Lecture 2. June 7 2019

Recap

We described the following ABC method:

1. Generate $\theta \sim \pi(\cdot)$

Recap

We described the following ABC method:

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $\mathcal{D}'$ from the model with parameter θ

Recap

We described the following ABC method:

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $\mathcal{D}'$ from the model with parameter θ
3. Choose a set of summary statistics S of the data
   - Compute $S \equiv S(\mathcal{D})$, and $S' = S(\mathcal{D}')$
   - Accept θ if $\rho(S', S) < \epsilon$, where
     - ρ is a metric on the space of Ss

   Return to [1.]

The connection with sufficiency

Note that

$$\rho(S(\mathcal{D}),\, S(\mathcal{D}')) = 0 \;\not\Longrightarrow\; \mathcal{D} = \mathcal{D}'$$

We are employing a dimension-reduction strategy here. There is a connection with sufficiency.

Recall that a statistic $S = S(\mathcal{D})$ is *sufficient* for the parameter θ if

$$\mathbb{P}(\mathcal{D}|S, \theta) \text{ is independent of } \theta$$

If S is sufficient for θ, then

$$f(\theta|\mathcal{D}) \quad \propto \quad \mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta)$$

If S is sufficient for θ, then

$$\begin{aligned} f(\theta|\mathcal{D}) &\propto \mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}, S(\mathcal{D})|\theta)\pi(\theta) \end{aligned}$$

If S is sufficient for θ, then

$$\begin{aligned} f(\theta|\mathcal{D}) &\propto \mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}, S(\mathcal{D})|\theta)\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}|S(\mathcal{D}), \theta)\mathbb{P}(S(\mathcal{D})|\theta)\pi(\theta) \end{aligned}$$

If S is sufficient for θ, then

$$\begin{aligned} f(\theta|\mathcal{D}) &\propto \mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}, S(\mathcal{D})|\theta)\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}|S(\mathcal{D}), \theta)\mathbb{P}(S(\mathcal{D})|\theta)\pi(\theta) \\ &\propto \mathbb{P}(S|\theta)\pi(\theta) \end{aligned}$$

If S is sufficient for θ, then

$$\begin{aligned} f(\theta|\mathcal{D}) &\propto \mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}, S(\mathcal{D})|\theta)\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}|S(\mathcal{D}), \theta)\mathbb{P}(S(\mathcal{D})|\theta)\pi(\theta) \\ &\propto \mathbb{P}(S|\theta)\pi(\theta) \\ &= f(\theta|S) \end{aligned}$$

If S is sufficient for θ, then

$$\begin{aligned} f(\theta|\mathcal{D}) &\propto \mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}, S(\mathcal{D})|\theta)\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}|S(\mathcal{D}), \theta)\mathbb{P}(S(\mathcal{D})|\theta)\pi(\theta) \\ &\propto \mathbb{P}(S|\theta)\pi(\theta) \\ &= f(\theta|S) \end{aligned}$$

We are really after a notion of *approximate sufficiency*

Research Question: If we had a measure of how close S is to sufficient, we should be able to metrise the difference between $f(\theta|\mathcal{D})$ and $f(\theta|S)$.

A Normal example – 1

Suppose $X_1, X_2, \ldots, X_n$ are iid N(μ, σ^2), with σ^2 known.

A Normal example – 1

Suppose $X_1, X_2, \ldots, X_n$ are iid N(μ, σ^2), with σ^2 known.

$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$ is sufficient for μ

A Normal example – 1

Suppose $X_1, X_2, \ldots, X_n$ are iid N(μ, σ^2), with σ^2 known.

$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ is sufficient for μ

Think of the prior as U(a, b), with $a \to -\infty, b \to \infty$

A Normal example – 1

Suppose $X_1, X_2, \ldots, X_n$ are iid N(μ, σ^2), with σ^2 known.

$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ is sufficient for μ

Think of the prior as U(a, b), with $a \to -\infty, b \to \infty$

The posterior is truncated N$(\bar{X}, \sigma^2/n)$, restricted to (a, b)

A Normal example – 1

Suppose $X_1, X_2, \ldots, X_n$ are iid N(μ, σ^2), with σ^2 known.

$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ is sufficient for μ

Think of the prior as U(a, b), with $a \to -\infty, b \to \infty$

The posterior is truncated N$(\bar{X}, \sigma^2/n)$, restricted to (a, b)

For the ABC method, assume we observed $\bar{X}_0 = 0$. Then

1. Generate $\mu \sim \mathrm{U}(a, b)$
2. Generate $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$
3. Accept μ if $\rho(\bar{X}, \bar{X}_0 = 0) = |\bar{X}| \leq \epsilon$

A Normal example – 2

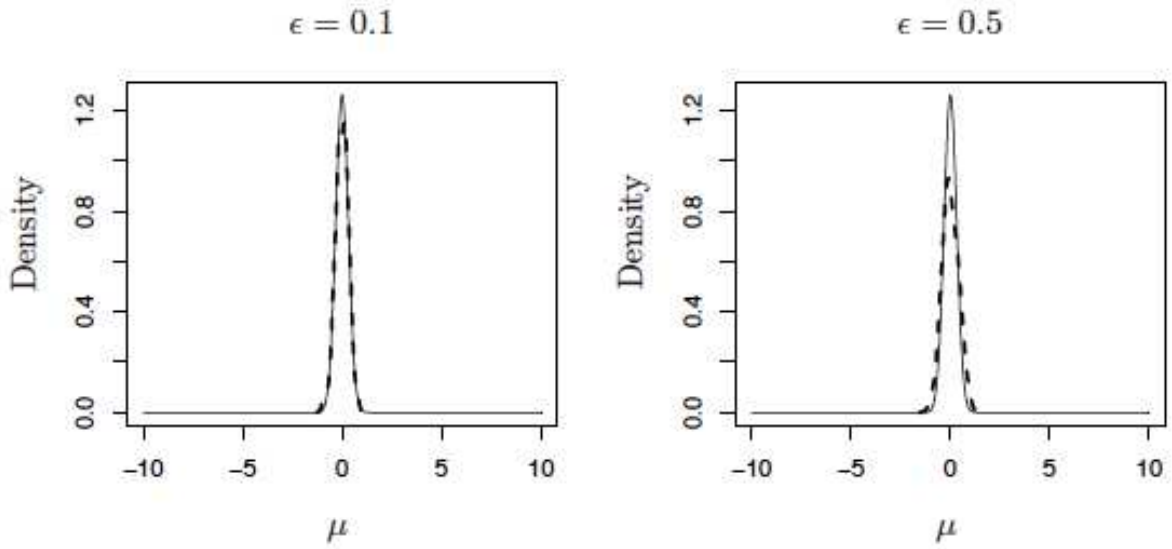


Figure 3.1: *Plots of the true posterior distribution for μ (solid line), and the ABC estimate using 1000 samples (dashed line). A value of $\sigma^2 = 1$ was used for all four plots.*

