Brown 2008: Prediction of Batting Averages

Tues go through 1st 7 pages of these notes (the context)

Original data $\{H_i, N_i\}$ # hits for player i
at bats for player i
(1st half of season) p players

Original model $H_i \sim \text{Binomial}(N_i, p_i)$

unknown success/hit rate parameter p_i

$$\frac{H_i}{N_i} = \frac{\text{Bin}(N_i, p_i)}{N_i} \xrightarrow{d} \mathcal{N}\Big(p_i, \underbrace{\frac{p_i(1-p_i)}{N_i}}\Big)$$

wish this didn't depend on the unknown p_i

CLT $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k \xrightarrow{d} N(0, \sigma^2)$ if $\{X_k\}$ iid mean 0 variance $\sigma^2 < \infty$.

$$\frac{1}{\sqrt{N_i}}\left(\sum_{k=1}^{N_i} \text{Bernoulli}(p_i) - N_i p_i\right) \xrightarrow{d} \mathcal{N}(0, p_i(1-p_i))$$

$$\sqrt{N_i}\left(\frac{\text{Bin}(N_i, p_i)}{N_i} - p_i\right) \xrightarrow{d} \mathcal{N}(0, p_i(1-p_i))$$

$$\boxed{\frac{H_i}{N_i} \xrightarrow{d} \mathcal{N}\left(p_i, \frac{p_i(1-p_i)}{N_i}\right)}$$

2.

Variance Stabilizing Transformation

$$\boxed{X_i = \arcsin\sqrt{\frac{H_i + 1/4}{N_i + 1/2}}}$$

transformed data $\{X_i\}$

Where does the form $\arcsin\sqrt{\frac{H}{N}}$ come from?

$H \sim \text{Binomial}(N, p)$.

$\mu = \mathbb{E}\left(\frac{H}{n}\right) = p$, $\quad \text{var}\left(\frac{H}{n}\right) = \frac{p(1-p)}{n} =: V(p)$.

Find transformation $T(\cdot)$ so that $\text{Var}(T(\frac{H}{n}))$ constant. $\rightarrow$ not depending on p.

1st order Taylor Approx. of T about mean p at point $\frac{H}{N}$:

$$\boxed{T\left(\frac{H}{N}\right) \approx T(p) + T'(p)\left(\frac{H}{N} - p\right)}$$

$$\text{var}\left(T\left(\frac{H}{N}\right)\right) = (T'(p))^2 \,\text{var}\left(\frac{H}{N}\right)$$

$$= (T'(p))^2 V(p) \overset{\text{want}}{=} C \text{ constant}$$

$\Rightarrow$ $\boxed{\text{solve } T'(p) = \sqrt{\frac{C}{V(p)}}}$

$\Rightarrow$ $T(p) = \int_0^p \frac{\sqrt{C}}{\sqrt{V(t)}}\,dt = \sqrt{nC}\int_0^p \frac{dt}{\sqrt{t(1-t)}}$

$$= \sqrt{nC}\int_0^{\arcsin\sqrt{p}} \frac{2\sin\theta\cos\theta\,d\theta}{\sqrt{\sin^2\theta\,(1-\sin^2\theta)}}$$

letting $t = \sin^2\theta$

$$= \underbrace{2\sqrt{nC}}_{\text{constant}} \; \arcsin\sqrt{p}$$

Why $X_i = \arcsin\sqrt{\dfrac{H_i + 1/4}{N_i + 1/2}}$? (Significance of the Constants 1/4, 1/2)

- Constant variance $\mathrm{Var}(X) = \frac{1}{4N} + O\left(\frac{1}{N^2}\right)$
- Nice Expectation $EX = \arcsin\sqrt{p} + O\left(\frac{1}{N^2}\right)$
- $\boxed{X_i \xrightarrow{d} N(\theta_i, \sigma_i^2)}$

+ term → 0

will be estimating this parameter

$$\boxed{\begin{array}{ll} \theta_i = \arcsin\sqrt{p_i} & \text{unknown} \\ \sigma_i^2 = 1/4N_i & \text{known} \end{array}}$$

Why Lots of Algebra & Taylor Series manipulations. I have a (poorly written) reference from Brown on this.

3. Measuring the Efficacy of an Estimator

Transformed Data $\boxed{X_i \sim N(\underbrace{\theta_i}_{\text{unknown}}, \underbrace{\sigma_i^2}_{\text{Known}})}$

Goal $\boxed{\text{Estimate } \{\theta_i\}}$ from the data $\{X_i\}$, which is equivalent to estimating the i^{th} player's Batting Average

Estimator of θ_i is a function of the data

Notation: $\hat{\theta}_i = \delta(X_i)$

Total Squared Error of the Estimator δ

$$TSE(\delta) := \sum_i (\delta_i - \theta_i)^2 \quad \text{if we know the } \theta_i$$

$$= \sum_i (\delta_i - X_i)^2 + \sum_i (X_i - \theta_i)^2 - \sum_i 2(\delta_i - X_i)(X_i - \theta_i)$$

Estimate of TSE

$$\widehat{TSE}(\delta) := \sum_i (\delta_i - X_i)^2 + \mathbb{E}\left[\sum_i (X_i - \theta_i)^2 - \sum_i 2(\delta_i - X_i)(X_i - \theta_i)\right]$$

$$\boxed{\widehat{TSE}(\delta) = \sum_i (\delta_i - X_i)^2 - \sum_i \frac{1}{4N_i}}$$

$\mathbb{E}$ over $X \sim N(\theta_i, \sigma_i^2)$

$\sigma_i^2 = \frac{1}{4N_i}$ known

$$\left\{\begin{array}{ll} X_i & \text{known} \\ \delta_i = \delta_i(X_i) & \text{known} \\ N_i & \text{known} \end{array}\right\}$$

so, we can calculate $\widehat{TSE}(\delta)$ to evaluate our estimator $\delta_i(X_i)$ of θ_i

Naive Estimator

$$\boxed{\delta_0(x) = x}$$

Identity: guess future batting average will be the exact same as the previous BA

Normalization

want to compare estimator δ to δ_0.

$$\boxed{\widehat{TSE}^*(\delta) := \frac{\widehat{TSE}(\delta)}{\widehat{TSE}(\delta_0)}}$$

$$\boxed{\widehat{TSE}^*(\delta) < 1 \rightarrow \delta \text{ better than } \delta_0}$$

4. Choosing our Estimators $\delta(x)$ for Θ

- Naive Estimator
- Overall Mean
- Parametric Empirical Bayes
  - Method of Moments
  - MLE
- Nonparametric Empirical Bayes
- Harmonic Bayes Estimator
- James Stein Estimator

5 nontrivial estimators

Next Week: Lets' try to understand these estimators.

- **Results**

  Page 133 of Brown 2008, Table 2, $\hat{TSE}^*$ column.

  Naive estimator did the worst!

  EB(MM), NP EB, JS did the best! $\hat{TSE}^* \approx 0.5$

- **Explanation**

  1. $\{X_i\}$ not normal (moderately)
  2. Sample values for $\{N_i\}$ and $\{X_i\}$ are moderately correlated. When separate pitchers & nonpitchers, this correlation disappears, and the other estimators do way better!