Lec 10

3/15/21

Consider the following dataset. There are 6,115 mothers. Each mother had >12 children, and we only consider their first 12 children. (Thus each mother has 12 children in this dataset.) We now count the # of boyes for each mother

# Boys	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
X	3	24	104	286	670	1033	1343	1112	829	478	181	45	7	6115
Binom Prediction	1	12	72	259	628	1065	1367	1266	854	410	152	26	2	6115
Beta-Binom Predic	2	23	105	311	656	1036	1258	1182	854	462	178	44	5	

How do we model this data ($\approx$). This example is beyond the scope of the course. E.g. $X \sim Bin(12, 50\%)$. It turns out, the sex ratio is not even: P(boy) is closed to 51.1% (not 50%). That diffrence is real. So let's examin the model $X \sim Bin(12, 51.1\%)$

How do we fit a betabinomial? We know $n=12$, what is α, β? We fit the alpha and Beta with Maximum likelyhood and find alphaMLE = 34 & betaMLE = 32. So now we have $X \sim BetaBinomial(12, 34, 32)$.

$$E[X] = \frac{12(34)}{(34)+32} = 0.515 \sim 51.1\% \text{ (the published avg)}$$

The betabinomial model fits better to human birth rate.

$P(\theta) = Beta(34, 32)$ $\quad Q[\theta, .5\%] = 36\%$

$\quad\quad\quad\quad\quad\quad\quad\quad\quad Q[\theta, .99\%] = 67\%$

Back to the curriculum... what about the following problem. You see data for n Bernoulli trials. What if you want to know about the *next, future* n_* trials you haven't seen?

Scanned with CamScanner

$\frac{0}{1}\ \frac{1}{2} \ldots \ldots \frac{1}{n}$ | $\overline{}\ \overline{} \ldots \ldots \overline{n_*}$

$X = \#$ 1's above past | $X_* = ?$ future

This problem is called the "prediction" problem ("forecasting"). In science there are generally two goals: (1) explaining phenomena which means finding a model $\mathcal{F}$ and estimating θ, it's parameters and (2) Predicting the future values of the phenomena. They are related.

Consider:

$$P(X_* \mid X = x) \overset{?}{=} \text{Bin}(n_*, \hat{\hat{\theta}}_{MLE}) = \text{Bin}\left(n_*, \frac{x}{n}\right)$$

If θ is known $\Rightarrow$ $=$ $\text{Bin}(n_*, \theta)$, but θ is never known!

Is using the MLE a reasonable idea? Yes

The problem with the above is $\hat{\theta}_{MLE}$ is not θ and there is uncertainty in its estimation that is not being accounted for. We know with n large, the MLE is approximately normally distributed. We can use this, but if n is small, it won't be accurate. So... Bayesian statistics to the rescue!

$$P(X_* \mid x) = \int_{\Theta} P(X_*, \theta \mid x)\, d\theta = \int_{\Theta} P(X_* \mid \theta, x)\, P(\theta \mid x)\, d\theta$$

If θ is known, X doesn't given you any more information.

Posterior Predictive distribution, a mixture/compound rv

$$= \int_{\Theta} \overbrace{P(X_* \mid \theta)}^{\text{Likelyhood}}\, \underbrace{P(\theta \mid x)}_{\text{Posterior}}\, d\theta$$

for $\mathcal{F}: \text{Bin}(n, \theta)$, Prior $P(\theta)$

$$\overset{\leftarrow}{=} \int_{\Theta} \underbrace{P(X_* \mid \theta)}_{\text{Bin}(n_*, \theta)}\, \underbrace{P(\theta \mid x)\, d\theta}_{\text{Beta}(\alpha + x, \beta + n - x)} = \text{BetaBinom}(n_*, \alpha + x, \beta + n - x)$$

Scanned with CamScanner

$= \text{BetaBinom}(n_*, \alpha + x, \beta + n - x)$

$$P(\theta) \xrightarrow{x} P(\theta|X) \text{ but also } P(X_*) \xrightarrow{x} P(X_*|X)$$

$$\overset{\parallel}{\int_{\Theta} P(x|\theta)P(\theta)d\theta} \to \int P(X_*|\theta)P(\theta|x)d\theta$$

Let's see a concrete example. A new baseball player has $n=10$ bats, and he get $x=6$ hits. Assuming each at bat is iid $Bern(\theta)$, what is the probability he will have $X_*=17$ hits in the next $n_*=32$ bats. Assume a uniform prior. $P(\theta) = Beta(1,1)$

$$P(X_*|x=6) = \text{BetaBinomial}(32, 6+1, 1+10-6)$$

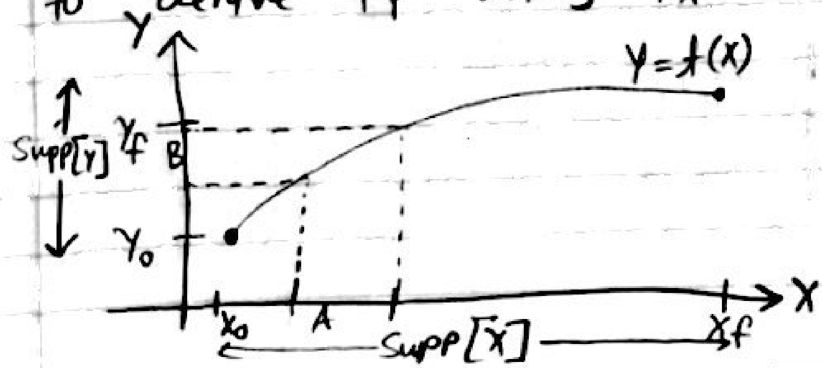
$$P(X_*=17|x=6) = \frac{\binom{32}{17}}{B(7,5)} B(24, 20) = \text{dbetabinom}(17, 32, 7, 5)$$

What's the probability he gets 17 or less hits on the next at bats?

$$P(X_* \leq 17|X=6) = \sum_{Y=0}^{17} \frac{\binom{32}{7}}{B(7,5)} B(Y+7, 32-Y+5)$$

$$= \text{Pbetabinom}(17, 32, 7, 5)$$

Back to probability land... Let X, Y be continuous rv's where f_X is known and $Y = t(X)$ where t is a known invertible function. We want to derive f_Y using f_X and t.



Scanned with CamScanner

Based on the graph:

$P(X \in A) = P(Y \in B)$

If A, B small,

$P(X \in A) \approx f_X(x)\,|dx|$

$P(Y \in B) \approx f_Y(y)\,|dy|$

$\Downarrow$

$f_X(x)\,|dx| = f_Y(y)\,|dy| \Rightarrow f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$

$y = f(x) \Rightarrow x = f^{-1}(y) \Rightarrow f_Y(y) = f_X(f^{-1}(y))\left|\frac{d}{dy}[f^{-1}(y)]\right|$

= This is called the change of variable formula for densities.