A Normal example – 3

The ABC density is proportional to

$$\mathbb{1}(a < \mu < b) \int_{-\epsilon}^{\epsilon} (2\pi\sigma^2/n)^{-1/2} \exp(-n(y-\mu)^2/2\sigma^2) dy$$

and the normalizing constant is $\int_a^b () d\mu$. We look at the case where $a \to -\infty, b \to \infty$. Then the normalising constant is 2ϵ, and the density is

$$f_\epsilon(\mu) = \frac{1}{2\epsilon} \left(\Phi\left(\frac{\epsilon - \mu}{\sigma/\sqrt{n}} \right) - \Phi\left(\frac{-\epsilon - \mu}{\sigma/\sqrt{n}} \right) \right)$$

Furthermore,

$$\mathbb{E}(\mu \,|\, |\bar{X}| \leq \epsilon) = 0, \quad \mathrm{Var}(\mu \,|\, |\bar{X}| \leq \epsilon) = \frac{\sigma^2}{n} + \frac{\epsilon^2}{3}$$

A Normal example – 4

Note that the variance shows *overdispersion*: the variance of the ABC method is larger than the true variance

Next we use a Taylor expansion to show that

$$d_{\mathrm{TV}}(f_\epsilon, f) := \frac{1}{2}\int_{-\infty}^{\infty} |f_\epsilon(\mu) - f(\mu)| d\mu \approx \frac{cn\epsilon^2}{\sigma^2} + o(\epsilon^2), \epsilon \to 0$$

Here, $c = \sqrt{\frac{2}{\pi}} e^{-1/2} \approx 0.4839$

Note: This example is from Richard Wilkinson's (2007) DAMTP PhD thesis

Relevant theory from population genetics

Population genetics principles

- Study the effects of mutation, selection, recombination . . . on the structure of genetic variation in natural populations
- Allow for demographic effects such as migration, admixture, subdivision and fluctuations in population size

With the advent of molecular data in the early 1970s, paradigm changed from prospective to retrospective

Drosophila allozyme frequency data

- *D. tropicalis* Esterase-2 locus [$n = 298$]

$$234, 52, 4, 4, 2, 1, 1$$

- *D. simulans* Esterase-C locus [$n = 308$]

$$91, 76, 70, 57, 12, 1, 1$$

What is expected under neutrality?

Sewall Wright (vol. 4, p303) argued that

> *. . . the observations do not agree at all with the equal frequencies expected for neutral alleles in enormously large populations*

What is expected under neutrality?

Sewall Wright (vol. 4, p303) argued that

> *. . . the observations do not agree at all with the equal frequencies expected for neutral alleles in enormously large populations*

What was needed . . .

What is expected under neutrality?

Sewall Wright (vol. 4, p303) argued that

> *. . . the observations do not agree at all with the equal frequencies expected for neutral alleles in enormously large populations*

What was needed . . .

. . . the null distribution of the numbers $\boldsymbol{C}(n) = (C_1(n), \ldots, C_n(n))$ of alleles represented j times in a sample of size n

The Ewens Sampling Formula [1972]

The distribution of the counts in a sample of size n with mutation parameter θ:

$$\mathbb{P}[C_j(n) = c_j, j = 1, 2, \ldots, n]$$

$$= \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!}$$

The Ewens Sampling Formula [1972]

The distribution of the counts in a sample of size n with mutation parameter θ:

$$\mathbb{P}[C_j(n) = c_j, j = 1, 2, \ldots, n]$$

$$= \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!}$$

- *D. tropicalis* Esterase-2 locus [$n = 298$]

  234 52 4 4 2 1 1
- *D. simulans* Esterase-C locus [$n = 308$]

  91 76 70 57 12 1 1

Consequences

- Number of types, $K = C_1(n) + \cdots + C_n(n)$, is sufficient for θ

Consequences

- Number of types, $K = C_1(n) + \cdots + C_n(n)$, is sufficient for θ
- Therefore can use conditional distribution of $\boldsymbol{C}(n)$ given K for testing adequacy of model

Consequences

- Number of types, $K = C_1(n) + \cdots + C_n(n)$, is sufficient for θ
- Therefore can use conditional distribution of $\boldsymbol{C}(n)$ given K for testing adequacy of model

Watterson (1977,1978) suggested using the homozygosity statistic $F = \sum_{j=1}^{n} (j/n)^2 C_j(n)$ as a test statistic for neutrality.

Consequences

- Number of types, $K = C_1(n) + \cdots + C_n(n)$, is sufficient for θ
- Therefore can use conditional distribution of $\boldsymbol{C}(n)$ given K for testing adequacy of model

Watterson (1977,1978) suggested using the homozygosity statistic $F = \sum_{j=1}^{n} (j/n)^2 C_j(n)$ as a test statistic for neutrality.

