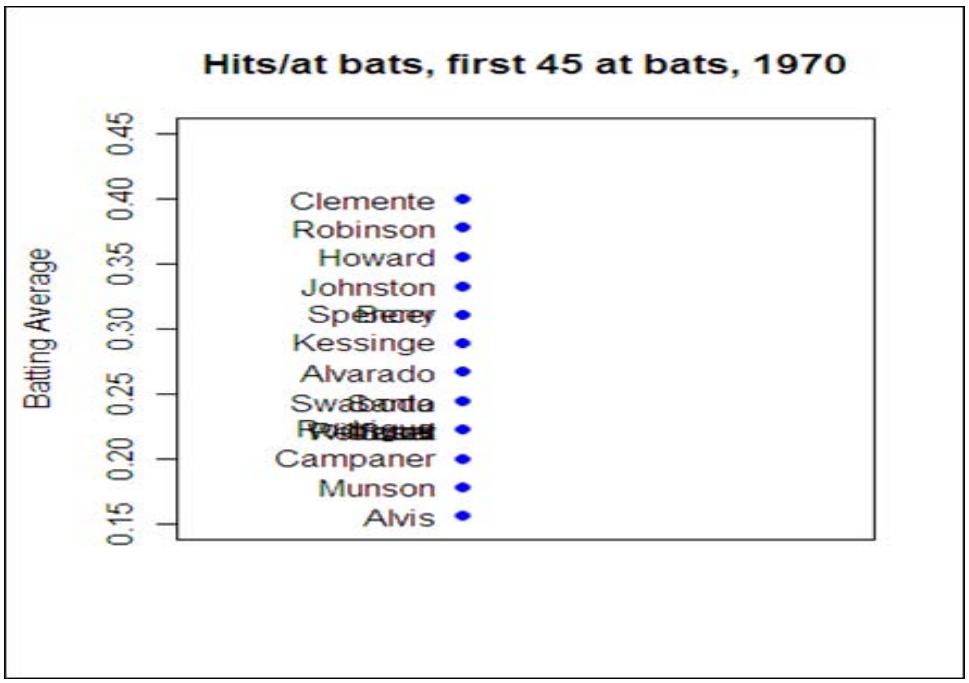
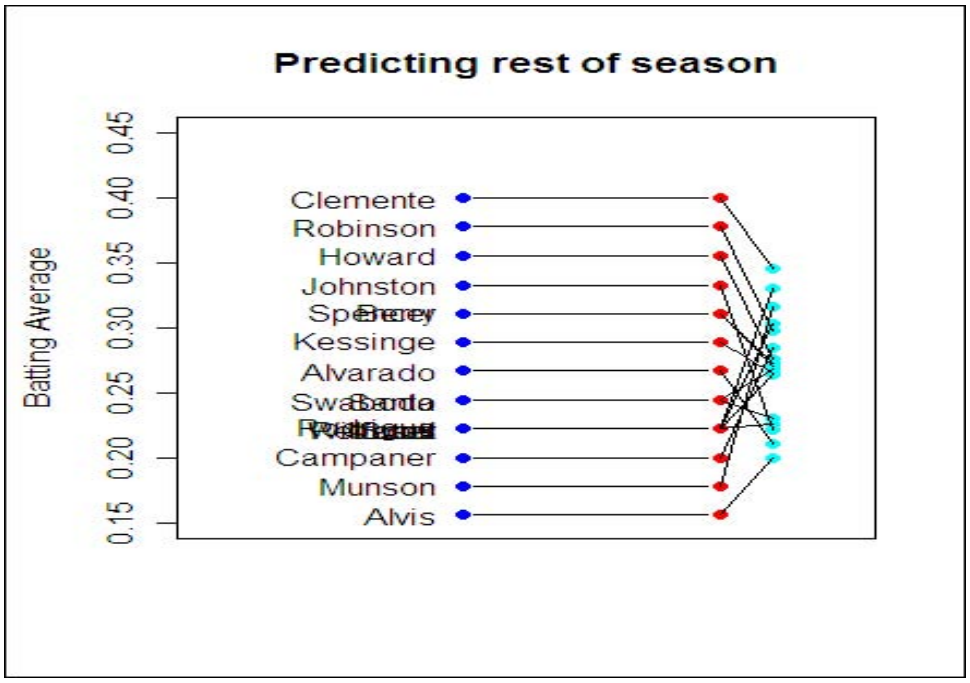
Key distinction between fixed and random effects:

- Estimate means of fixed effects
- Estimate variance of random effects

But in some instances, want to predict FUTURE values of a random effect

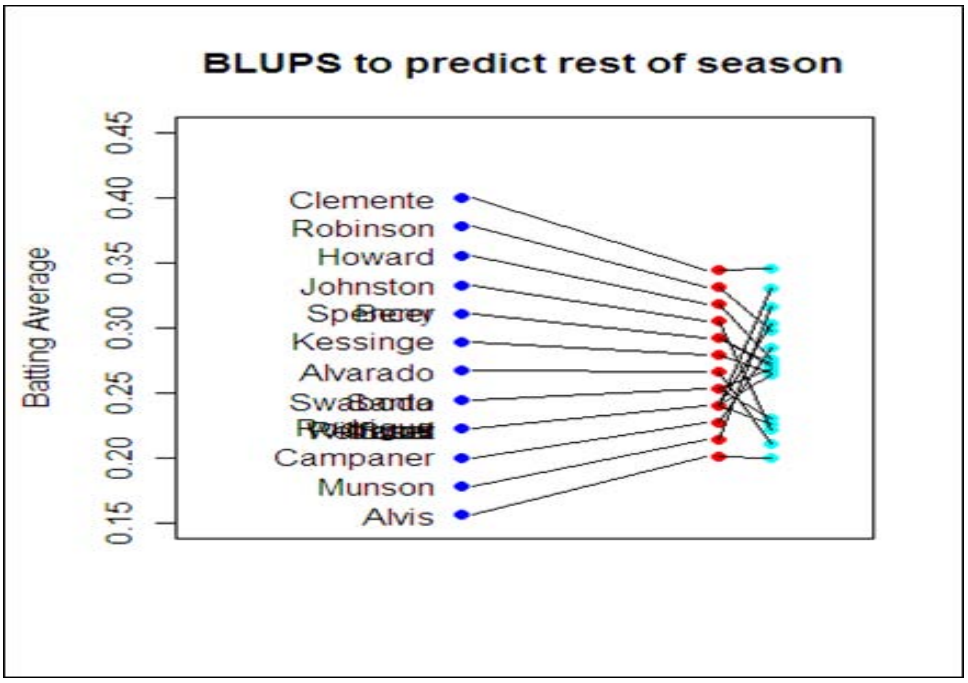
- Example (from Efron and Morris, 1975, JASA 70:311-319):
- US Baseball players. Given a player's performance in the beginning of the season, predict performance in rest of season.
- Following pictures illustrate the issues





Imagine knowing "the true" value for each player:

- If all the "true values" were identical, all the observed variation is random noise
  - The best predictor is the overall average - same value for all players
- If there were no random noise, the observed values are the "true values"
  - Best prediction is the observed value
- Reality is something in between
  - The true values are not identical, and there is random noise
  - Best predictor is found by "Shrinking" obs. performance towards overall mean.



- So, how much shrinkage is needed? How do we compute optimal predictor?
- Consider a simple mixed model - subsampling only

$$\begin{aligned} Y_{ij} &= \mu + b_i + \varepsilon_{ij} \\ b_i &\sim N(0, \sigma_b^2) \\ \varepsilon_{ij} &\sim N(0, \sigma^2) \ b_i \text{ and } \varepsilon_{ij} \text{ are independent} \end{aligned}$$

- Given data (i.e., all the y_{ij}), what is our best guess for the unobserved b_i ?
- Answer is the Best Linear Unbiased Predictor (BLUP)
  - linear function of observations
  - unbiased, averaged value of the estimate (or prediction) = true value
  - has a variance no larger than the variance of any other unbiased linear predictor
- Answer is the expected value of $\boldsymbol{\beta}_i$ given the data y_{ij}, $E(b_i|\boldsymbol{y})$.
- joint distribution of b_i and e_{ij} is multivariate normal for all i, j.

