# Lecture 2

* basic Bayes
* two-level model for rates
* two-level normal-normal model

# Lecture 2 Outline

Today we will focus on the case where we have a outcome $Y_{ij}$ measured for $i = 1, \ldots, m$ clusters and $n_i$ observations within i.

The analysis goals are to:

1. Understand the relative size of the within and between-cluster variation in Y
2. Estimate the cluster specific mean of Y

To accomplish these goals we will

- Review Bayes theorem within context of diagnostic testing
- Look at the simplest binary two-level model
- Introduce the two-level normal-normal model
- Review the calculations for random effects within an example: on testing in schools

# Bayes Rule in Diagnostic Testing

# Diagnostic Testing

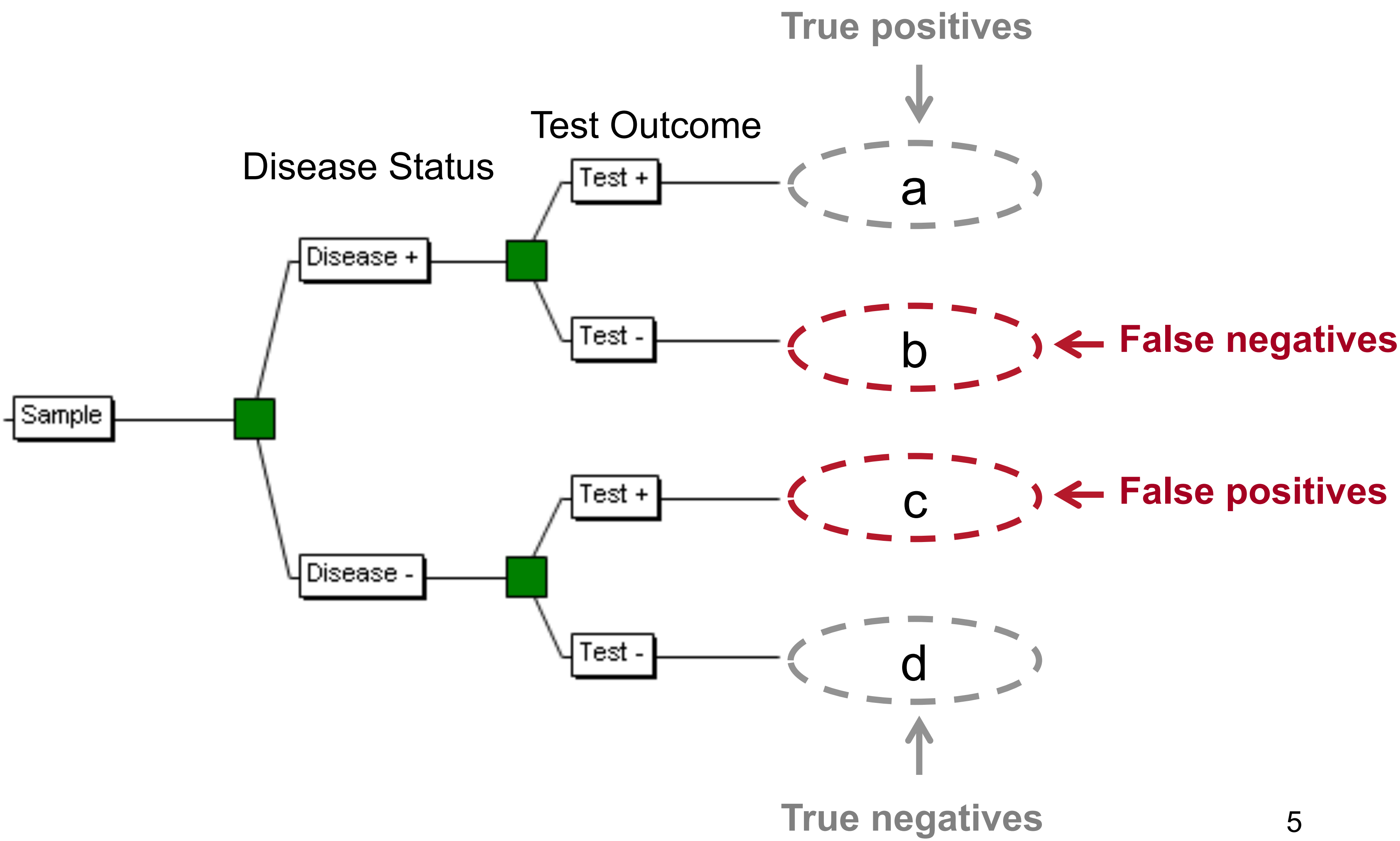
Ask

Marilyn®

BY MARILYN VOS SAVANT

**A particularly interesting and important question today is that of testing for drugs. Suppose it is assumed that about 5% of the general population uses drugs. You employ a test that is 95% accurate, which we'll say means that if the individual is a user, the test will be positive 95% of the time, and if the individual is a nonuser, the test will be negative 95% of the time. A person is selected at random and is given the test. It's positive. What does such a result suggest? Would you conclude that the individual is a drug user? What is the probability that the person is a drug user?**

# Diagnostic Testing



# Diagnostic Testing

- "The workhorse of Epi": The $2 \times 2$ table

| | **Disease +** | **Disease -** | **Total** |
|---|---|---|---|
| **Test +** | ***a*** | ***b*** | ***a* + *b*** |
| **Test -** | ***c*** | ***d*** | ***c* + *d*** |
| **Total** | ***a* + *c*** | ***b* + *d*** | ***a* + *b* + *c* + *d*** |

# Diagnostic Testing

- "The workhorse of Epi": The 2 × 2 table

| | **Disease +** | **Disease -** | **Total** |
|---|---|---|---|
| **Test +** | ***a*** | ***b*** | ***a* + *b*** |
| **Test -** | ***c*** | ***d*** | ***c* + *d*** |
| **Total** | ***a* + *c*** | ***b* + *d*** | ***a* + *b* + *c* + *d*** |

$Sens = P(+ \mid D) = \dfrac{a}{a+c}$ $\quad$ $Spec = P(- \mid \overline{D}) = \dfrac{d}{b+d}$