Its distribution can be simulated very rapidly using Chinese Restaurant Process or Feller Coupling, and a rejection method with $\theta = K/\log(n)$

Consequences

- *D. tropicalis* Esterase-2 locus [$n = 298$]

  234 52 4 4 2 1 1 $F = 0.647, 95\% \ (0.226, 0.818)$

- *D. simulans* Esterase-C locus [$n = 308$]

  91 76 70 57 12 1 1 $F = 0.236, 95\% \ (0.244, 0.834)$

Consequences

- *D. tropicalis* Esterase-2 locus [$n = 298$]

  234 52 4 4 2 1 1 $\quad F = 0.647, 95\% \ (0.226, 0.818)$

- *D. simulans* Esterase-C locus [$n = 308$]

  91 76 70 57 12 1 1 $\quad F = 0.236, 95\% \ (0.244, 0.834)$

Beginning of statistical theory for estimators of population quantities like θ: geneticists were using that part of the data least informative for θ

Consequences

- *D. tropicalis* Esterase-2 locus [$n = 298$]

  234 52 4 4 2 1 1 $\quad F = 0.647, 95\%\ (0.226, 0.818)$

- *D. simulans* Esterase-C locus [$n = 308$]

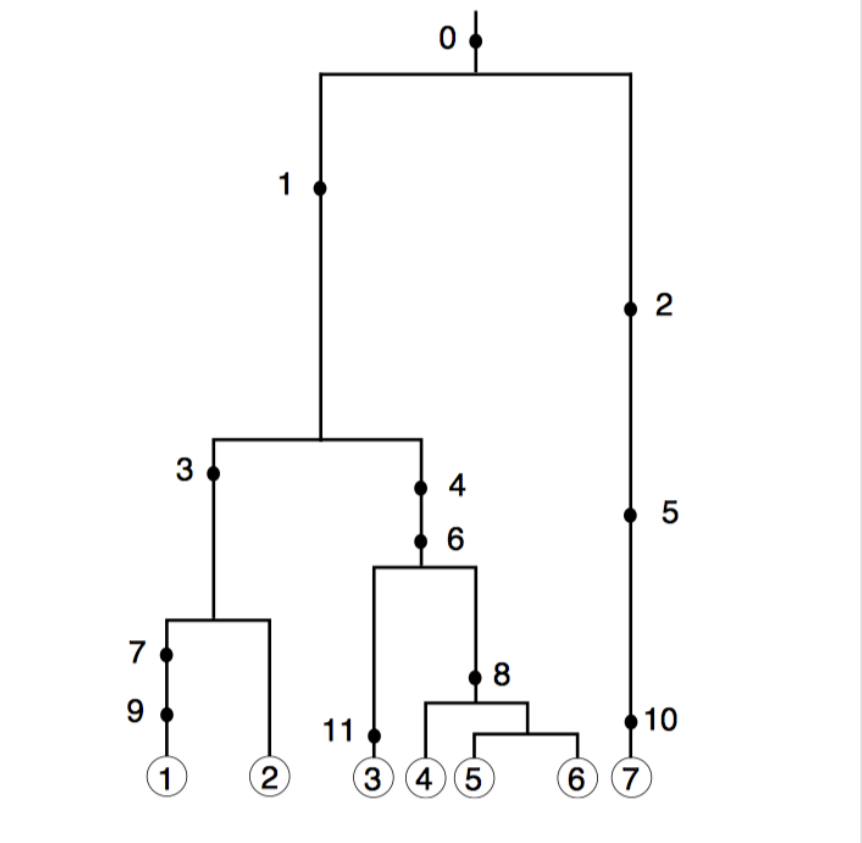
  91 76 70 57 12 1 1 $\quad F = 0.236, 95\%\ (0.244, 0.834)$

Beginning of statistical theory for estimators of population quantities like θ: geneticists were using that part of the data least informative for θ

The Maximum Likelihood Estimator $\hat{\theta}_n$ of θ is asymptotically Normal with mean θ variance proportional to $1/\log n$. The slow rate is because of dependence . . .

The coalescent (Kingman 1982)

Infinitely many sites/alleles model



gene 1	(9,7,3,1,0)
gene 2	(3,1,0)
gene 3	(11,6,4,1,0)
gene 4	(8,6,4,1,0)
gene 5	(8,6,4,1,0)
gene 6	(8,6,4,1,0)
gene 7	(10,5,2,0)

What this led to . . .

- Measure-valued processes, lambda coalescents
- Inference for stochastic processes, including ABC
- Coalescent (and related branching process) simulators
- Applications to statistical genetics, human disease, evolutionary biology
- Non-parametric Bayesian inference, clustering
- Useful paradigm for cancer, in which the individuals are cells (and we model cell division)

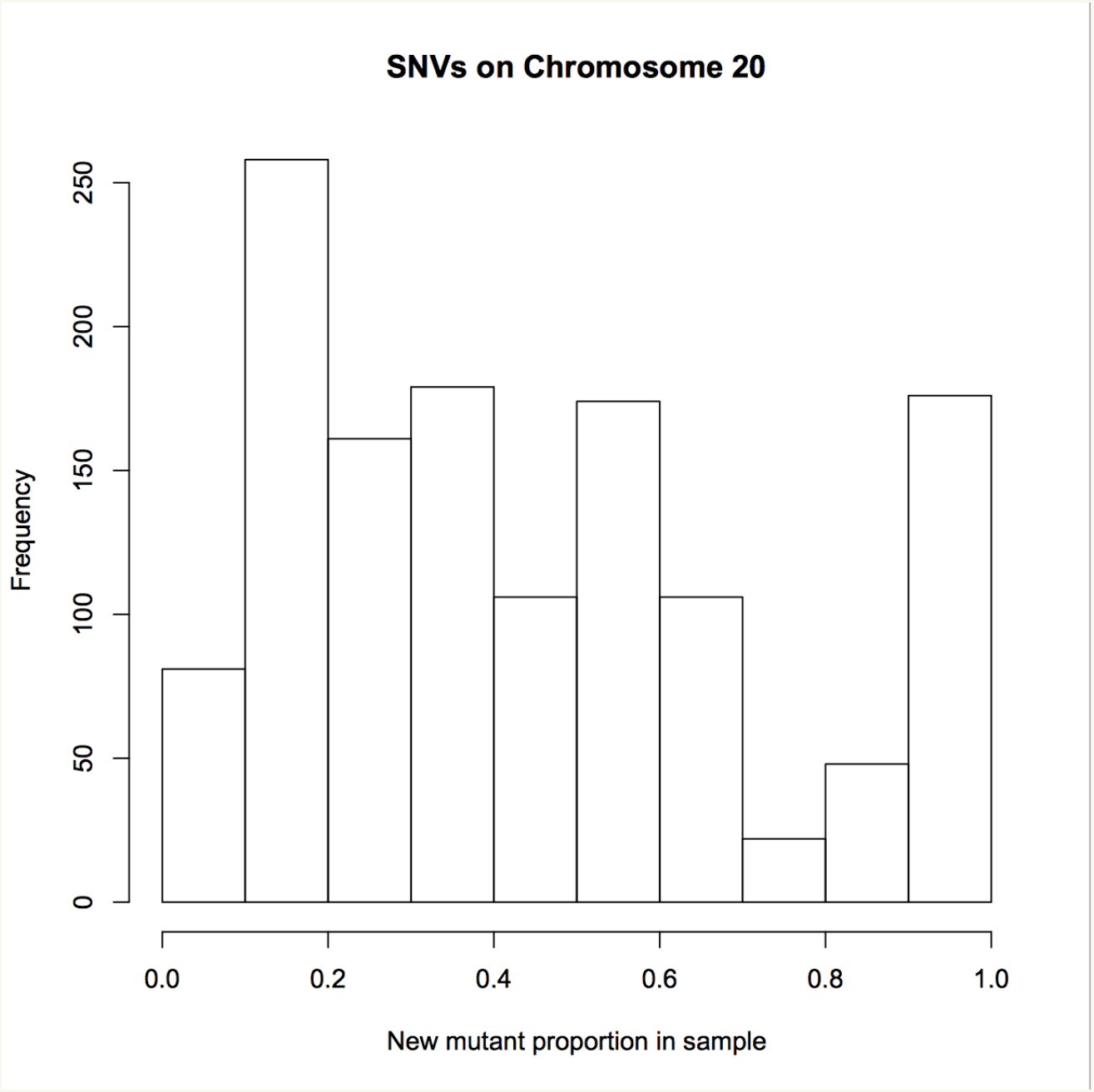
A substantive example from the cancer genomics world (in progress)

