Bioinformatics Research Center
Aarhus University

Asger Hobolth
23. september 2016

Supplerende materiale til:

Hvem er bedst?
Rangering af straffekastskytter i amerikansk basketball

Grundlæggende statistisk model

En naturlig model for antallet af scoringer x_i for spiller i med antal scoringsforsøg n_i er binomialfordelingen

$$X_i \sim \text{Bin}(n_i, p_i).$$

Middelværdi og varians er

$$\text{E}(X_i) = n_i p_i \quad \text{og} \quad \text{Var}(X_i) = n_i p_i (1 - p_i),$$

og dermed har vi også at

$$\text{E}(X_i/n_i) = p_i \quad \text{og} \quad \text{Var}(X_i/n_i) = p_i(1 - p_i)/n_i.$$

Derfor er scoringsfrekvensen $\hat{p}_i = x_i/n_i$ et rimeligt estimat for scoringssandsynligheden p_i. Udfordringen er, at variansen på estimatet varierer meget mellem spillerne, fordi de har meget forskellige antal scoringsforsøg.

(a) ESPN rangering

ESPN rangerer ved først at have en kvalifikationstærskel på 125 scoringer, og dernæst bruge scoringsfrekvensen til rangering. Hvis scoringsfrekvensen er f.eks. 0.6 skal spilleren altså have skudt mindst 125/0.6=209 gange for at være kvalificeret.

(b) P-værdi rangering

Alternativt kan spillerne rangeres ved p-værdien for hypotesen $H : p = p_0$, hvor p_0 er scoringsfrekvensen for alle forsøgene i hele sæsonen. Binomialfordelingen kan approksimeres med en normalfordeling

$$X_i \approx N\big(n_i p_i, n_i p_i (1 - p_i)\big),$$

og så bliver teststørrelsen

$$t_i = \frac{x_i - n_i p_0}{\sqrt{n_i p_0 (1 - p_0)}},$$

hvor en stor værdi af t_i er kritisk for hypotesen. Eftersom p_0 er fast bliver t_i proportional med

$$t_i \propto \frac{x_i - n_i p_0}{\sqrt{n_i}} = \sqrt{n_i}\Big(\frac{x_i}{n_i} - p_0\Big) = \sqrt{n_i}\Big(\hat{p}_i - p_0\Big).$$

Bemærk hvordan effektstørrelsen $(\hat{p}_i - p_0)$ bliver forøget af kvadratroden af antal straffekastforsøg.

I biomedicinsk forskning bruges p-værdier til rangering af enheder (en enhed er i dette tilfælde en spiller).

(c) Bayesiansk statistik: Posterior middelværdi rangering

En tredje mulighed er at bruge Bayesiansk statistik til rangering. I Bayesiansk statistik specificeres en prior på modellens parametre. Beta-fordelingen er en fleksibel fordeling på intervallet fra 0 til 1, så den bruges ofte som prior på sandsynligheder.

En Beta-fordeling $Y \sim \text{Beta}(a, b)$ med formparametre a og b har tæthed

$$f(y) = \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)}, \quad 0 \leq y \leq 1.$$

Her er

$$B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1} dy$$

en normaliseringskonstant, der sørger for, at tætheden integrerer til 1. Middelværdi og varians for Beta-fordelingen er

$$\text{E}(Y) = \frac{a}{a+b} \quad \text{og} \quad \text{Var}(Y) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Prior for scoringssandsynligheden er altså $p_i \sim \text{Beta}(a, b)$, hvor parametrene a og b bestemmes så middelværdien og variansen har samme værdi som den empiriske middelværdi og den empiriske varians for scoringsfrekvenserne.

Posterior fordeling er prior fordeling for scoringssandsynligheden opdateret med data fra den enkelte spiller. I dette tilfælde er posterior fordeling en Beta fordeling med parametre $a + x_i$ og $b + n_i - x_i$. Middelværdien for posterior fordeling er altså

$$\tilde{p}_i = \frac{a + x_i}{a + b + n_i}.$$

I tabellen nedenfor har jeg opsummeret de forskellige metoder.

Metode	Rangering	Betingelser og konstanter
ESPN	$\hat{p}_i = x_i/n_i$	Hvis $x_i \geq 125$; ellers ikke kvalificeret
P-værdi	$t_i = \sqrt{n_i}(\hat{p}_i - p_0)$	$p_0 = 0.75$
Bayes	$\tilde{p}_i = (a + x_i)/(a + b + n_i)$	$a = 7.5$ og $b = 2.5$

En omskrivning af middelværdien for posterior fordeling giver

$$\tilde{p}_i = \frac{a + x_i}{a + b + n_i} = \frac{(a+b)}{(a+b+n_i)} \frac{a}{(a+b)} + \frac{n_i}{(a+b+n_i)} \frac{x_i}{n_i} = w_i \frac{a}{(a+b)} + (1 - w_i)\frac{x_i}{n_i},$$

hvor vægten $w_i = (a+b)/(a+b+n_i)$ er tæt på 1 hvis n_i er lille, og tæt på 0 hvis n_i er stor. Hvis n_i er lille er posterior middelværdi altså tæt på prior middelværdi, og hvis n_i er stor er posterior middelværdi tæt på scoringsfrekvensen. Derfor er ESPN rangering og Posterior middelværdi rangering næsten identiske for store n_i.

Referencer

Wikipedia siden

`https://en.wikipedia.org/wiki/Beta-binomial_distribution`

indeholder ovenstående teori.