# Diagnostic Testing

- "The workhorse of Epi": The $2 \times 2$ table

| | Disease + | Disease - | Total |
|---|---|---|---|
| **Test +** | ***a*** | ***b*** | ***a + b*** |
| **Test -** | ***c*** | ***d*** | ***c + d*** |
| **Total** | ***a + c*** | ***b + d*** | ***a + b + c + d*** |

$$PPV = P(D \mid +) = \frac{a}{a+b}$$

$$NPV = P(\overline{D} \mid -) = \frac{d}{c+d}$$

$$Sens = P(+ \mid D) = \frac{a}{a+c} \qquad Spec = P(- \mid \overline{D}) = \frac{d}{b+d}$$

# Diagnostic Testing

- Marilyn's Example: Sens = 0.95, Spec = 0.95

| | Disease + | Disease - | Total |
|---|---|---|---|
| **Test +** | ***48*** | ***47*** | ***95*** |
| **Test -** | ***2*** | ***903*** | ***905*** |
| **Total** | ***50*** | ***950*** | ***1000*** |

PPV = 51%

NPV = 99%

P(D) = 0.05

# Diagnostic Testing

- Marilyn's Example
  - Sens = 0.95
  - Spec = 0.95

| | Disease + | Disease - | Total |
|---|---|---|---|
| **Test +** | ***190*** | ***40*** | ***230*** |
| **Test -** | ***10*** | ***760*** | ***770*** |
| **Total** | ***200*** | ***800*** | ***1000*** |

PPV = 83%

NPV = 99%

P(D) = 0.20

Point: PPV depends on prior probability of disease in the population

# Diagnostic Testing Linked to Bayes Theorem

- P(D): prior distribution, that is prevalence of disease in  the population
- P(+|D): likelihood function, that is the probability of observing a positive test given that the person has the disease (sensitivity) or doesn't (1-specificity)
- P(D|+), P(D|-) : posterior distribution of the unknown (latent) disease state given the test result

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+)}$$

$$P(+) = P(+ \mid D)P(D) + P(+ \mid \overline{D})P(\overline{D})$$

# Bayes & MLMs…

# Terminology

- **Two stage normal normal model**
- **Variance component model**
- **Two-way random effects ANOVA**
- **Hierarchical model with a random intercept and no covariates**

*Are all the same thing!*

# Testing in Schools

- Goldstein and Spiegelhalter JRSS (1996)
- Outcome: Standardized Test Scores
- Sample: 1978 students from 38 schools
  - Hierarchy: students (level 1) clustered within schools (level 2)
- Goal: Estimate overall “quality” of the schools using the results of the standardized test scores
- Possible Analyses:
  1. Calculate each school’s observed average score (approach A)
  2. Calculate an overall average for all schools (approach B)
  3. Borrow strength across schools to improve individual school estimates (Approach C)

# Two-stage normal normal model

$$y_{ij} = \theta_i + \varepsilon_{ij}$$

$$i = 1,\ldots, I,\ j = 1,..,n_i$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\boxed{\theta_i \sim N(\theta, \tau^2)}$$

i represents schools

j represents students

So that there are j =1,…, $n_i$ students within school i

$y_{ij}$ is the standardized test score for student j within school i

# Two-stage normal normal model

$$y_{ij} = \theta_i + \varepsilon_{ij}$$

$$i = 1, \ldots, I, j = 1, \ldots, n_i$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\boxed{\theta_i \sim N(\theta, \tau^2)}$$

In Bayesian approach, we would specify prior distributions for $\theta$, and $\tau^2$.

$\sigma^2$,

# Two-stage normal normal model

$$y_{ij} = \theta_i + \varepsilon_{ij}$$

$$i = 1,\ldots,I,\; j = 1,\ldots,n_i$$

$$\varepsilon_{ij} \sim N(0,\sigma^2)$$

$$\boxed{\theta_i \sim N(\theta,\tau^2)}$$

In **EMPIRICAL** Bayesian approach, we estimate $\theta$ and $\tau^2$ from the observed data and make no additional assumptions on these parameters

# Shrinkage estimation

- Goal: estimate the school-specific average score $\theta_i$
- Two simple approaches:
  - A) No shrinkage $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$

Maximum likelihood estimates of the cluster means

  - B) Total shrinkage $\hat{\theta} = \frac{\sum_{i=1}^{I} \frac{n_i}{\sigma^2} \bar{y}_i}{\sum_{i=1}^{I} \frac{n_i}{\sigma^2}}$

Inverse variance weighted average

NOTE: (a modified version of) this our estimate of θ in a random intercept model

# ANOVA and the F test

- To decide which estimate to use, a traditional approach is to perform a classic F test for differences among means
- If the group-means appear significantly different from each other then use A
- If the variance between groups is not significantly greater than what could be explained by individual variability within groups, then use B

# Shrinkage Estimation: Approach C

We are not forced to choose between A and B

- An alternative is to use a weighted combination between A and B

$$\hat{\theta}_i = \lambda_i \bar{y}_i + (1-\lambda_i)\hat{\theta}$$

$$\hat{\theta}_i = \hat{\theta} + \lambda_i(\bar{y}_i - \hat{\theta})$$

Empirical Bayes estimate

$$\lambda_i = \frac{\tau^2}{\tau^2 + \sigma_i^2}; \sigma_i^2 = \sigma^2 / n_i$$

NOTE: assumed variance within groups is constant.

# Shrinkage Estimation: Approach C

Empirical Bayes Estimates; THINKING 1st

- First way to think about it
- Take the "some" of the MLE + (1 – "some") of the overall mean
- Result is to get a biased estimate of the cluster specific mean BUT less variable since we are "shrinking" the cluster means towards the overall mean

$$\hat{\theta}_i = \lambda_i \bar{y}_i + (1 - \lambda_i)\hat{\theta}$$

$$\lambda_i = \frac{\tau^2}{\tau^2 + \sigma_i^2}; \sigma_i^2 = \sigma^2 / n_i$$

# Shrinkage Estimation: Approach C

Empirical Bayes Estimates

- Second way to think about it
- Application of Bayes rule!
- Take the estimate of the overall population mean
- Add an update based on the separation of variation into between clusters and within clusters

$$\hat{\theta}_i = \hat{\theta} + \lambda_i(\bar{y}_i - \hat{\theta})$$

$$\lambda_i = \frac{\tau^2}{\tau^2 + \sigma_i^2}; \sigma_i^2 = \sigma^2 / n_i$$

# Shrinkage estimation

- Approach C reduces to approach A (no pooling) when the shrinkage factor is equal to 1, that is, when the variance between groups is very large
- Approach C reduces to approach B, (complete pooling) when the shrinkage factor is equal to 0, that is, when the variance between group is close to be zero

$$\hat{\theta}_i = \lambda_i \bar{y}_i + (1 - \lambda_i)\hat{\theta}$$

$$\hat{\theta}_i = \hat{\theta} + \lambda_i (\bar{y}_i - \hat{\theta})$$

$$\lambda_i = \frac{\tau^2}{\tau^2 + \sigma_i^2}$$

$$\sigma_i^2 = \sigma^2 / n_i$$

# Back to the school example:

- Why borrow across schools?
  - Median # of students per school: 48, Range: 1-198
- Suppose small school (N=3)
  - Scores: 90, 90,10, average = 63
  - Difficult to say, small N $\Rightarrow$ highly variable estimates
- For larger schools we have good estimates, for smaller schools we may be able to borrow information from other schools to obtain more accurate estimates