Cancer data: Nik-Zainal et al, *Cell*, 2012

Full data: number of SNVs s = 27,719 (the read counts have been processed to provide allele frequency estimates at each SNV)

chr	1	2	3	4	5	6	
#snv	1110	4885	4032	1728	3725	27	
chr	8	9	10	12	13	14	16
#snv	5	164	2702	1	1	3	1253
chr	17	18	19	20	21	22	X
#snv	1286	320	1387	1311	232	243	3304

Binned site frequency spectrum



Another look at sequence data

$$
\begin{array}{c}
 & \begin{array}{ccccccc} m_1 & m_2 & m_3 & m_4 & m_5 & \cdots & m_s \end{array} \\
\begin{array}{c} h_1 \\ h_2 \\ h_3 \\ h_4 \\ \vdots \\ h_n \end{array} &
\left[\begin{array}{ccccccc}
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 1 & \cdots & 1 \\
1 & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & 1 & 0 & \cdots & 0
\end{array}\right]
\end{array}
$$

0 denotes ancestral type, 1 the mutant type

Another look at sequence data

$$
\begin{array}{c}
 & \begin{array}{ccccccc} m_1 & m_2 & m_3 & m_4 & m_5 & \cdots & m_s \end{array} \\
\begin{array}{c} h_1 \\ h_2 \\ h_3 \\ h_4 \\ \vdots \\ h_n \end{array} &
\left[\begin{array}{ccccccc}
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 1 & \cdots & 1 \\
1 & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & 1 & 0 & \cdots & 0
\end{array}\right]
\end{array}
$$

0 denotes ancestral type, 1 the mutant type

The *rows* are haplotypes, the *columns* are sites of mutations

Another look at sequence data

$$
\begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ \vdots \\ h_n \end{array}
\begin{array}{c}
\begin{array}{ccccccc} m_1 & m_2 & m_3 & m_4 & m_5 & \cdots & m_s \end{array} \\
\left[\begin{array}{ccccccc}
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 1 & \cdots & 1 \\
1 & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\
1 & 0 & 0 & 1 & 0 & \cdots & 0
\end{array}\right]
\end{array}
$$

0 denotes ancestral type, 1 the mutant type

The *rows* are haplotypes, the *columns* are sites of mutations

f_i = number of mutant copies at ith site (known)

Another look at sequence data

$$
\begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ \vdots \\ h_n \end{array}
\begin{array}{c}
\begin{array}{ccccccc} m_1 & m_2 & m_3 & m_4 & m_5 & \cdots & m_s \end{array} \\
\left[\begin{array}{ccccccc}
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 1 & \cdots & 1 \\
1 & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\
1 & 0 & 0 & 1 & 0 & \cdots & 0
\end{array}\right]
\end{array}
$$

0 denotes ancestral type, 1 the mutant type

The *rows* are haplotypes, the *columns* are sites of mutations

f_i = number of mutant copies at ith site (known)

K = number of *distinct* haplotypes (not known)

The site frequency spectrum (SFS)

$$
\begin{array}{c}
 & \begin{array}{ccccccc} m_1 & m_2 & m_3 & m_4 & m_5 & \cdots & m_s \end{array} \\
\begin{array}{c} h_1 \\ h_2 \\ h_3 \\ h_4 \\ \vdots \\ h_n \end{array} &
\left[\begin{array}{ccccccc}
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 1 & 0 & 1 & \cdots & 1 \\
1 & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & 1 & 0 & \cdots & 0
\end{array}\right]
\end{array}
$$

Compute

$$
\begin{aligned}
M_b &= \text{number of sites carrying } b \text{ copies of the mutant} \\
&= \#\{l | f_l = b\}, b = 1, 2, \ldots, n-1
\end{aligned}
$$

$\boldsymbol{M}(n) = (M_1, M_2, \ldots, M_{n-1})$ is the observed SFS

Some theory

Some explicit results are known for features of coalescent models, including constant and varying population size

Some theory

Some explicit results are known for features of coalescent models, including constant and varying population size

For example, in the constant population size setting, the number $C_j(n)$ of haplotypes appearing j times in the sample has the ESF with parameter θ

Some theory

Some explicit results are known for features of coalescent models, including constant and varying population size

For example, in the constant population size setting, the number $C_j(n)$ of haplotypes appearing j times in the sample has the ESF with parameter θ

We call $\boldsymbol{C}(n)$ the *haplotype frequency spectrum* (HFS)

Some theory

Some explicit results are known for features of coalescent models, including constant and varying population size

For example, in the constant population size setting, the number $C_j(n)$ of haplotypes appearing j times in the sample has the ESF with parameter θ

We call $\boldsymbol{C}(n)$ the *haplotype frequency spectrum* (HFS)

For many settings, including most branching processes, simulation is required

Some theory

For the SFS, there are corresponding results, the simplest of which is

$$\mathbb{E}M_b = \frac{\theta}{b}, b = 1, 2, \ldots, n-1$$

and many papers have discussed other features of the counts. For example, Dahmer and Kersting (2015) showed that

$$\boldsymbol{M}(n) \Rightarrow (P_1, P_2, \ldots) \quad \text{as } n \to \infty$$

where the P_i are independent Poisson random variables with $\mathbb{E}(P_i) = \theta/i$.

