lec 10
3/15/21

Consider the following dataset. There are 6,115 mothers. Each mother had >=13 children. and we only consider their first 13 children (thus each mother has 13 children in the dataset). We now count # of boys for each mother.

# of 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
Binomial / predicted	1	12	72	259	628	1085	1367	1266	854	410	152	26	2	6115
Betabinom	2	23	105	311	656	1036	1258	1182	854	462	178	44	5	6115

How do we model this data (curly F). This example is beyond the scope of the course. E.g. $X \sim B(12, 50\%)$. It turns out, the sex ratio is not even: P(boy) is closer to 51.1% (not 50%). That difference is real. so let's examine the model $X \sim Bin(12, 51.1\%)$

How do we fit a betabinomial? We know $n=12$. What is alpha and beta? We fit the alpha and beta with maximum likehood and find alphaMLE = 34 and beta = 32. So now we have $X \sim BetaBin(12, 34, 32)$ $\longrightarrow$

$$E[X] = \frac{12 * 34}{34+32} = 6.5 \text{ is} \sim 51.1\% \text{ (the publish}$$

The beta binomial model fits better to human birth data.

$P(\theta) = \text{Beta}(34, 32)$ $\quad$ $Q[\theta, 0.5\%] = 36\%$

$Q[\theta, 99.5\%] = 67\%$

Back to the cirriculum... 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?

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

$x = \#$ 1s 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 and estimating θ, it's parameter and (2) predicting the future values of the phenomea. They are related.

Consider :

$$P(X_* | X=x) \overset{?}{=} Bin(n_*, \hat{\theta}_{MLE})$$

if θ know $= Bino(n_*, \theta)$ but θ is never known!

But it's true.

is using the MLE a resonable idea? Sure.... We can do better! The problem with the above is $\hat{\theta}_{MLE}$ is not theta and there is uncertainty in it's 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.

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

If θ is known, X doesn't give you any information

Posterior predictive distribution, $= \int_{\Theta} \underbrace{P(X_*|\theta)}_{\text{likelihood}} \underbrace{P(\theta|x)}_{\text{Posterior}} d\theta$

for F: $B(n, \theta)$, prior $P(\theta)$ beta

$$= \int_{\Theta} \underbrace{P(x_*|\theta)}_{Bin(n_*, \theta)} \underbrace{P(\theta|x)}_{Beta(\alpha + x, \beta + n - x)} d\theta$$

$P(\theta) \longrightarrow P(\theta|x)$ but also $P(X_*) \xrightarrow{x} P(X_*|x) =$

$= \int_{\Theta} P(x_*|\theta) P(\theta) d\theta \longrightarrow \int_{\Theta} P(x_*|\theta) P(\theta|x) d\theta$

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

$P(X_*|X=6) = BetaBinomial(32, 1+6, 1+4)$

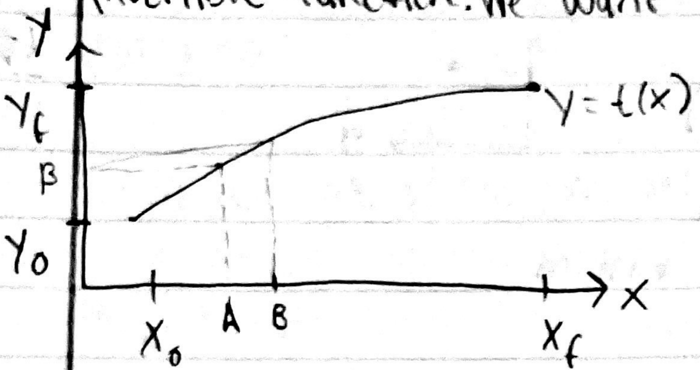
$$P(X_* = 17 | X=6) = \frac{\binom{32}{17}}{B(7,5)} B(24, 20) = dbetabinomial(17, 32, 7, 5)$$

what is the probability he gets 17 or less hits on the next 32 at bats?

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

$= pbeta(17, 32, 7, 5)$

Back to probability land... let x and y be a continous random variable where f_X and f_X are known and $Y = t(X)$ where t is known and $Y = t(X)$ where t is known invertible function. We want to derive f_Y using f_X and t.



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

If A, B Small

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

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

Therefore,

$f_X(x)|dx| = f_Y(y)|dy|$

$Y = t(X)$
$X = t^{-1}(Y)$

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

This is called the changes of variable formula for densities: → $\boxed{f_Y(y) = f_X(t^{-1}(y))\left[\frac{d}{dy}[t^{-1}(y)]\right]}$

formula