# Two-stage normal normal model

$$y_{ij} = \theta_i + \varepsilon_{ij}$$

$$y_{ij} = \theta + b_i + \varepsilon_{ij}$$

$$i = 1,\ldots,I,\ j = 1,\ldots,n_i$$

$$\varepsilon_{ij} \sim N(0,\sigma^2)$$

$$\theta_i \sim N(\theta,\tau^2), b_i \sim N(0,\tau^2)$$

# Two-stage Normal-Normal Model

$$y_{ij} = \theta_i + \varepsilon_{ij}$$

$$y_{ij} = \theta + b_i + \varepsilon_{ij}$$

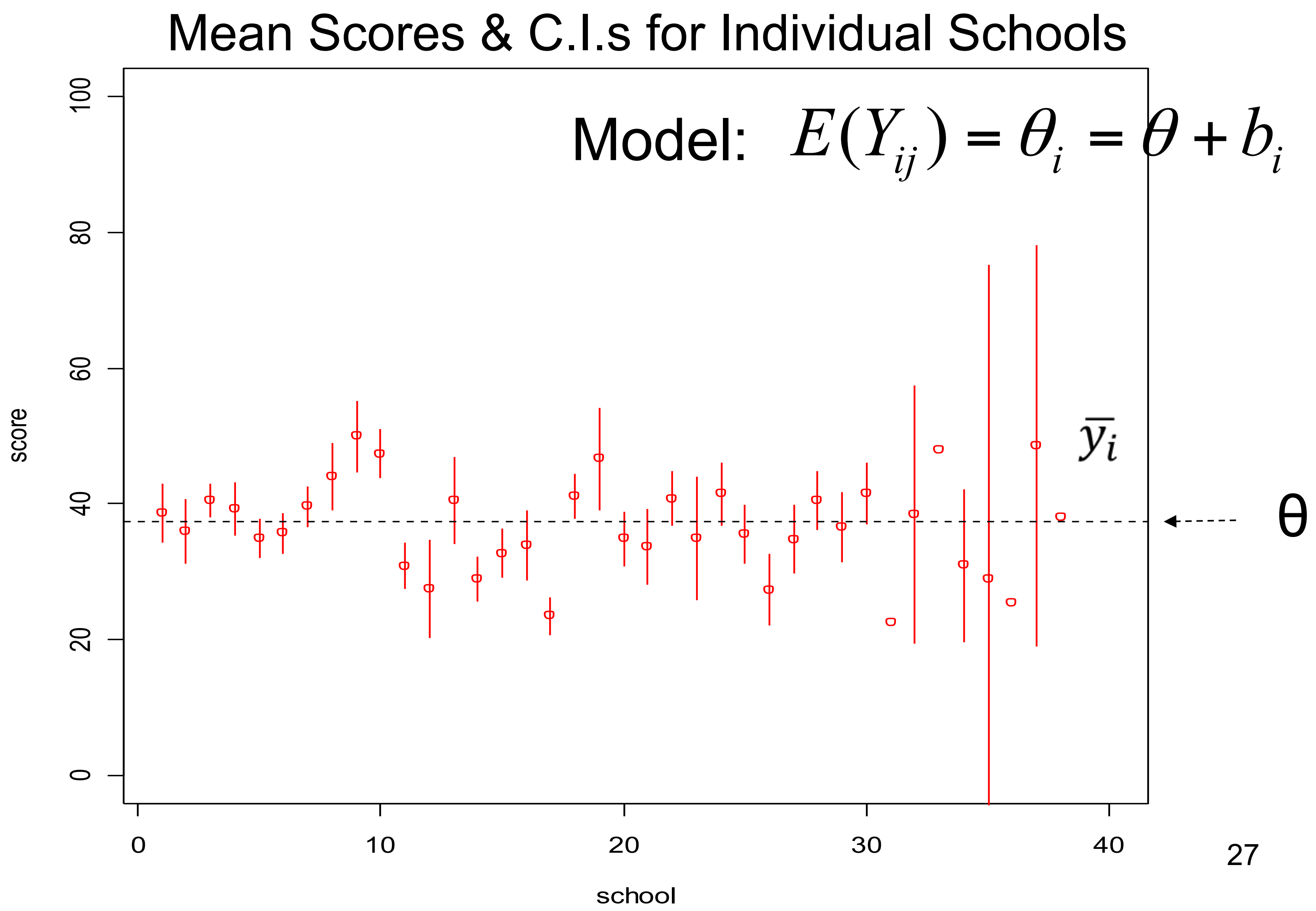
$\hat{\theta} = \text{inverse-variance weighted mean}$