Inference from the SFS

Inference based on the SFS summary statistics arise in several settings:

Inference from the SFS

Inference based on the SFS summary statistics arise in several settings:

One is inference for *Pool-Seq* data (Schloetterer et al, 2012 –)

> This is a whole-genome method for sequencing pools of individuals, and provides a cost- effective alternative to sequencing individuals separately

Inference from the SFS

Inference based on the SFS summary statistics arise in several settings:

One is inference for *Pool-Seq* data (Schloetterer et al, 2012 –)

> This is a whole-genome method for sequencing pools of individuals, and provides a cost- effective alternative to sequencing individuals separately

From such data, we can compute the SFS, but we do not know haplotypes

Inference from the SFS

Inference based on the SFS summary statistics arise in several settings:

One is inference for *Pool-Seq* data (Schloetterer et al, 2012 –)

> This is a whole-genome method for sequencing pools of individuals, and provides a cost- effective alternative to sequencing individuals separately

From such data, we can compute the SFS, but we do not know haplotypes

Problem: infer distribution of K, θ from the SFS

Inference from the SFS

Inference based on the SFS summary statistics arise in several settings:

One is inference for *Pool-Seq* data (Schloetterer et al, 2012 –)

> This is a whole-genome method for sequencing pools of individuals, and provides a cost- effective alternative to sequencing individuals separately

From such data, we can compute the SFS, but we do not know haplotypes

Problem: infer distribution of K, θ from the SFS

We do this in a Bayesian setting (is there any other way?)

Inference from the SFS

We want to compute $\mathcal{L}(K, \theta \mid \boldsymbol{M}(n))$

Inference from the SFS

We want to compute $\mathcal{L}(K, \theta \mid \boldsymbol{M}(n))$

We will focus on a slightly different version, motivated by the fact that in the human sequencing setting, the number of mutations S is often very large

Inference from the SFS

We want to compute $\mathcal{L}(K, \theta \mid \boldsymbol{M}(n))$

We will focus on a slightly different version, motivated by the fact that in the human sequencing setting, the number of mutations S is often very large

Thus we scale and bin the SFS by computing the fractions p_i of sites which have relative frequency of the mutant in ranges determined by bins $B_1, B_2, \ldots, B_r$

Inference from the SFS

We want to compute $\mathcal{L}(K, \theta \mid \boldsymbol{M}(n))$

We will focus on a slightly different version, motivated by the fact that in the human sequencing setting, the number of mutations S is often very large

Thus we scale and bin the SFS by computing the fractions p_i of sites which have relative frequency of the mutant in ranges determined by bins $B_1, B_2, \ldots, B_r$

We are then after

$$\mathcal{L}(K, \theta \mid (p_1, \ldots, p_r))$$

Inference from the SFS

We want to compute $\mathcal{L}(K, \theta \mid \boldsymbol{M}(n))$

We will focus on a slightly different version, motivated by the fact that in the human sequencing setting, the number of mutations S is often very large

Thus we scale and bin the SFS by computing the fractions p_i of sites which have relative frequency of the mutant in ranges determined by bins $B_1, B_2, \ldots, B_r$

We are then after

$$\mathcal{L}(K, \theta \mid (p_1, \ldots, p_r))$$

The combinatorics are not simple here, and explicit results seem hard to come by ...

Back to cancer sequencing: some challenges

- Pooled samples of cells of mixed type are sequenced – a mixture of tumor and non-tumor cells
  - Experiments are rather like Pool-Seq, but the sample size n is not known

- Want to know about number of clones in the sample (akin to K)

Inference from SFS

Now need to model joint prior for (n, θ), and choose a metric

Inference from SFS

Now need to model joint prior for (n, θ), and choose a metric

- Priors for $\theta \sim U(200, 300)$, $n \sim U(100, 300)$
- Bins are $(0.0, 0.1], (0.1, 0.2], \ldots, (0.9, 1.0)$

Inference from SFS

Now need to model joint prior for (n, θ), and choose a metric

- Priors for $\theta \sim U(200, 300)$, $n \sim U(100, 300)$
- Bins are $(0.0, 0.1], (0.1, 0.2], \ldots, (0.9, 1.0)$
- Metric:

$$\rho(p_i, p_i^{\text{obs}}) = \frac{1}{2} \sum_{i=1}^{10} |p_i - p_i^{\text{obs}}|$$

Inference from SFS

Now need to model joint prior for (n, θ), and choose a metric

- Priors for $\theta \sim U(200, 300)$, $n \sim U(100, 300)$
- Bins are $(0.0, 0.1], (0.1, 0.2], \ldots, (0.9, 1.0)$
- Metric:

$$\rho(p_i, p_i^{\text{obs}}) = \frac{1}{2} \sum_{i=1}^{10} |p_i - p_i^{\text{obs}}|$$

- Want $\mathcal{L}(K, n, \theta \mid (p_1, \ldots, p_{10}))$
  - Generated 500,000 samples

Inference from SFS

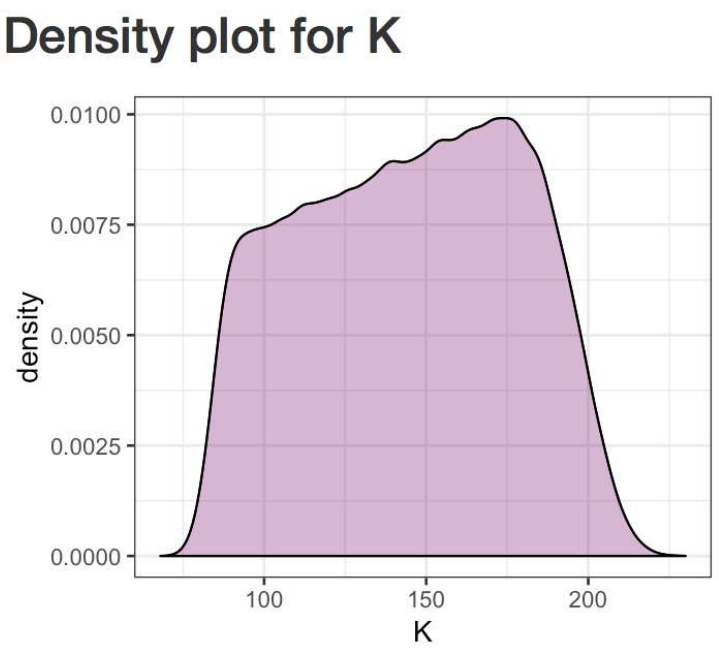
Now need to model joint prior for (n, θ), and choose a metric

