Caution: These lecture notes are under construction. You may find parts that are incomplete.

1 Pythagorean Formula

1.1 When do we switch to preferring actual winning percentage?

n_i is the number of games played by team i
X_i is the Pythag W% of team i
Y_i is the actual W% of team i
Z_i is the residual W% of team i ($Y_i = X_i + Z_i$)

Intuition

$$\begin{aligned}\text{Actual W\%} &= \text{Pythag W\%} + \text{Residual}\\ \text{Outcome} &= \text{Skill} + \text{Luck}\\ Y_i &= X_i + Z_i\end{aligned}$$

Model

$$\begin{aligned}X_i &\sim \text{ind. Normal}(\mu_i,\, \sigma_X^2/n_i)\\ \mu_i &\sim \text{i.i.d. Normal}(\mu_0,\, \sigma_\mu^2)\end{aligned}$$

Option #1 (Z_i is all luck):

$$Z_i \sim \text{i.i.d. Normal}(0,\, \sigma_Z^2/n_i)$$

Option #2 (Z_i is not purely luck):

$$\begin{aligned}Z_i &\sim \text{ind. Normal}(\eta_i,\, \sigma_Z^2/n_i)\\ \eta_i &\sim \text{i.i.d. Normal}(0,\, \sigma_\eta^2)\end{aligned}$$

For Z_i, the signal variance is σ_η^2, and the noise variance is σ_Z^2/n. The total variance is $\sigma_\eta^2 + \sigma_Z^2/n$. When $n = \sigma_Z^2/\sigma_\eta^2$, the variance in Z_i is half signal, half noise.

A common measurement of interest for evaluating metrics in baseball is the *split-half correlation*. A high split-half correlation close to one tells you that a metric is stable and reliable. A lower split-half correlation close to zero tells you that a metric is noisy and unreliable. We can imagine splitting the season into two halves and calculating the residual winning percentages Z_i^1 and Z_i^2 in the first half and second half respectively. Assuming equal sample sizes $n_i = n_i^1 = n_i^2$,

$$\begin{aligned}\text{Corr}(Z_i^1, Z_i^2) &= \frac{\text{Cov}(Z_i^1, Z_i^2)}{\sqrt{\text{Var}(Z_i^1)\text{Var}(Z_i^2)}} = \frac{\text{Cov}(\eta_i, \eta_i)}{\sqrt{(\sigma_\eta^2 + \sigma_Z^2/n_i^1)(\sigma_\eta^2 + \sigma_Z^2/n_i^2)}}\\ &= \frac{\text{Var}(\eta_i)}{\sqrt{(\sigma_\eta^2 + \sigma_Z^2/n_i)(\sigma_\eta^2 + \sigma_Z^2/n_i)}} = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_Z^2/n_i}.\end{aligned}$$

Again we see the significance of $n = \sigma_Z^2/\sigma_\eta^2$ because this sample size makes the split-half correlation 0.5.

Lastly, if our goal is to use observed results to predict future results, then $\mu_i + \eta_i$ is what we want to estimate. We can compare how well X_i and Y_i achieve this goal.

$$\begin{aligned}E[(X_i - (\mu_i + \eta_i))^2] &= E[((X_i - \mu_i) + \eta_i)^2]\\ &= E[(X_i - \mu_i)^2] + 2E[(X_i - \mu_i)\cdot \eta_i] + E[\eta_i^2]\\ &= \sigma_X^2/n + 0 + \sigma_\eta^2 = \sigma_X^2/n + \sigma_\eta^2.\end{aligned}$$

By contrast,

$$\begin{aligned}
E[(Y_i - (\mu_i + \eta_i))^2] &= E[((X_i - \mu_i) + (Z_i - \eta_i))^2] \\
&= E[(X_i - \mu_i)^2] + 2E[(X_i - \mu_i) \cdot (Z_i - \eta_i)] + E[(Z_i - \eta_i)^2] \\
&= \sigma_X^2/n + 0 + \sigma_Z^2/n = \sigma_X^2/n + \sigma_Z^2/n.
\end{aligned}$$

We see that $E[(Y_i - (\mu_i + \eta_i))^2] < E[(X_i - (\mu_i + \eta_i))^2]$ when $n > \sigma_Z^2/\sigma_\eta^2$. In other words, actual record (Y_i) becomes a stronger prediction of future record than Pythagorean record (X_i) when the number of games observed is at least σ_Z^2/σ_η^2.