$$\hat{\theta}_i = \lambda_i \bar{y}_i + (1 - \lambda_i)\hat{\theta}$$

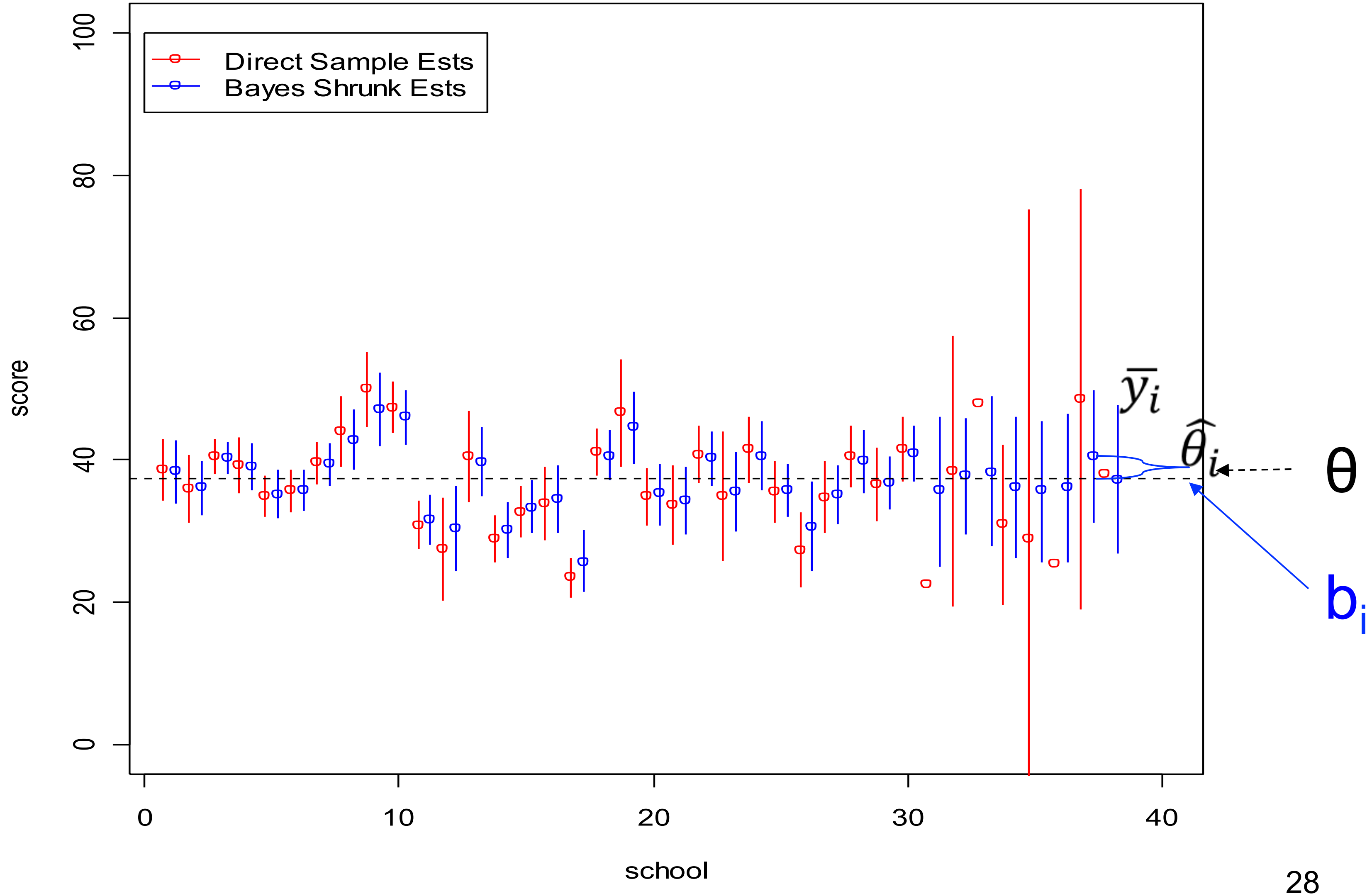
$$\hat{\theta}_i = \hat{\theta} + \lambda_i(\bar{y}_i - \hat{\theta})$$

$$\hat{\theta}_i = \hat{\theta} + \hat{b}_i$$

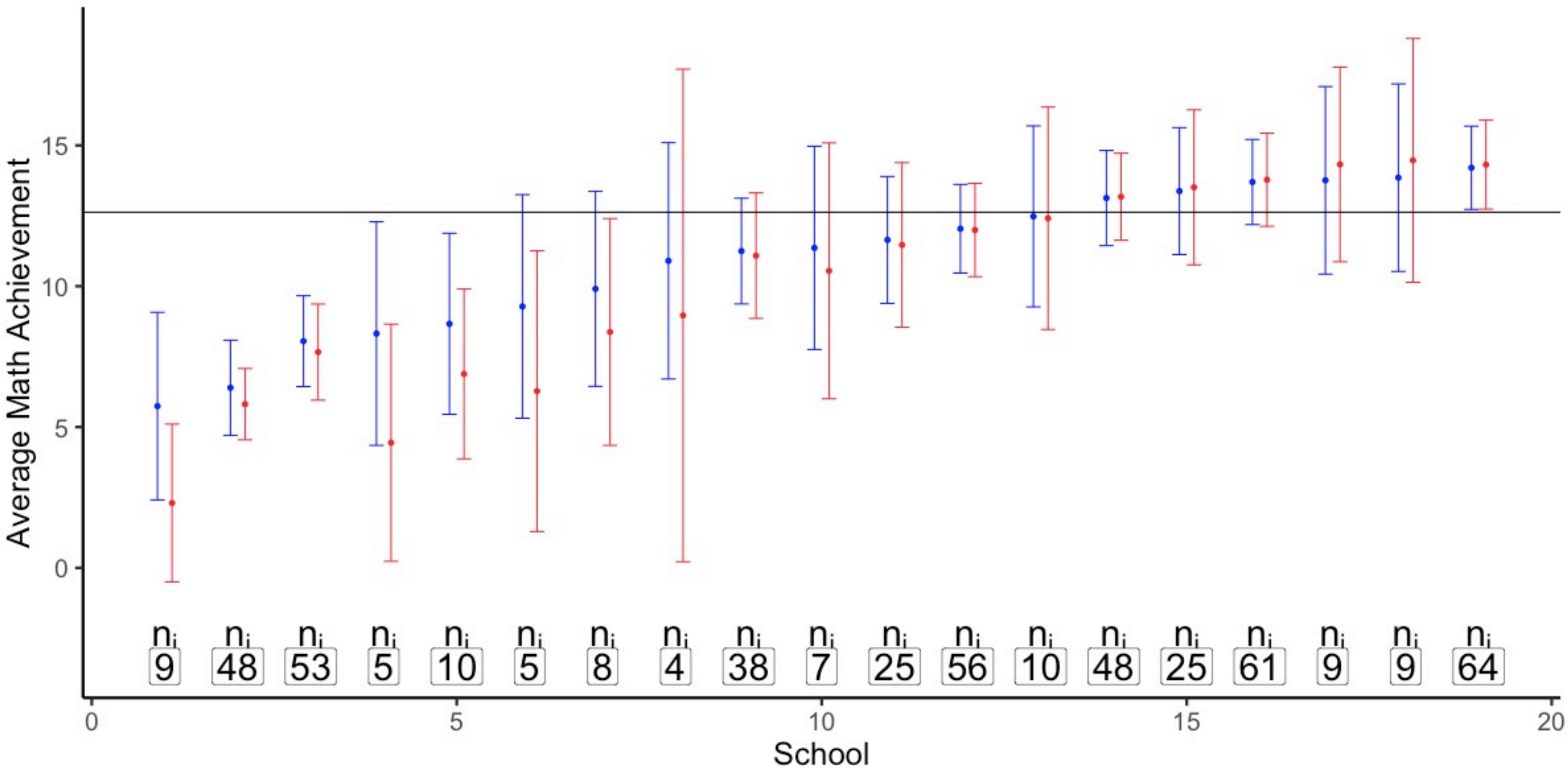
# Testing in Schools: MLE School Specific Mean