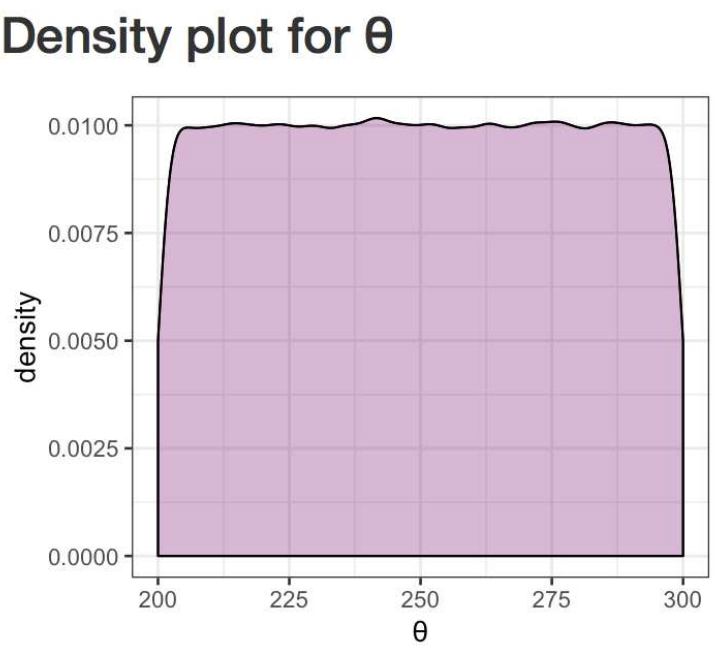
- Priors for $\theta \sim U(200, 300)$, $n \sim U(100, 300)$
- Bins are $(0.0, 0.1], (0.1, 0.2], \ldots, (0.9, 1.0)$
- Metric:

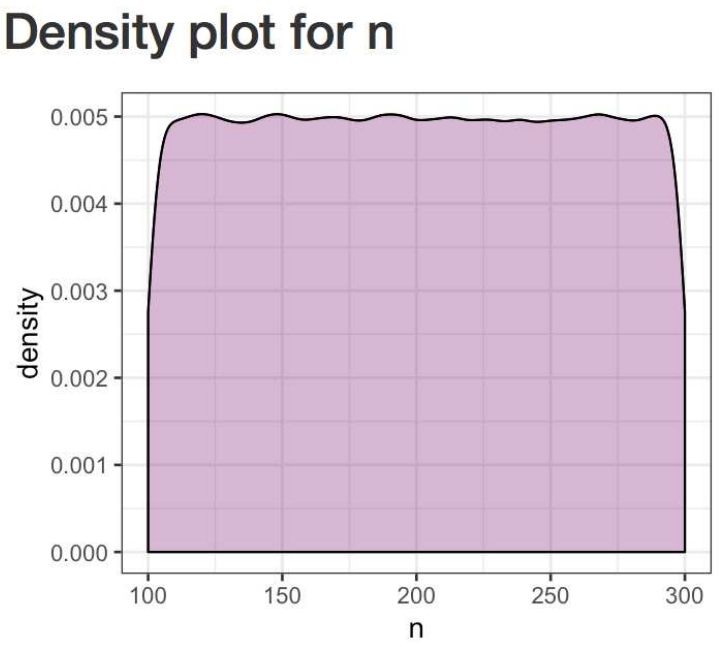
$$\rho(p_i, p_i^{\text{obs}}) = \frac{1}{2} \sum_{i=1}^{10} |p_i - p_i^{\text{obs}}|$$

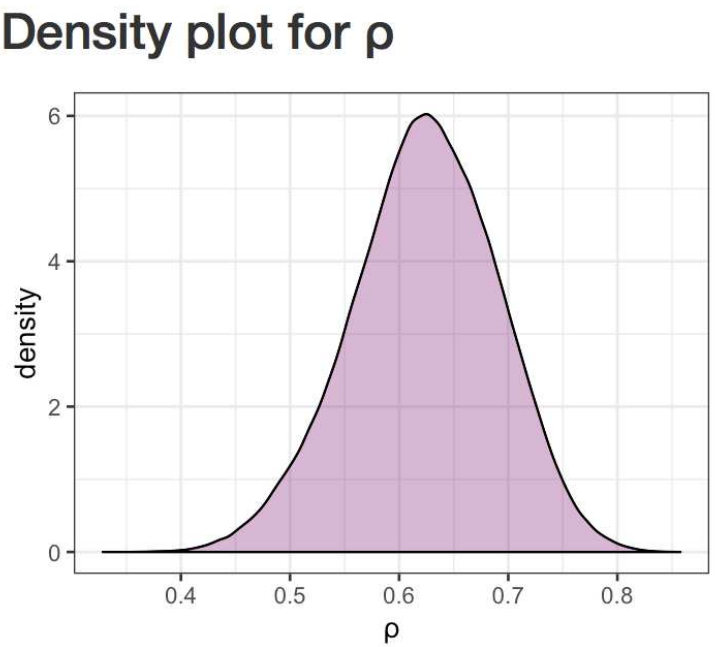
- Want $\mathcal{L}(K, n, \theta \mid (p_1, \ldots, p_{10}))$
  - Generated 500,000 samples
  - The 1% point of ρ values is 0.46