BLUP of b_i for subsampling model

- Have n_i observations for "subject" i, e.g., baseball player
- Compute $\overline{Y}_i$ and $\overline{Y}$, subject average and overall average
- Variance component for subjects = σ_b^2

- Variance component for observations within subjects = σ^2
- p_i = BLUP of b_i

$$\begin{aligned} p_i &= \overline{Y} + k_i(\overline{Y}_i - \overline{Y}) \\ k_i &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2/n_i} \end{aligned}$$

- Example, Roberto Clemente: $\overline{Y}_i = 0.4$, All players: $\overline{Y} = 0.265$
  - No variation among players: $\sigma_b^2 = 0$, $k_i = 0$, $p_i = 0.265 + (0)(0.4 - 0.265) = 0.265$
  - No error variation: $\sigma = 0$, $k_i = 1$, $p_i = 0.265 + (1.0)(0.4 - 0.265) = 0.4$
  - BLUP: $\hat{\sigma}_b^2 = 0.002854$, $\hat{\sigma}^2 = 0.0020$, $k_i = 0.588$, $\hat{p}_i = 0.265 + (0.588)(0.4 - 0.265) = 0.344$
  - Rest of season: observed batting average = 0.346
- For all 18 players:
  - players as fixed effect (use observed mean as the prediction): MSEP = $\sum(\hat{p}_i - o_i)^2/18 = 0.084$
  - BLUPs for players: MSEP = 0.041
  - BLUPs are twice as good (on variance scale) as the observed means

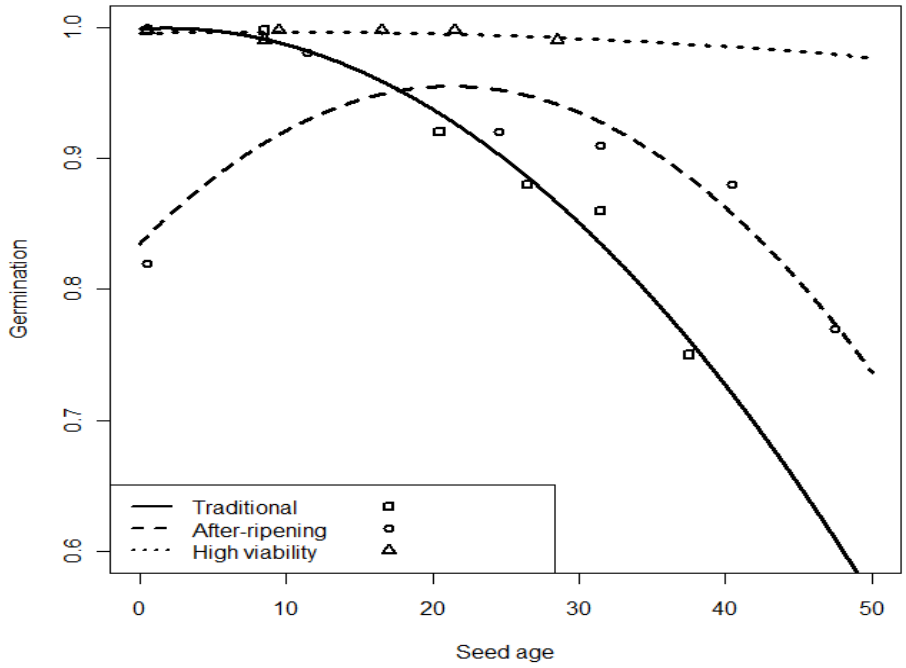
Notes:

- Theory based on known variances. Variances usually estimated,
  - Predictions more correctly called eBLUPs: estimated BLUPS
- When do I use means and when do I use BLUPs?
  - Means much more commonly used
  - BLUPs not well understood
    - "This wasn't in my statistic class" (taken 20+ years ago)
    - Exception: plant or animal breeding.
      BLUPs now widely used as estimated breeding values
  - My approach: what do I want to do with the number?
    - Describe the past (what happened): use the mean
    - Predict what will happen in the future / on a different farm: use the BLUP
- Concepts of BLUPs can be extended to any linear random effects
- One example: Modeling seed germination over time
  - maize seed lots stored the Ames Plant Introduction Center

– 2833 have 3 or more germination tests

– ca 40% tested 3 times, 33% tested 5 times, remainder 4, 6, or 7 times

– 3 seed lots



– Start with a quadratic model: Germ $= \beta_0 + \beta_1 Age + \beta_2 (Age)^2$

– Allow each coefficient to vary between seed lots

$$\text{Germ} = (\beta_0 + b_0) + (\beta_1 + b_1) Age + (\beta_2 + b_2)(Age)^2$$

– b's are correlated random effects, each with its own variance

– Matrix expressions for the model $\Rightarrow$ predictions of b's, and

– Predictions of germinability in the future