# Testing in Schools: Shrinkage Plot



# Using (a subset of) the data from Lab 0



# Some Bayes Concepts

- Frequentist: Parameters are “the truth”
- Bayesian: Parameters have a distribution
- In Bayes,
  – we “Borrow Strength” from other clusters
  – we “Shrink Estimates” towards overall averages
  – we compromise between model & data
  – Incorporate prior/other information in estimates
  – Account for other sources of uncertainty that are not measured

# Lecture 2 Summary

- Reviewed idea of updating information within Bayes theorem using classic 2x2 table
- Considered idea of "borrowing strength across similar problems" - risk estimation
- Walked slowly through the 2-level normal-normal model
  - The simplest case of multi-level model
  - Model allows for two sources of variation: within and between clusters
  - Overall mean is inverse-variance weighted average
  - Cluster specific estimates are weighted averages of observe cluster specific estimate and the overallmean
    - Weighting depends on the relative size of within and between cluster variance.

# Improved Assessment by *Borrowing Strength*; Estimating Local Coverage Rates (and penalty kick scoring) *Using* *Empirical Bayes Estimation* *(aka Multi-level models)*

# Many Related Questions

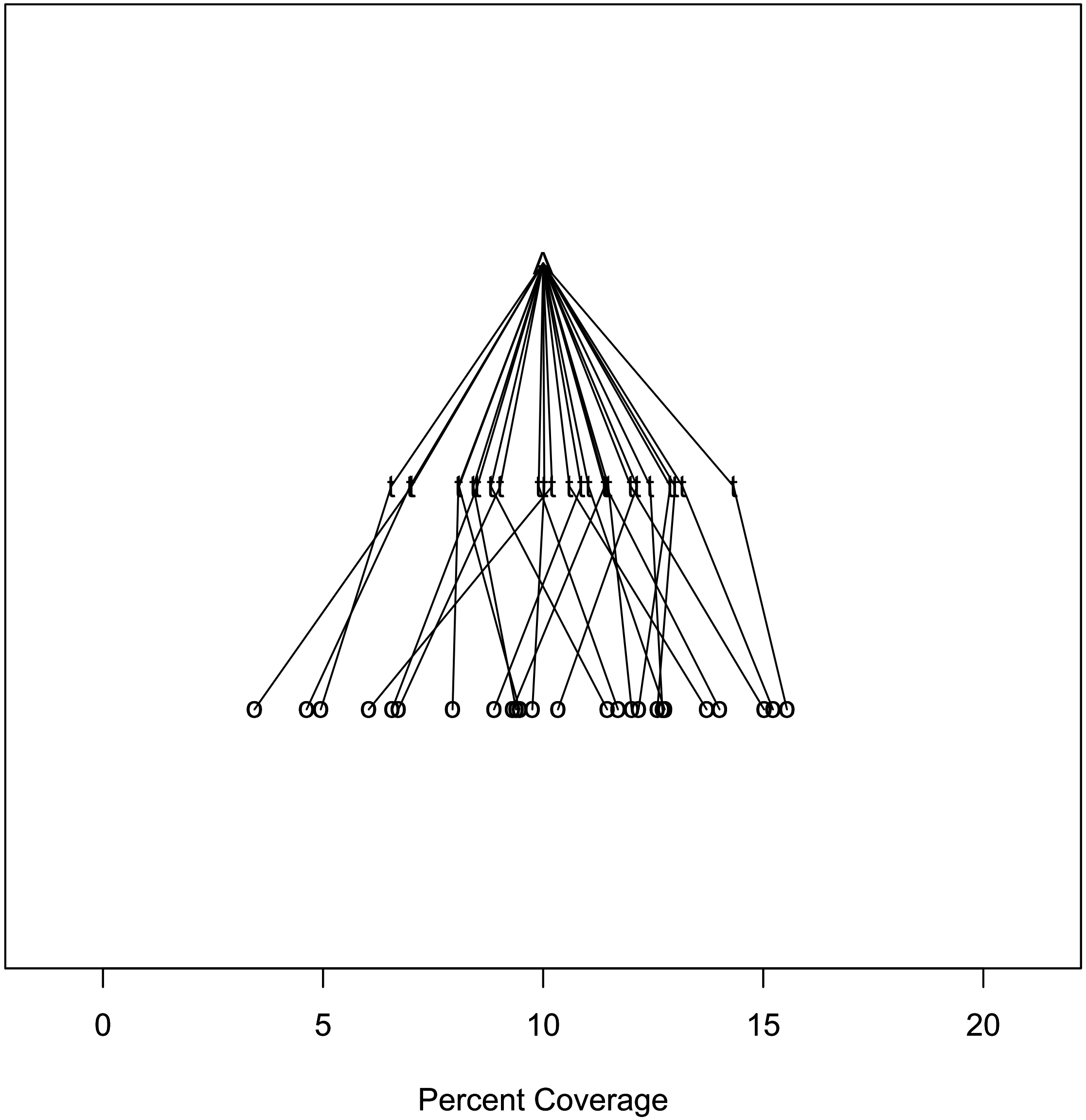
**One question**:  What is the rate of vaccine coverage  for children <24 mo in Mali?

**Many related questions**: What are the rates of measles vaccine coverage for children <24 mo for the 65 districts of Mali?

