Y-chromosome haplotypes

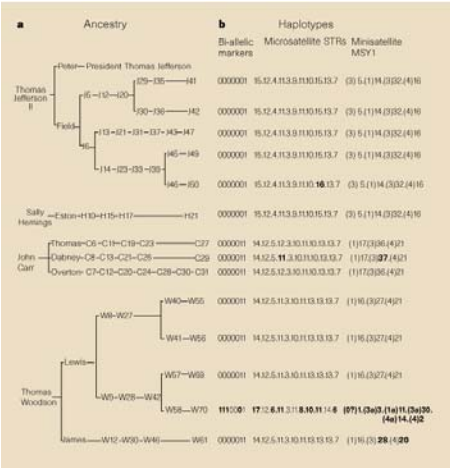


Table I
Dates of Jefferson's Visits to Monticello.

Arrival Date	*Departure Date*	*Stay Length*
Jan. 16, 1794	Feb. 20, 1797	1,131
March 20,1797	May 5, 1797	46
July 11, 1797	Dec. 4, 1797	146
July 3, 1798	Dec. 18, 1798	168
March 8, 1799	Dec. 21, 1799	288
May 29, 1800	Nov. 24, 1800	179
April 5, 1801	April 26, 1801	21
August 2, 1801	Sept. 27, 1801	56
May 8, 1802	May 27, 1802	19
July 25, 1802	Oct. 1, 1802	68
March 11, 1803	March 31, 1803	20
July 24, 1803	Sept. 22, 1803	60
April 5, 1804	May 11, 1804	36
July 27, 1804	Sept. 27, 1804	62
March 17, 1805	April 14, 1805	28
July 18, 1805	Sept. 29, 1805	73
May 9, 1806	June 4, 1806	26
July 24, 1806	Oct. 11, 1806	69
April 11, 1807	May 13, 1807	32
Aug. 5, 1807	Oct. 1, 1807	57
May 12, 1808	June 8, 1808	27

Table II

Birthdays and Associated Statistics for Hemings's Children.

Name	*Birthday*	*IBI (days)*	*Estimated conception date*	*Days from Jefferson's arrival to estimated conception date*	*TJ Age*	*SH Age*
Harriet 1	Oct. 5, 1795	.	Jan. 11, 1795	360	53	22
Beverly	April 1, 1798	909	July 8, 1797	-3	55	25
Daughter	Dec. 7, 1799	615	March 15, 1799	7	57	27
Harriet 2	May 15, 1801	524	Aug. 21, 1800	84	58	28
Madison	Jan. 19, 1805	1345	April 27, 1804	22	62	32
Eston	May 21, 1808	1218	Aug. 28, 1807	23	65	35

Monte Carlo Models

Model 1:
- -Birth date of first child (and conception date) fixed.
- -Randomly pick MC IBIs: ~*U* (*524, 1545*)
- -Use MC IBI's to construct MC conception dates.
- -How many MC conception dates was TJ present for?

Model 2:
- -Same as 1.
- -sort MC IBI's before constructing MC conceptions

Model 3:
- -Choose six MC birth dates, ~*U* (*Oct. 5 1795, May 21, 1808)*
- -MC IBI's > *523.*
- -Use MC IBI's to construct MC conception dates.
- -How many MC conception dates was TJ present for?

Model 4:
- -Same as 3.
- -sort MC IBI's before constructing MC conceptions

Table III
Relative Frequency Distributions for the Number of Conceptions that Fall during or three Days before a Jefferson Visit, for the four Monte-Carlo Models.

Number of Conceptions	*Model 1*	*Model 2*	*Model 3*	*Model 4*
0	0.0%	0.0%	0.1%	0.0%
1	3.6	0.0	6.3	4.3
2	19.1	13.8	24.6	18.3
3	37.3	37.7	36.8	34.3
4	29.2	34.3	24.2	30.6
5	9.7	12.7	7.2	11.1
6	1.2	1.5	0.8	1.3

Bayes' Theorem

How should we use new evidence to modify prior beliefs about the world?

