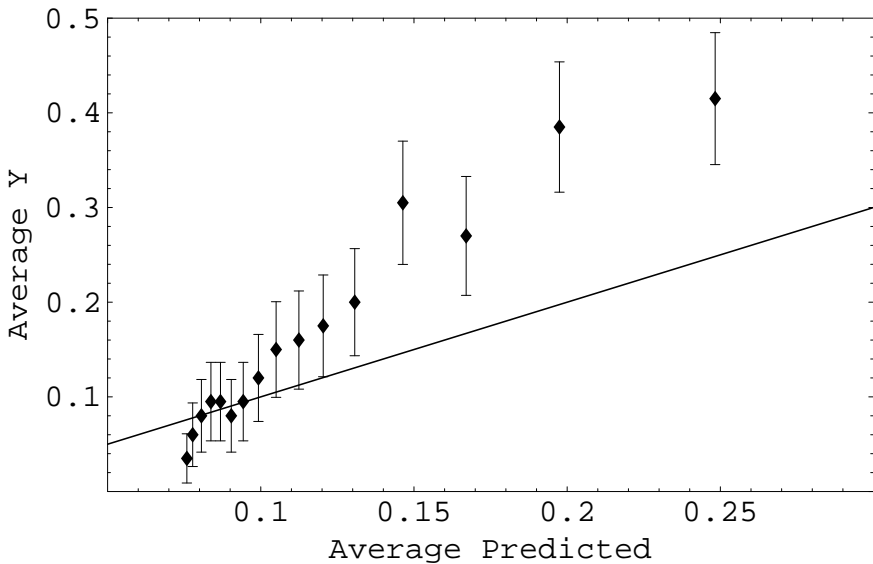


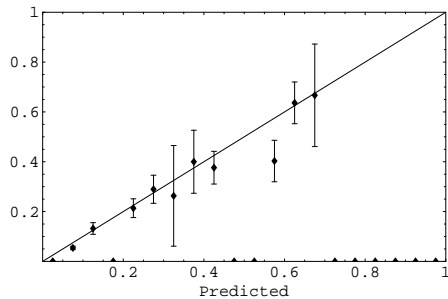
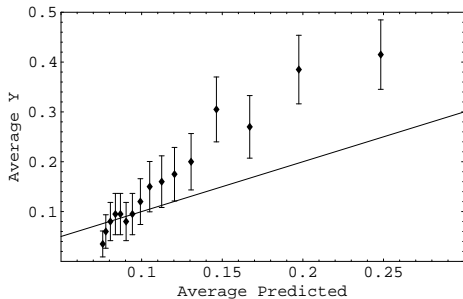
## Calibrating

Dean P. Foster

# This is not calibrated



# Anything easily fixed isn't calibrated



Fix the obvious problems!

# Calibration is unbiasedness

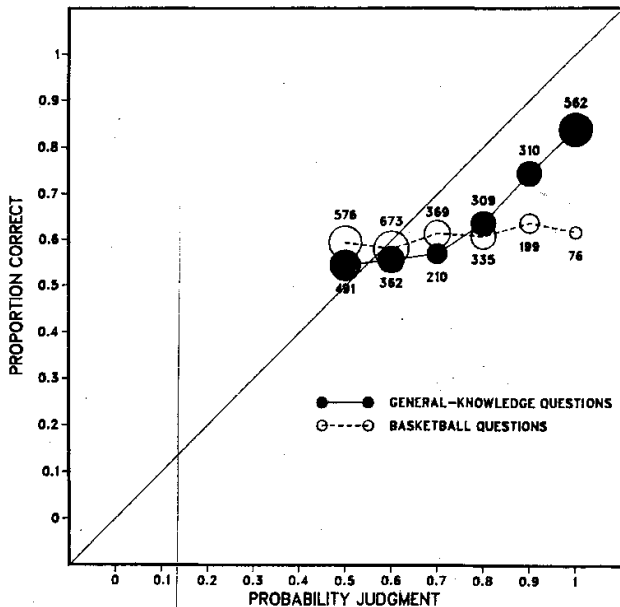
- Simple unbiasedness:  $E(Y - \hat{Y}) \approx 0$ .