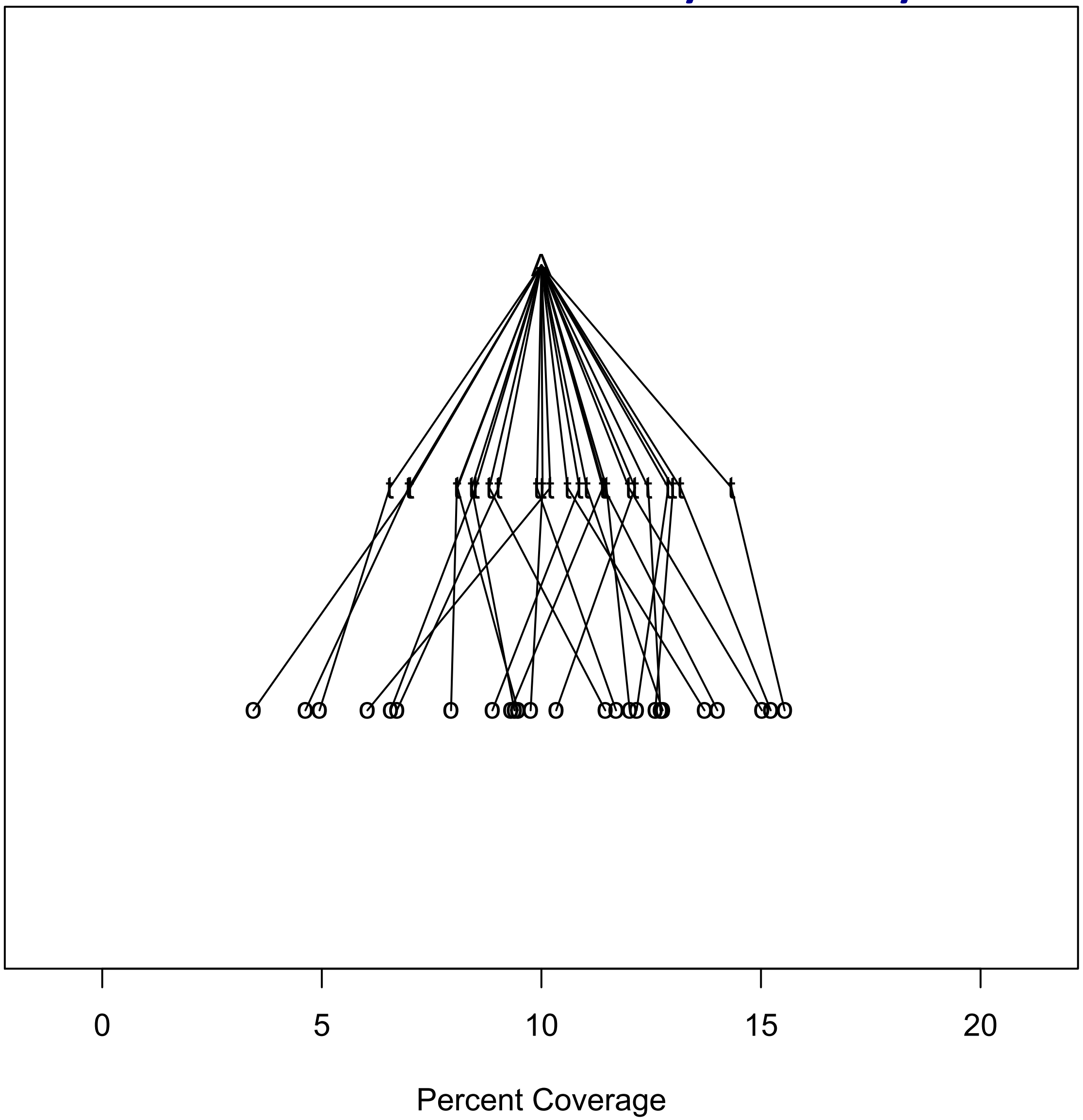
**One question**: What is the rate of penalty kick scoring in the world cup?

**Many related questions**: What are the rates of penalty kick scoring for the 30 top players in the world cup?

# DHS Observed Coverage Rates for 25 Imaginary Districts



# Q1. Are the observed rates more or less variable across districts than the true rates? Why and by how much?



# Hierarchical (or Multi (2)-level) Statistical Model

Level I: Observed rate (o) = true rate (t) for given district + deviation attributable to the sampling of only a subset of the population

Level II: True value for given district = average true value for the country + real deviation specific to the district

# Understanding "Borrowing Strength" Idea
# You do it every day; it is how our minds work

Q2. Identify a prediction you have made recently in answering a question and explain what "data" you used. How did you "borrow strength" from related questions and their data?

Answer

- Question: Will my boss give me a >3% raise this year?
- Related questions: Will my boss give each of my co-workers >3% raises this year
- Data
  - My last 3 raises have been <1%
  - The 4 other people in my group have all gotten <3% in each of the last three years; the boss favors Francine who told me she got a 2.5% raise
- Prediction: Probably, I will not get a raise >3%!

# Q3. What is the best way to estimate the true population rate for a particular district?

- Two extremes to consider
  - What if the true rates of coverage all are the same for all districts?
    - Best estimator of that common rate is:
  - What if there is large heterogeneity among districts so that one teaches us little about the other
    - Best estimator of a districts true rate is:

# Optimal Solution - Big Scientific Idea

- Compromise between the two extremes
  - Estimate:
    1. average rate (more precise; more biased); and
    2. district specific rate (less precise; less biased)
  - Estimate the degree of heterogeneity among districts in their **true** rates – (tricky)
  - Combine the average and district specific estimates, upweight the average when the heterogeneity is estimated to be small; upweight the district rate when heterogeneity is large


KUMHO

| | goals | shots | p |
|---|---|---|---|
| Valdivia | 0 | 1 | 0.00 |
| Forlan | 2 | 6 | 0.33 |
| Hernandez | 4 | 9 | 0.44 |
| Cavani | 11 | 17 | 0.65 |
| Lukaku | 3 | 5 | 0.60 |
| Samara | 0 | 3 | 0.00 |
| Benzema | 2 | 3 | 0.67 |
| Ozil | 1 | 6 | 0.17 |
| Giroud | 5 | 10 | 0.50 |
| Dempsey | 4 | 7 | 0.57 |
| Odemwinge | 2 | 4 | 0.50 |
| Johannsson | 2 | 8 | 0.25 |
| Martinez | 6 | 9 | 0.67 |
| Hulk | 12 | 23 | 0.52 |
| Robben | 7 | 10 | 0.70 |
| Ruaz | 3 | 5 | 0.60 |
| van Persie | 5 | 17 | 0.29 |
| Messi | 22 | 41 | 0.54 |
| Rodriguez | 4 | 7 | 0.57 |
| Muller | 8 | 14 | 0.57 |
| Aguero | 5 | 10 | 0.50 |
| Hazard | 13 | 19 | 0.68 |
| Vidal | 11 | 21 | 0.52 |
| Neymar | 5 | 6 | 0.83 |
| Guardado | 3 | 6 | 0.50 |
| Borges | 1 | 1 | 1.00 |
| Salpingidis | 0 | 1 | 0.00 |
| Moses | 3 | 3 | 1.00 |
| Drmie | 2 | 2 | 1.00 |
| Shaquiri | 0 | 1 | 0.00 |