- $$P(H\mid D)=\frac{P(H)\times P(D\mid H)}{P(H)\times P(D\mid H)+P(\overline{H})\times P(D\mid \overline{H})}$$
- $$P(tj\mid 6)=\frac{P(tj)\times P(6\mid tj)}{P(tj)\times P(6\mid tj)+P(\overline{tj})\times P(6\mid \overline{tj})}$$
- $$P(tj\mid 6)=\frac{P(tj)\times 1}{P(tj)\times 1+P(\overline{tj})\times .02}$$

Let's say $P(tj)=.5$

- $$P(tj\mid 6)=\frac{.5}{.5+.02}=.98$$

Two kinds of probability questions:

1. What is the probability of getting a 6 and a 1?

2. What is the probability that TJ was the father of all Hemings's children?

Bayes' Theorem
Derivation

For any two events, H and D…

- $P(H \,\&\, D) = P(H) \times P(D \mid H)$
- $P(D \,\&\, H) = P(D) \times P(H \mid D)$

  Now
- $P(D \,\&\, H) = P(H \,\&\, D)$

  so...
- $P(H) \times P(D \mid H) = P(D) \times P(H \mid D)$

  rearranging...
- $P(P \mid D) = \dfrac{P(H) \times P(D \mid H)}{P(D)}$
- $P(D) = P(H) \times P(D \mid H) + P(\overline{H}) \times P(D \mid \overline{H})$

  so...
- $P(P \mid D) = \dfrac{P(H) \times P(D \mid H)}{P(H) \times P(D \mid H) + P(\overline{H}) \times P(D \mid \overline{H})}$

Ta Da!

Bayes' Theorem
Why the derivation matters…

- 1% of women in their 40's who participate in routine screening have breast cancer.
- 80% of women with breast cancer will get positive mammographies.
- 9.6% of women without breast cancer will also get positive mammographies.
- A woman in this age group had a positive mammography in a routine screening.
- What is the probability that she actually has breast cancer?

- $P(P \mid D) = \dfrac{P(P) \times P(D \mid P)}{P(P) \times P(D \mid P) + P(\overline{P}) \times P(D \mid \overline{P})}$

Bayes' Theorem
Why the derivation matters…

- 1% of women in their 40's who participate in routine screening have breast cancer.
- 80% of women with breast cancer will get positive mammographies.
- 9.6% of women without breast cancer will also get positive mammographies.
- A woman in this age group had a positive mammography in a routine screening.
- What is the probability that she actually has breast cancer?

- $P(P \mid D) = \dfrac{P(P) \times P(D \mid P)}{P(P) \times P(D \mid P) + P(\overline{P}) \times P(D \mid \overline{P})}$
- $P(c \mid m) = \dfrac{P(c) \times P(m \mid c)}{P(c) \times P(m \mid c) + P(\overline{c}) \times P(m \mid \overline{c})}$

Bayes' Theorem
Why the derivation matters…

- 1% of women in their 40's who participate in routine screening have breast cancer.
- 80% of women with breast cancer will get positive mammographies.
- 9.6% of women without breast cancer will also get positive mammographies.
- A woman in this age group had a positive mammography in a routine screening.
- What is the probability that she actually has breast cancer?

- $P(P \mid D) = \dfrac{P(P) \times P(D \mid P)}{P(P) \times P(D \mid P) + P(\overline{P}) \times P(D \mid \overline{P})}$
- $P(c \mid m) = \dfrac{P(c) \times P(m \mid c)}{P(c) \times P(m \mid c) + P(\overline{c}) \times P(m \mid \overline{c})}$
- $P(c \mid m) = \dfrac{.01 \times .8}{.01 \times .8 + .99 \times .096} = \dfrac{.008}{0.008 + .095} = .078$

Bayes' Theorem

- $P(H \mid D) = \dfrac{P(H) \times P(D \mid H)}{P(H) \times P(D \mid H) + P(\overline{H}) \times P(D \mid \overline{H})}$
- **A simpler version:**

$$\underbrace{P(H \mid D)}_{\text{Posterior probability}} \propto \underbrace{P(H)}_{\text{Prior probability}} \times \underbrace{P(D \mid H)}_{\text{Likelihood function}}$$

Posterior probability: P(Hypothesis | Data)

Prior probability: P(Hypothesis)