# Calibration is unbiasedness

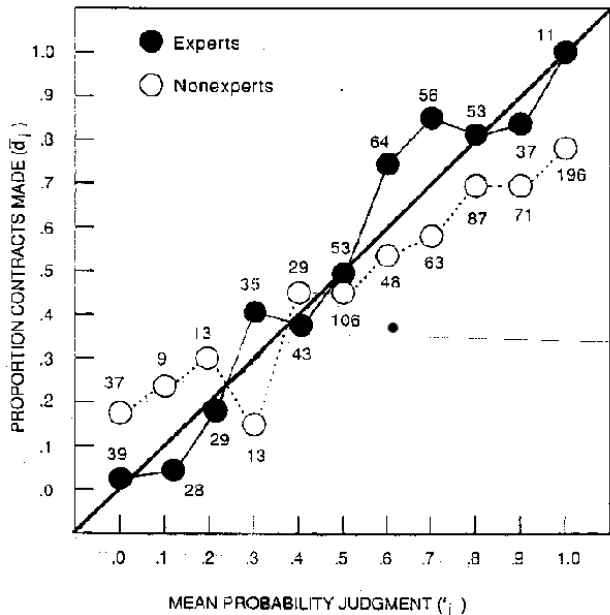
- Simple unbiasedness:  $E(Y - \hat{Y}) \approx 0$ .
- We want more:

$$E(Y - \hat{Y} | \hat{Y} \approx c) \approx 0$$

# Human behavior: without incentives



# Human behavior: With incentives!



# Calibration theory: example

Calibration is a minimal condition for performance

- On sequence: 0 1 0 1 0 1 0 ...
- The constant forecast of .5 is calibrated
- The constant forecast of .6 is not calibrated
- The variable forecast of .1 .9 .1 .9 .1 .9 ... is not calibrated



# Calibration theory: example

Calibration is a minimal condition for performance

- On sequence: 0 1 0 1 0 1 0 ...
- The constant forecast of .5 is calibrated
- The constant forecast of .6 is not calibrated
- The variable forecast of .1 .9 .1 .9 .1 .9 ... is not calibrated
  - But the forecast .1 .9 .1 .9 .1 .9 ... is pretty good!
  - Yes, it has better “refinement.”
  - But, it isn't calibrated.
  - Our goal: Keep this refinement, but make it calibrated

# Calibration is achievable

## Theorem

*A calibrated forecast exists.*

# Calibration is achievable

## Theorem

*A calibrated forecast exists.*

### **proof:**

Apply mini-max theorem.

(Sergiu Hart)

# Calibration is achievable

## Theorem

*A calibrated forecast exists.*

### Detailed proof:

- Game: between the statistician and Nature.
  - Nature's strategy is a distribution over sequences of rain. (A distribution over the  $2^T$  sequences.)
  - The statistician's strategy is a forecasting function. (A function mapping  $2^T$  to  $\{\epsilon, 2\epsilon, \dots, 1\}$ .)
  - This is a two person, zero sum game with a finite set of actions.
- If the statistician knew the process she could easily "win."
  - Compute  $E(X_t | X_1, \dots, X_{t-1})$
  - round to the nearest  $\epsilon$  grid point
  - Play that forecast
  - By LLN the empirical average is close to the forecast
- By the mini-max theorem the statistician can always win.

# Calibration exists: So what?

- Predicting the “grand average” is calibrated
  - But it is a crappy forecast.
- We have lots of ways of generating good forecasts:
  - probabilistic models
  - Time series: ARIMA, etc
  - on-line least squares regression
  - Combining experts
- None are guaranteed to be calibrated

# Calibration exists: So what?

- Predicting the “grand average” is calibrated
  - But it is a crappy forecast.
- We have lots of ways of generating good forecasts:
  - probabilistic models
  - Time series: ARIMA, etc
  - on-line least squares regression
  - Combining experts
- None are guaranteed to be calibrated

Goal: Find a way to convert these good forecasts into calibrated forecasts without removing their goodness.