What happened? . . . priors

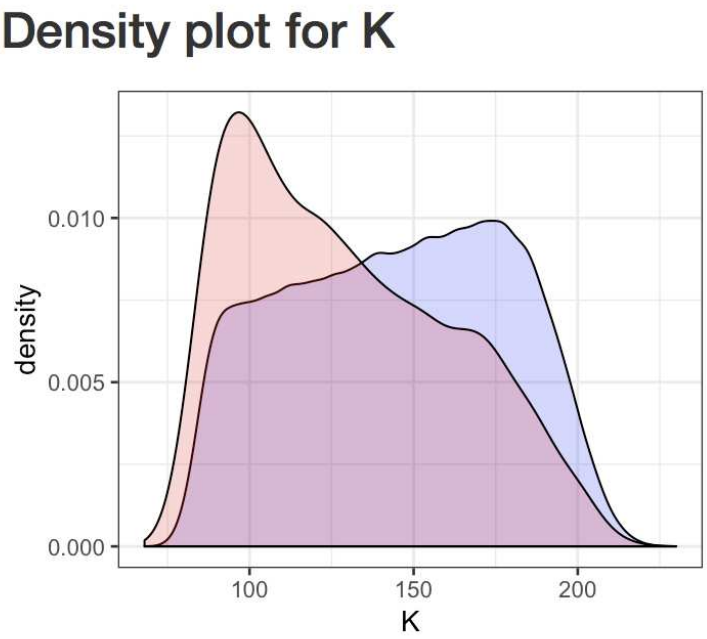


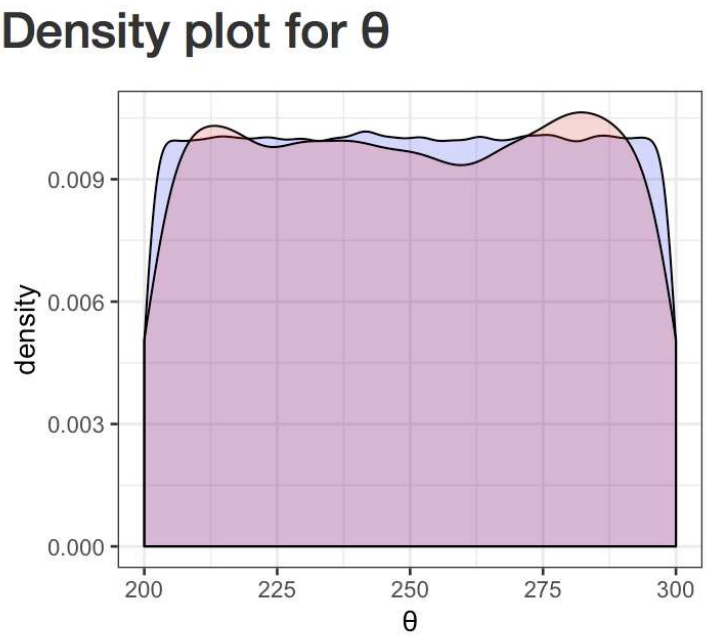


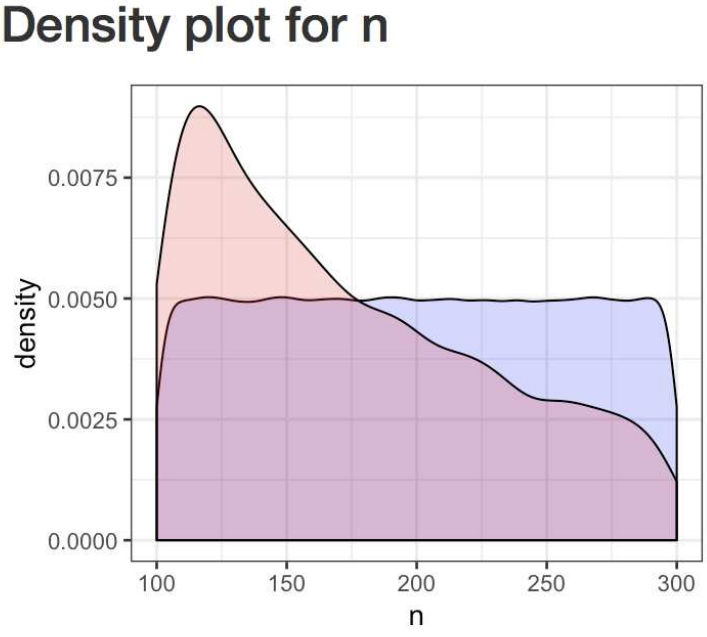


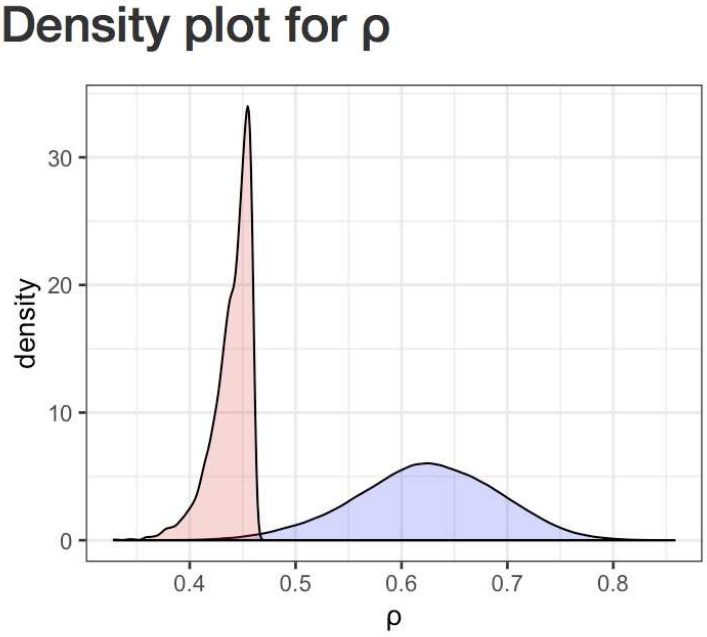


What happened? . . . posteriors









Adding other summary statistics

We want the number of mutations (S) in the simulations to be roughly what was observed in the data

Adding other summary statistics

We want the number of mutations (S) in the simulations to be roughly what was observed in the data

θ is the rate at which mutations arise. In simplest model, $s \sim \theta \log(n)$

Adding other summary statistics

We want the number of mutations (S) in the simulations to be roughly what was observed in the data

θ is the rate at which mutations arise. In simplest model, $s \sim \theta \log(n)$

Recall n is not known (the data come from a pooled sample of cells), but we can estimate the fraction of mutant cells at each site