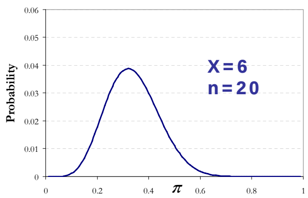
Likelihood function: P(Data | Hypothesis)

- **Or…** *Posterior* $\propto$ *Likelihood* $\times$ *Prior*

Bayesian Methods for Proportions

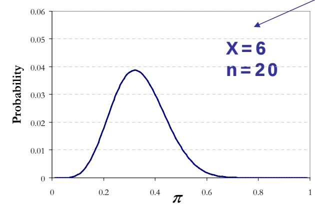
- Beth collects 20 sherds from a 5-foot quadrat.
  6 of them are Creamware
- N=20, X=6
- Assume the data come from a binomial distribution with parameters *x, n,* and π
  x=6
  n=20
  π =?
- so the **likelihood function** is the **binomial distribution function**:

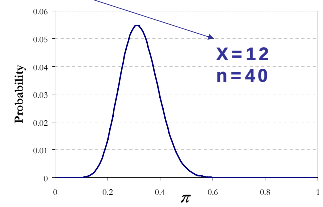
$$P(D\mid H)\sim P(x,n\mid\pi)=\frac{n!}{n!(n-x)!}\pi^{x}(1-\pi)^{n-x}$$

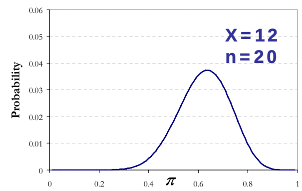


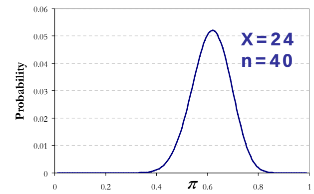
The Likelihood Function

$$P(D\mid H)\sim P(x,n\mid\pi)=\frac{n!}{n!(n-x)!}\pi^{x}(1-\pi)^{n-x}$$



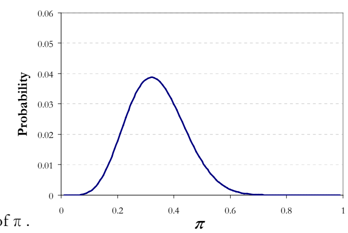






The Likelihood Function

- We estimate the likelihood function using empirical values of x=6, n=20
- What value of π maximizes the likelihood of getting the observed values for x and n.



- $p=\hat{\pi}=x/n$
- This is the Maximum Likelihood Estimate of π .
- To compute (frequentist) confidence limits, using the normal (Gaussian) approximation:

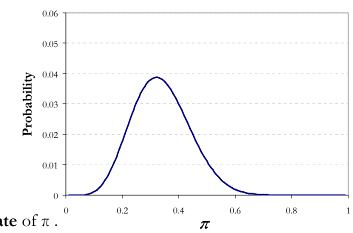
$$s^2=p(1-p)$$

$$p\pm1.96\sqrt{\frac{s^2}{n}}$$

- If we are frequentists, we are done. If Bayesians, we are not! Why?

The Likelihood Function

- We estimate the likelihood function using empirical values of x=6, n=20
- What value of π maximizes the likelihood of getting the observed values for x and n.



- $p=\hat{\pi}=x/n$
- This is the **Maximum Likelihood Estimate** of π .
- To compute (frequentist) confidence limits, using the normal (Gaussian) approximation:

$$s^2=p(1-p)$$

$$p\pm1.96\sqrt{\frac{s^2}{n}}$$

- If we are frequentists, we are done. If Bayesians, we are not! Why?

$$P(H\mid D)\propto P(H)\times P(D\mid H)$$

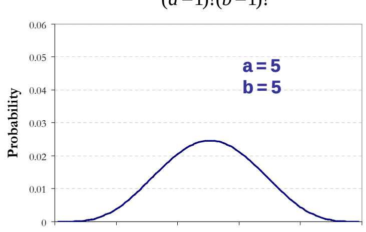
Bayesian Methods for Proportions

- We used the **binomial distribution function** to model the **likelihood:**

$$P(D\mid H)\sim P(x,n\mid\pi)=\frac{n!}{n!(n-x)!}\pi^{x}(1-\pi)^{n-x}$$

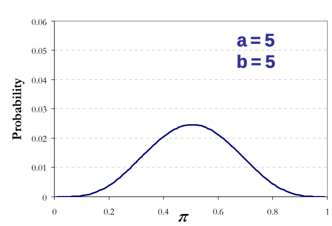
- We use the **beta density function** to model the **prior:**