# Bias / Variance decomposition

- bias:

$$\beta \equiv E(Y|\hat{Y}) - \hat{Y}$$

- variance:

$$\text{VAR} = \text{Var}(Y - E(Y|\hat{Y}))$$

- Mean Squared error:

$$\text{MSE} = E(Y - \hat{Y})^2 = E(\beta^2) + \text{VAR}$$

- For binary sequences:
  - Bias is called *Calibration*
  - Variance is called *Refinement*
  - MSE is called *Brier Score*

# Brier score

- Conditional expectation:

$$\rho(x) = \frac{\sum_t Y_t I_{\hat{y}_t=x}}{\sum I_{\hat{y}_t=x}}$$

- Bias:  $\beta(x) = \rho(x) - x$
- Brier score / MSE:

$$BS = \frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)^2$$

- Decomposition (MSE = bias + Variance):

$$\underbrace{\frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)^2}_{BS} = \underbrace{\frac{1}{T} \sum_{t=1}^T (\hat{Y} - \rho(\hat{Y}))^2}_{\text{Calibration}} + \underbrace{\frac{1}{T} \sum_{t=1}^T (Y_t - \rho(\hat{Y}_t))^2}_{\text{Refinement}}$$



# Defining calibeating

Calibration is fixable after the fact. But, can we fix it as we go along?

- Start with a forecast  $\hat{y}_t$ 
  - Calibration  $K(\hat{y})$
  - Refinement  $R(\hat{y})$
- Find a new forecast  $\tilde{y}_t$  that doesn't pay the calibration costs of  $\hat{y}$

## Definition (Calibeating)

$\tilde{y}$  calibeats  $\hat{y}$  if:

$$BS(\tilde{y}) \leq R(\hat{y}).$$

- $\tilde{y}$  keeps any patterns found by  $\hat{y}$
- $\tilde{y}$  doesn't "pay" the calibration error

# Calibeating many forecasters

We can extend this to calibeating many forecasters.

## Definition (Calibeating)

$\tilde{y}$  calibeats a collection of forecasts  $\{\hat{y}^1, \dots, \hat{y}^n\}$  if for all  $i$ :

$$\text{BS}(\tilde{y}) \leq R(\hat{y}^i).$$

# Calibrating is easy

- Consider a family of forecasts:  $\hat{y}_t^i$ 
  - Break up the interval  $[0, 1]$  into small buckets  $B_j$ .
  - Knowing which bucket each forecast is in is enough information to approximately compute the refinement of the forecast
  - Make these buckets into regression variables:

$$X_t^{ij} = I_{\hat{y}_t^i \in B_j}$$

- $\tilde{y}_t$  is generated by an on line regression:  $Y \sim X$ .

## Theorem

*The forecast combination  $\tilde{y}_t$  will  $\epsilon$ -calibeat  $\hat{y}_t^i$  if we use buckets with width less than  $\epsilon$ .*

# Calibeating is easy, but it can be calibeaten!

We can find  $\tilde{y}$  that calibeats  $\hat{y}$ . But, there is no reason for  $\tilde{y}$  to be calibrated. So it can be calibeaten. The result likewise isn't calibrated, so it can be calibeaten.

# Calibeating is easy, but it can be calibeaten!

We can find  $\tilde{y}$  that calibeats  $\hat{y}$ . But, there is no reason for  $\tilde{y}$  to be calibrated. So it can be calibeaten. The result likewise isn't calibrated, so it can be calibeaten.

- This can go on ad infinitum

# Stopping the infinite regress

We can have  $C_t$  calibeat  $A_t$  and  $B_t$ .

- Suppose at each time  $t$  we pick  $B_t = C_t$ .
- Requires a fixed point computation
- $C_t$  calibeats  $A_t$
- $C_t$  calibeats  $C_t$ :

$$BS(C_t) \leq R(C_t)$$

So  $C_t$  is calibrated.