## Optimal Solution- Implemented for World Cup Futbol

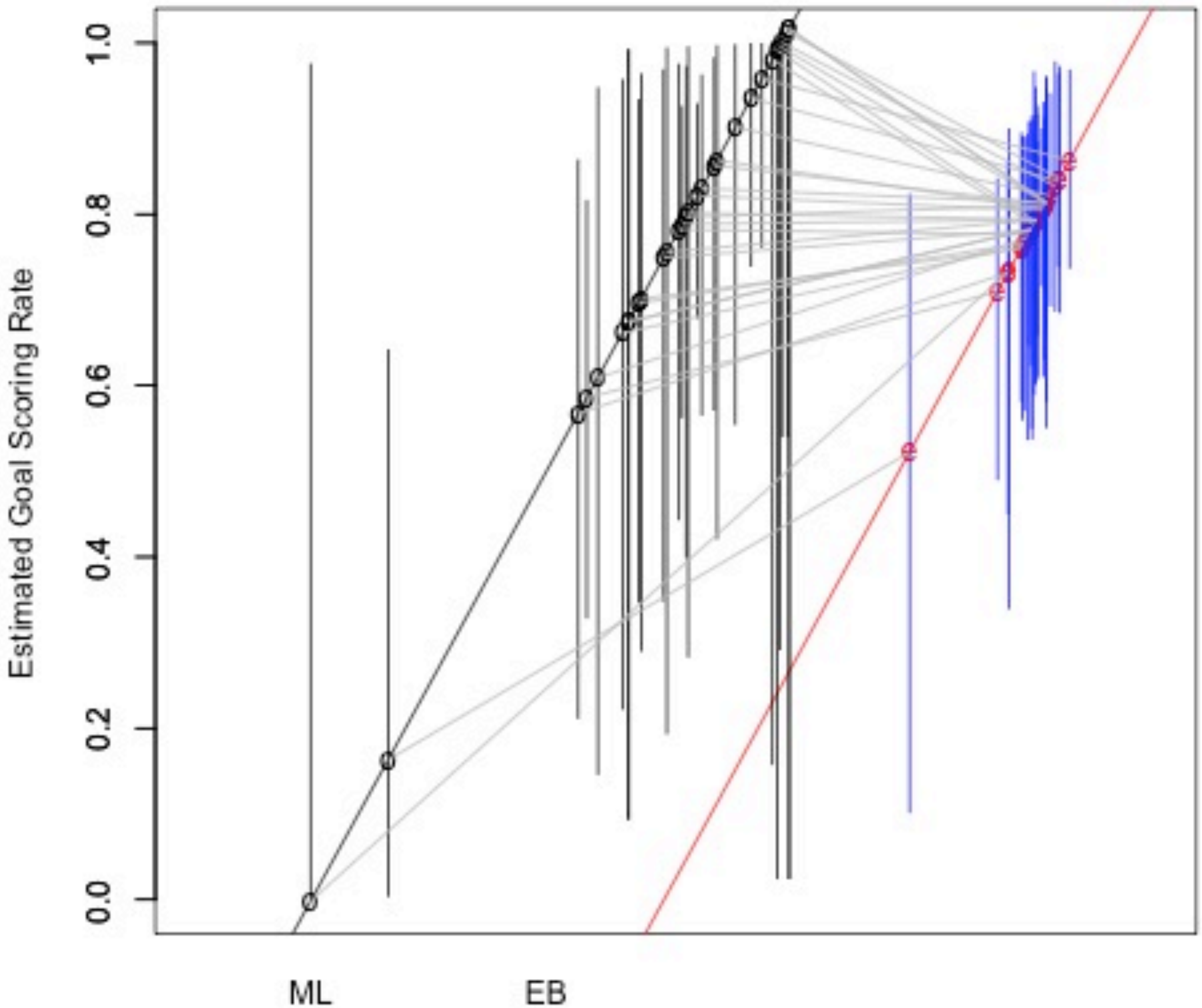
- Estimate the average goal scoring and the player specific rates
- Estimate the degree of heterogeneity among players in their true rates
- Compromise between the average and player-specific estimates, closer to the average when the player-to-player heterogeneity is estimated to be small.