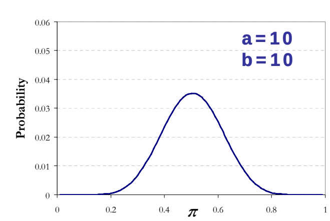
$$P(H)\sim P(\pi)=\frac{(a+b+1)!}{(a-1)!(b-1)!}\pi^{a-1}(1-\pi)^{b-1}$$

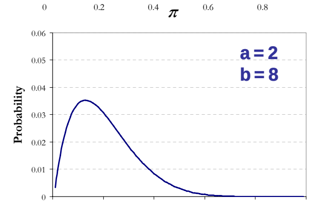


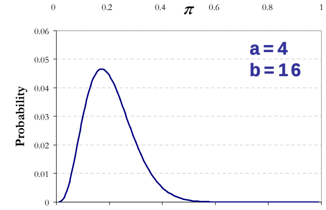
Bayesian Methods for Proportions

- $P(H)\sim P(\pi)=\frac{(a+b+1)!}{(a-1)!(b-1)!}\pi^{a-1}(1-\pi)^{b-1}$



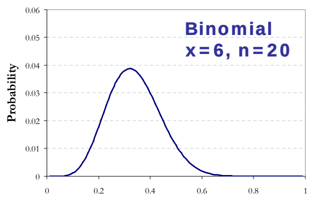




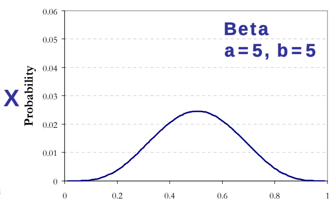


Bayesian Methods for Proportions

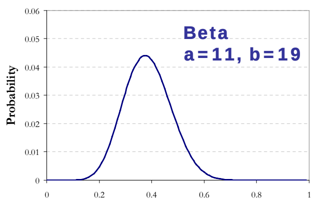
- $P(H\mid D) \propto P(H) \times P(D\mid H)$



X



=



← The posterior distribution

Bayesian Methods for Proportions

Computing the posterior distribution

- $P(H\mid D) \propto P(D\mid H) \times P(H)$
- $P(\pi\mid x,n) \propto P(x,n\mid \pi) \times P(\pi)$
- $P(\pi\mid x,n) = \dfrac{n!}{n!(n-x)!}\pi^{x}(1-\pi)^{n-x} \times \dfrac{(a+b+1)!}{(a-1)!(b-1)!}\pi^{a-1}(1-\pi)^{b-1}$

$$= \frac{(n+a+b+1)!}{(n+a-1)!(n-x+b-1)!}\pi^{x+a-1}(1-\pi)^{n-x+b-1}$$

- The beta is the "conjugate prior" for the binomial.

Bayesian Methods for Proportions

Computing the moments (mean and variance) of the likelihood, prior, and posterior

- The likelihood: $p = x/n$

$$s^2 = p(1-p)$$

- The prior: $\pi = \dfrac{a}{a+b}$

$$\sigma^2 = \frac{\pi(1-\pi)}{a+b+1}$$

- The posterior:

$$\pi' = \frac{x+a}{n+a+b}$$

$$\sigma'^2 = \frac{\pi'(1-\pi')}{n+a+b+1}$$

- The Bayesian posterior 95% ***probability*** interval:

$$\pi' \pm 1.96\sigma'$$

Bayesian Methods for Proportions

- Back to Beth and her surface collection…
  She knows x=6 and n=20.

