Basic Probability

Probability: Quantifying our belief in the possible outcomes of future event

Event 1: First roll of 6-sided die (DR1)

$$P(DR1 = 2) = \frac{1}{6} \qquad P(DR1 = 5) = \frac{1}{6}$$

Independent Events: The outcome of one event doesn't influence outcome of another

Event 2: Second roll of 6-sided die (DR2)

$$P(DR2 = 3 \mid DR1 = 5) = P(DR2 = 3) = \frac{1}{6}$$

"given"

Joint probability for independent events

$$P(DR1 = 5 \text{ AND } DR2 = 2) = P(DR1 = 5) \cdot P(DR2 = 2) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Mutually Exclusive Outcomes: when one outcome happens the other can't happen at same time

$$P(DR1 = 5 \text{ OR } DR1 = 4) = P(DR1 = 5) + P(DR1 = 4) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Maximum Likelihood Parameter Estimation

Parameter: A true (as opposed to estimated) measure of an individual or population

Maximum Likelihood: Method for estimating unknown parameter from data & statistical Model

ML Steps

1. Determine mathematical relationship between parameter being estimated and data being observed (build model)
2. Calculate Probability of observing data assuming a particular parameter value
3. Repeat step #2 for each possible parameter value
4. Best parameter estimate is value that maximized the probability of data

$$\text{Likelihood}(\text{Parameter} = X) = P(\text{Data} \mid \text{Parameter} = X)$$

Calculating Genotype Likelihoods (1 read)

Parameter: Genotype of Individual 01 @ chr 16 pos 18,716 (C/G SNP)

Data: 1 C read, sequencing error rate = 3%

Possible Parameter Values: G = CC, G = CG, G = GG

Likelihood (G = CC) = P(Data | G = CC)

Likelihood (G = CG) = P(Data | G = CG)

Likelihood (G = GG) = P(Data | G = GG)

P(C read | G = CG) = 0.49

sample C allele & no error OR sample G allele & G→C error

$(0.5 \cdot 0.97) + (0.5 \cdot 0.03(\frac{1}{3}))$

P(C read | G = CC) = 0.97

sample C allele & no error

$(1 \cdot 0.97)$

P(C read | G = GG) = 0.01 G→C

sample G allele & error

$(1 \cdot 0.03(\frac{1}{3}))$

Calculating Genotype Likelihoods (Multiple reads)

Data: 2 C reads, sequencing error rate = 3%

$P(C,C \text{ reads} \mid G=CC) = 0.97 \cdot 0.97 = 0.94$

read 1 C & read 2 C

$P(C \text{ read} \mid G=CC) \cdot P(C \text{ read} \mid G=CC)$

$P(C,C \text{ reads} \mid G=CG) = 0.49 \cdot 0.49 = 0.24$

$P(C \text{ read} \mid G=CG) \cdot P(C \text{ read} \mid G=CG)$

$P(C,C \text{ reads} \mid G=GG) = P(C \text{ read} \mid G=GG)^2 = 0.01^2 = 0.0001$

Most likely Genotype = CC

More Data

$P(C,C,G,C \text{ reads} \mid G=CC) =$

$P(C \text{ read} \mid G=CC) \cdot P(C \text{ read} \mid G=CC) \cdot P(G \text{ read} \mid G=CC) \cdot P(C \text{ read} \mid G=CC) =$

$0.97 \cdot 0.97 \cdot 0.01 \cdot 0.97 = 0.0091$

Bayesian Genotype Calling

Likelihoods: P(Data | Parameter)

$P(C,C \text{ reads} \mid G=CC) = 0.94$

$P(C,C \text{ reads} \mid G=CG) = 0.24$

$P(C,C \text{ reads} \mid G=GG) = 0.0001$

Don't sum to 1 - not Intuitive

More Intuitive

$P(\text{Parameter} \mid \text{Data})$

$$P(G=CC \mid C,C \text{ reads}) = \frac{0.94}{0.94+0.24+0.0001} = 0.797$$

$$P(G=CG \mid C,C \text{ reads}) = \frac{0.24}{0.94+0.24+0.0001} = 0.203$$

$$P(G=GG \mid C,C \text{ reads}) = \frac{0.0001}{0.94+0.24+0.0001} = 0$$

Much more satisfying Legit?

Individual came from Population X

$f_x(C) = 0.1 \quad f_x(G) = 0.9$

$f_x(CC) = 0.1^2 = 0.01$

$f_x(CG) = 2(0.1)(0.9) = 0.18$

$f_x(GG) = 0.9^2 = 0.81$

Bayes' Theorem

likelihoods

Priors

$$P(P_K \mid D) = \frac{P(D \mid P_K)\, P(P_K)}{\sum_{i=1}^{n} P(D \mid P_i)\, P(P_i)}$$

Posterior Probabilities

allele freq & HWE

likelihoods

$P(D \mid G=CC) = 0.94$

$P(D \mid G=CG) = 0.24$

$P(D \mid G=GG) = 0.0001$

Priors (uniform)

$P(G=CC) = \frac{1}{3}$

$P(G=CG) = \frac{1}{3}$

$P(G=GG) = \frac{1}{3}$

Priors (allele freq)

$P(G=CC) = 0.01$

$P(G=CG) = 0.18$

$P(G=GG) = 0.81$

$$P(G=CC \mid D) = \frac{0.94\left(\frac{1}{3}\right)}{0.94\left(\frac{1}{3}\right) + 0.24\left(\frac{1}{3}\right) + 0.0001\left(\frac{1}{3}\right)} = 0.797$$

$P(G=CG \mid D) = 0.203$

$P(G=GG \mid D) = 0$

$$P(G=CC \mid D) = \frac{0.94(0.01)}{0.94(0.01) + 0.24(0.18) + 0.0001(0.81)} = 0.178$$

$$P(G=CG \mid D) = \frac{0.24(0.18)}{0.94(0.01) + 0.24(0.18) + 0.0001(0.81)} = 0.820$$

$$P(G=GG \mid D) = \frac{0.0001(0.81)}{0.94(0.01) + 0.24(0.18) + 0.0001(0.81)} = 0.002$$