## Theorem

*For any set of forecasts, there is a combination forecast which calibeats each element in the set, and is also calibrated.*

# Freebie: Calibrating yourself is calibrated

If we use this theorem with an empty set then  $C$  is calibrated:

## Corollary

*If  $C$  calibrates itself, then  $C$  is calibrated.*

## About fixed points

Suppose we will forecast  $C_t$ . The calibrating algorithm would say we should instead forecast  $g(A_t, C_t)$ . If this happens to be  $C_t$ , we are done. Ignoring  $A_t$  this means we want  $C_t = g(A_t)$ .



## About fixed points

Suppose we will forecast  $C_t$ . The calibrating algorithm would say we should instead forecast  $g(A_t, C_t)$ . If this happens to be  $C_t$ , we are done. Ignoring  $A_t$  this means we want  $C_t = g(A_t)$ .

### Theorem (Outgoing distribution)

*There exists a probability distribution on  $C$  such that:*

$$E(|x - C|^2 - |x - g(C)|^2) \leq \delta^2$$

*for all  $x$ .*

Proof is via the mini-max theorem (so linear programming can find the answer.)

- This means the BS of using  $C$  is better than the BS of using the correct answer  $g(C)$ .

# Tension between calibration and BS

- We know never to randomize when minimizing a quadratic loss function
- All three proofs (this mornings, mini-max and the outgoing distribution) involve randomization
- In fact, they involve LARGE randomizations, eg:

$$P(\hat{y}_t = .2) = P(\hat{y}_t = .5) = P(\hat{y}_t = .9) = 1/3$$

# Tension between calibration and BS

- We know never to randomize when minimizing a quadratic loss function
- All three proofs (this mornings, mini-max and the outgoing distribution) involve randomization
- In fact, they involve LARGE randomizations, eg:

$$P(\hat{y}_t = .2) = P(\hat{y}_t = .5) = P(\hat{y}_t = .9) = 1/3$$

## Theorem (with Johnson 2013)

*Randomly rounding an exponential smooth to the nearest grid point is almost calibrated.*

But this is merely calibrated, and doesn't easily extend to calibrating arbitrary forecasts.

# True fixed points

## Theorem (Outgoing fixed point)

*For any smooth  $g(\cdot)$  and any closed convex set  $S$ , there exists a point  $C \in S$  such that:*

$$E(|x - C|^2 - |x - g(C)|^2) \leq 0$$

*for all  $x \in S$ .*

Proof is via the Brouwer's fixed point. In fact, it is equivalent to Brouwer's fixed point theorem.

## Theorem (Outgoing fixed point)

*For any smooth  $g(\cdot)$  and any closed convex set  $S$ , there exists a point  $C \in S$  such that:*

$$E(|x - C|^2 - |x - g(C)|^2) \leq 0$$

*for all  $x \in S$ .*

- Can create a deterministic “weak” calibration

# True fixed points

## Theorem (Outgoing fixed point)

*For any smooth  $g()$  and any closed convex set  $S$ , there exists a point  $C \in S$  such that:*

$$E(|x - C|^2 - |x - g(C)|^2) \leq 0$$

*for all  $x \in S$ .*

- Using rounding, it can create a local random calibrated forecast
  - Randomly round to nearest grid point
  - First few digits aren't random, just the least significant one
  - Need this minimal amount of rounding to avoid impossibility result mentioned this morning

# True fixed points

## Theorem (Outgoing fixed point)

*For any smooth  $g()$  and any closed convex set  $S$ , there exists a point  $C \in S$  such that:*

$$E(|x - C|^2 - |x - g(C)|^2) \leq 0$$

*for all  $x \in S$ .*

- Fixed points are hard to find
- Basically need to do exhaustive search at every time period
- CS people call complexity class PPAD

# Forms of calibeating

We've have four forms of calibeating:

simple	Distribution	local random	deterministic
LS or average	LP	Fixed point	Fixed point
calibrated	classic calibration	Both classic and weak	Weak
quadratic safe	Not quadratic safe	quadratic safe	quadratic safe



# Forms of calibeating

We've have four forms of calibeating:

simple	Distribution	local random	deterministic
LS or average	LP	Fixed point	Fixed point
calibrated	classic calibration	Both classic and weak	Weak
quadratic safe	Not quadratic safe	quadratic safe	quadratic safe

Thanks!