Adding other summary statistics

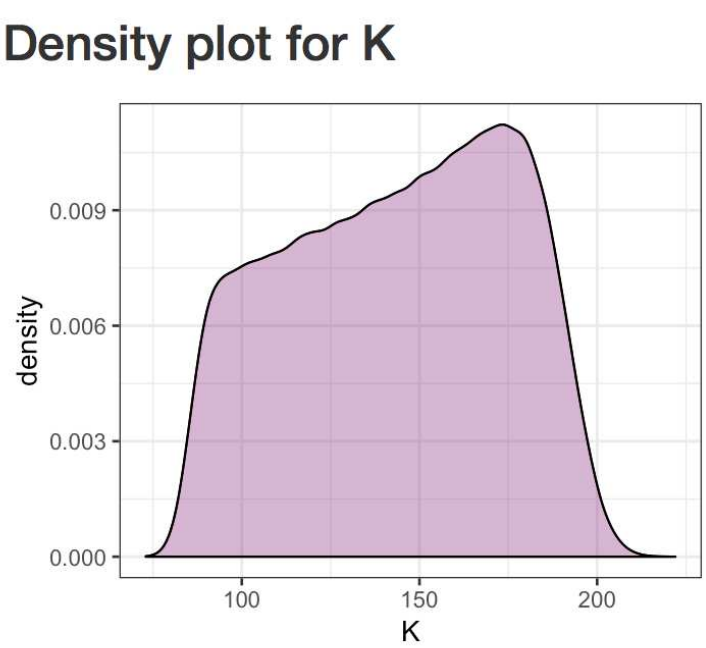
We want the number of mutations (S) in the simulations to be roughly what was observed in the data

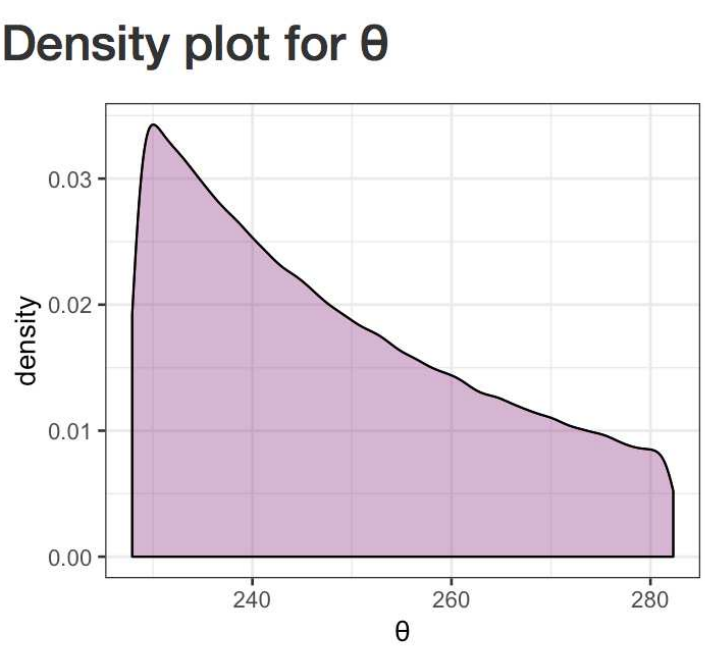
θ is the rate at which mutations arise. In simplest model, $s \sim \theta \log(n)$

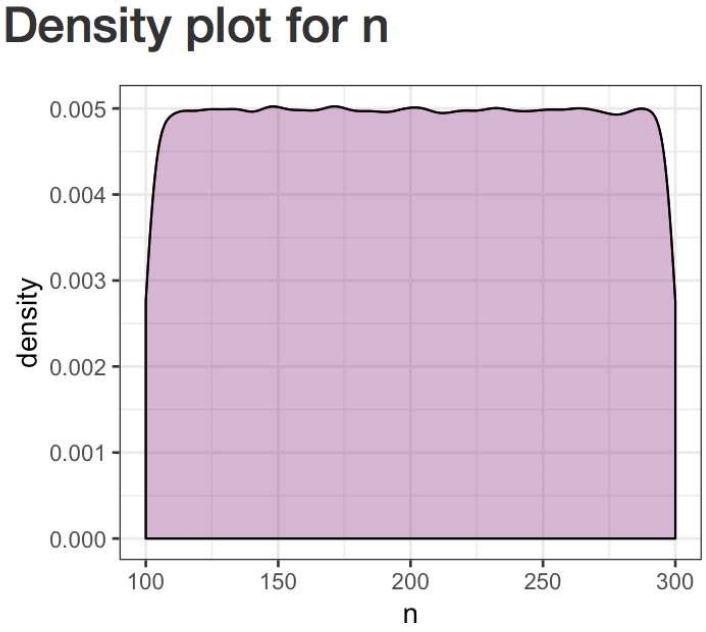
Recall n is not known (the data come from a pooled sample of cells), but we can estimate the fraction of mutant cells at each site

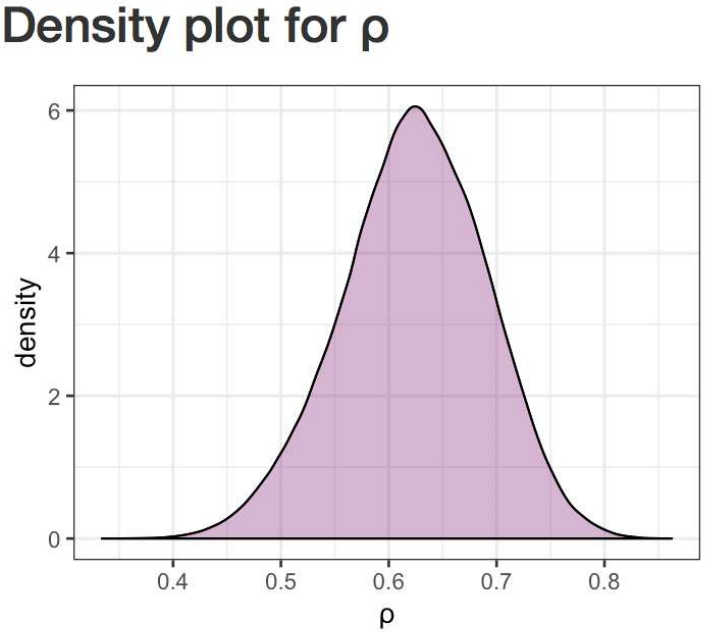
- $s \approx 1300$
- Prior for n: uniform on (100,300)
- $\theta = s/\log(n)$

What happened? . . . priors









What happened? . . . posteriors