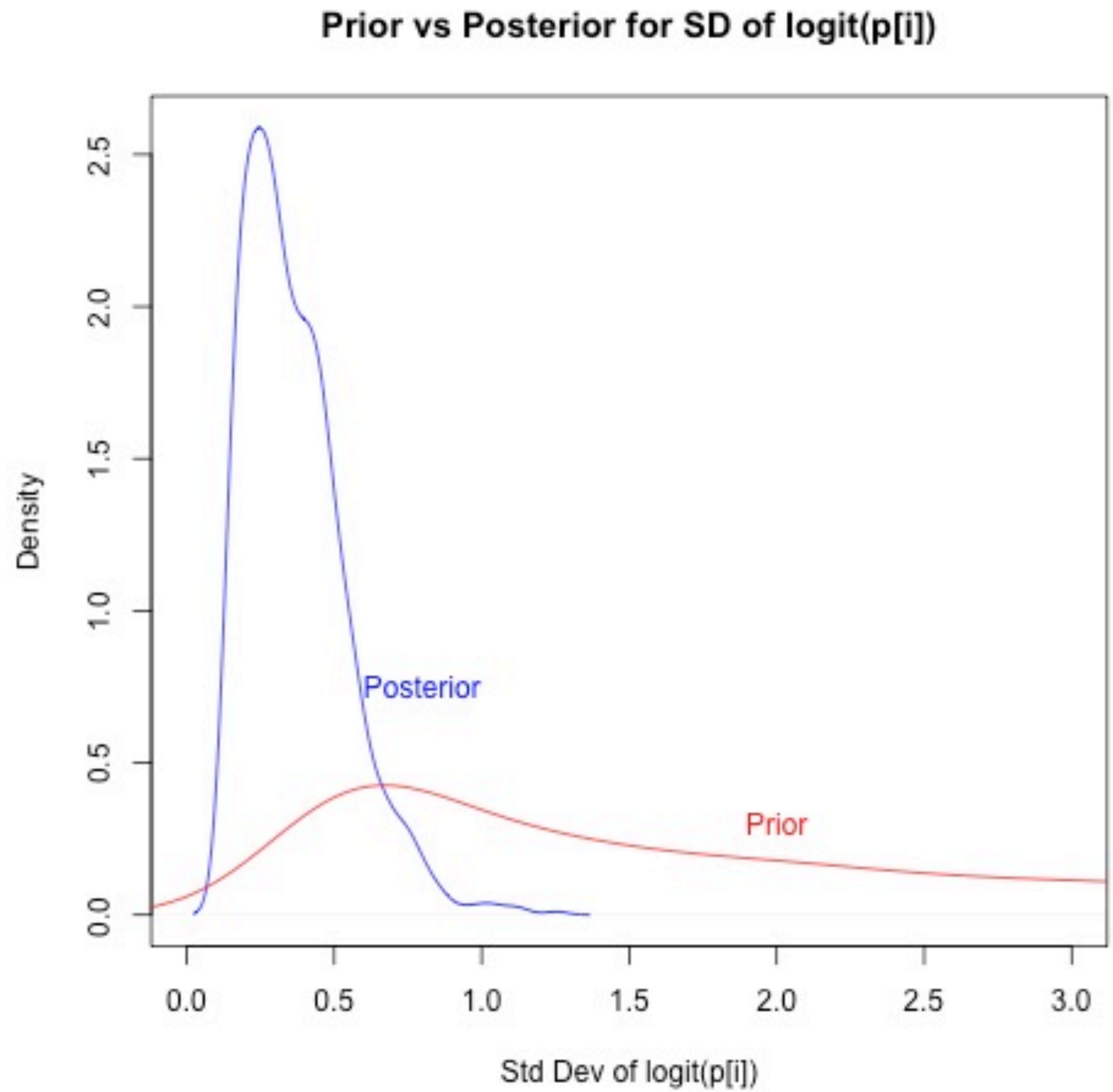
# Do the Empirical Bayes Calculation

| Task | Calculations and Results |
|---|---|
| 1. Calculate total variance, $V_T$, of observed rates, $p_i$ | 0.0792 |
| 2. Calculate variance among true rates = $V_T$ – mean of person-specific sampling variances ($se_i^2$) | 0.0792 - .0673 |
| 3. Calculate $\alpha_i = V_T / (V_T + se_i^2)$ – weight to give $p_i$ | |
| 4. Calculate $p.eb_i = (1-\alpha_i)\ p.bar + \alpha_i\ p_i$ | p.bar = 146 total goals out of 275 shots = 0.53 |

| | shots | goals | p | se | alpha | p.eb |
|---|---|---|---|---|---|---|
| Valdivia | 1 | 0 | 0.00 | 0.499 | 0.045 | 0.51 |
| Forlan | 6 | 2 | 0.33 | 0.204 | 0.222 | 0.49 |
| Hernandez | 9 | 4 | 0.44 | 0.166 | 0.300 | 0.50 |
| Cavani | 17 | 11 | 0.65 | 0.121 | 0.447 | 0.58 |
| Lukaku | 5 | 3 | 0.60 | 0.223 | 0.192 | 0.54 |
| Samara | 3 | 0 | 0.00 | 0.288 | 0.125 | 0.46 |
| Benzema | 3 | 2 | 0.67 | 0.288 | 0.125 | 0.55 |
| Ozil | 6 | 1 | 0.17 | 0.204 | 0.222 | 0.45 |
| Giroud | 10 | 5 | 0.50 | 0.158 | 0.323 | 0.52 |
| Dempsey | 7 | 4 | 0.57 | 0.189 | 0.250 | 0.54 |
| Odemwinge | 4 | 2 | 0.50 | 0.250 | 0.160 | 0.53 |
| Johannsson | 8 | 2 | 0.25 | 0.176 | 0.276 | 0.45 |
| Martinez | 9 | 6 | 0.67 | 0.166 | 0.300 | 0.57 |
| Hulk | 23 | 12 | 0.52 | 0.104 | 0.523 | 0.53 |
| Robben | 10 | 7 | 0.70 | 0.158 | 0.323 | 0.59 |
| Ruaz | 5 | 3 | 0.60 | 0.223 | 0.192 | 0.54 |
| van Persie | 17 | 5 | 0.29 | 0.121 | 0.447 | 0.42 |
| Messi | 41 | 22 | 0.54 | 0.078 | 0.661 | 0.53 |
| Rodriguez | 7 | 4 | 0.57 | 0.189 | 0.250 | 0.54 |
| Muller | 14 | 8 | 0.57 | 0.133 | 0.400 | 0.55 |
| Aguero | 10 | 5 | 0.50 | 0.158 | 0.323 | 0.52 |
| Hazard | 19 | 13 | 0.68 | 0.114 | 0.475 | 0.60 |
| Vidal | 21 | 11 | 0.52 | 0.109 | 0.500 | 0.53 |
| Neymar | 6 | 5 | 0.83 | 0.204 | 0.222 | 0.60 |
| Guardado | 6 | 3 | 0.50 | 0.204 | 0.222 | 0.52 |
| Borges | 1 | 1 | 1.00 | 0.499 | 0.045 | 0.55 |
| Salpingidis | 1 | 0 | 0.00 | 0.499 | 0.045 | 0.51 |
| Moses | 3 | 3 | 1.00 | 0.288 | 0.125 | 0.59 |
| Drmie | 2 | 2 | 1.00 | 0.353 | 0.087 | 0.57 |
| Shaquiri | 1 | 0 | 0.00 | 0.499 | 0.045 | 0.51 |

## Empirical Bayes Estimation




Prior vs Posterior for SD of logit(p[i])
Density
0.0
0.5
1.0
1.5
2.0
2.5
Posterior
Prior
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Std Dev of logit(p[i])

# Multi-level Model – hint of Bayes

- Level I. $[y_i | t_i, n_i]$: likelihood of the data given unknowns
  - Given true rate $t_i$ for player $i = 1,\ldots,m$, the observed number of scored penalty shots, $y_i$ out of $n_i$ attempts is *binomial*$(t_i, n_i)$
- Level II. $[t_i | n_i, m, v]$: prior distribution of unknowns
  - The true rates $t_i$ are independent and identically distributed random variables with distribution *Beta*$(m, v)$
- Multi-level model calculates $[t_i, m, v | n_i, y_i]$ : posterior of the unknowns given the data – using Bayes theorem