- How can she figure out *a* and *b* for the prior?
  1. Confess total ignorance: Let *a*=1, *b*=1 – a uniform prior.
  2. Choose values that look cool on the graph.
  3. Base the prior on prior experience – say the samples she collected at 5 adjacent sites:

Site ID	Creamware	Total	p
1	6	10	0.60
2	8	9	0.89
3	10	15	0.67
4	5	6	0.83
5	14	20	0.70

Estimate the mean and variance of the prior, from the data.
Robertson (1999) uses:

$$\hat{\pi} = \frac{\sum x_i}{\sum n_i} \qquad \hat{\sigma}^2 = \frac{n_i\sum(p_i-\hat{\pi})^2}{\sum n_i} \ *$$

* For $n_i >= 50$. A crude attempt to get rid of effect of sampling error effects on $\hat{\sigma}^2$

Bayesian Methods for Proportions

- Recall that: $\pi = \dfrac{a}{a+b}$

$$\sigma^2 = \frac{\pi(1-\pi)}{a+b+1}$$

- So… after some heroic 8th-grade algebra:

$$\hat{a} = \hat{\pi}\left[\frac{\hat{\pi}(1-\hat{\pi})}{\hat{\sigma}^2} - 1\right]$$

$$\hat{b} = (1-\hat{\pi})\left[\frac{\hat{\pi}(1-\hat{\pi})}{\hat{\sigma}^2} - 1\right]$$

- Using estimate of *a* and *b,* estimate the mean and variance of the posterior:

$$\hat{\pi}' = \frac{x+\hat{a}}{n+\hat{a}+\hat{b}}$$

$$\hat{\sigma}'^2 = \frac{\hat{\pi}'(1-\hat{\pi}')}{n+\hat{a}+\hat{b}+1}$$

Question: What are the effects of variation in the sample size and variance of the prior on our posterior estimates?

Bayesian Methods for Proportions

Site ID i	Creamware x	Total n	p	$n_i\sum(p_i-\hat{\pi})^2$
1	6	10	0.60	0.1361
2	8	9	0.89	0.2669
3	10	15	0.67	0.0375
4	5	6	0.83	0.0817
5	14	20	0.70	0.0056
Total	43	60		
Weighted Mean			0.72	0.0088

- Estimate the moments (mean and variance) of the prior:

$$\hat{\pi} = \frac{\sum x_i}{\sum n_i} = .72 \qquad \hat{\sigma}^2 = \frac{n_i\sum(p_i-\hat{\pi})^2}{\sum n_i} = .0088$$

- Estimate the parameters of the prior:

$$\hat{a} = \hat{\pi}\left[\frac{\hat{\pi}(1-\hat{\pi})}{\hat{\sigma}^2} - 1\right] = .72\left[\frac{.72(1-.72)}{.0088} - 1\right] = 15.82$$

$$\hat{b} = (1-\hat{\pi})\left[\frac{\hat{\pi}(1-\hat{\pi})}{\hat{\sigma}^2} - 1\right] = (1-.72)\left[\frac{.72(1-.72)}{.0088} - 1\right] = 6.26$$

Bayesian Methods for Proportions

- Estimate the mean and variance of the posterior:

$$\hat{\pi}' = \frac{x + \hat{a}}{n + \hat{a} + \hat{b}} = \frac{6 + 15.82}{20 + 15.82 + 6.26} = .52$$

$$\hat{\sigma}'^2 = \frac{\hat{\pi}'(1-\hat{\pi}')}{n+\hat{a}+\hat{b}+1} = \frac{.52(1-.52)}{20+15.82+6.26+1} = .0057$$

- The Bayesian posterior 95% ***probability*** interval:

$$\pi' \pm 1.96\sigma' = .52 \pm 1.96 \times .076$$

- Bayesian spatial smoothing:
  1. data: the counts in a given quadrat.
  2. The prior: estimated empirically, based on counts from neighboring quadrats.
  3. The posterior: the smoothed value for the given quadrat.
  4. Iterate 1-3 over all quadrats.

Bayesian Methods for Proportions
A final complication

- Using $\hat{\sigma}^2 = \frac{n_i \sum (p_i - \hat{\pi})^2}{\sum n_i}$ may over-estimate the prior variance.

- VAR(p_i) has <u>two</u> components of variance:
  -sampling error as a function of n_i and p_i;
  -variance of the prior

- We could try: (Martuzzi and Elliott 1996):

$$\hat{\sigma}'^2 = \frac{\frac{n_i \sum (p_i - \hat{\pi})^2}{\sum n_i} - \frac{1}{\bar{n}}(\hat{\pi}(1-\hat{\pi}))}{1 - 1/\bar{n}}$$

- But sometimes the numerator is negative! Ugh…