Inferring ancient population sizes: An application of coalescent theory
Dem260 Math Demog
Spring 2020
Lecture 12

Joshua R. Goldstein

April 16, 2020

Agenda for today

1. Coalesence when population size is changing
2. Inference from MRCA times
3. ($n > 2$ comparisons and relative branch lengths)
4. Demo: Using real mtDNA to estimate population sizes.

Caveat: We're still using N (not $2N$) as the number of haploids

Last time

- We defined the coalescent as the stochastic process going back in time to common ancestors
- For constant population size, we proved that time to coalescence for a sample pair is exponential.
- We showed (math and simulation) that $E(T) = N$.
- We showed that we could estimate N from observed mutations if we knew the mutation rate
- Last time was models, today mostly application to making inferences about the real history of human populations.

Coalesence when population is changing

- Last time we said hazard of coalescence was $h = c = 1/N$.
- What is hazard of coalesence in one generation for two different populations: $N = 1000$? $N = 2000$?

Coalesence when population is changing

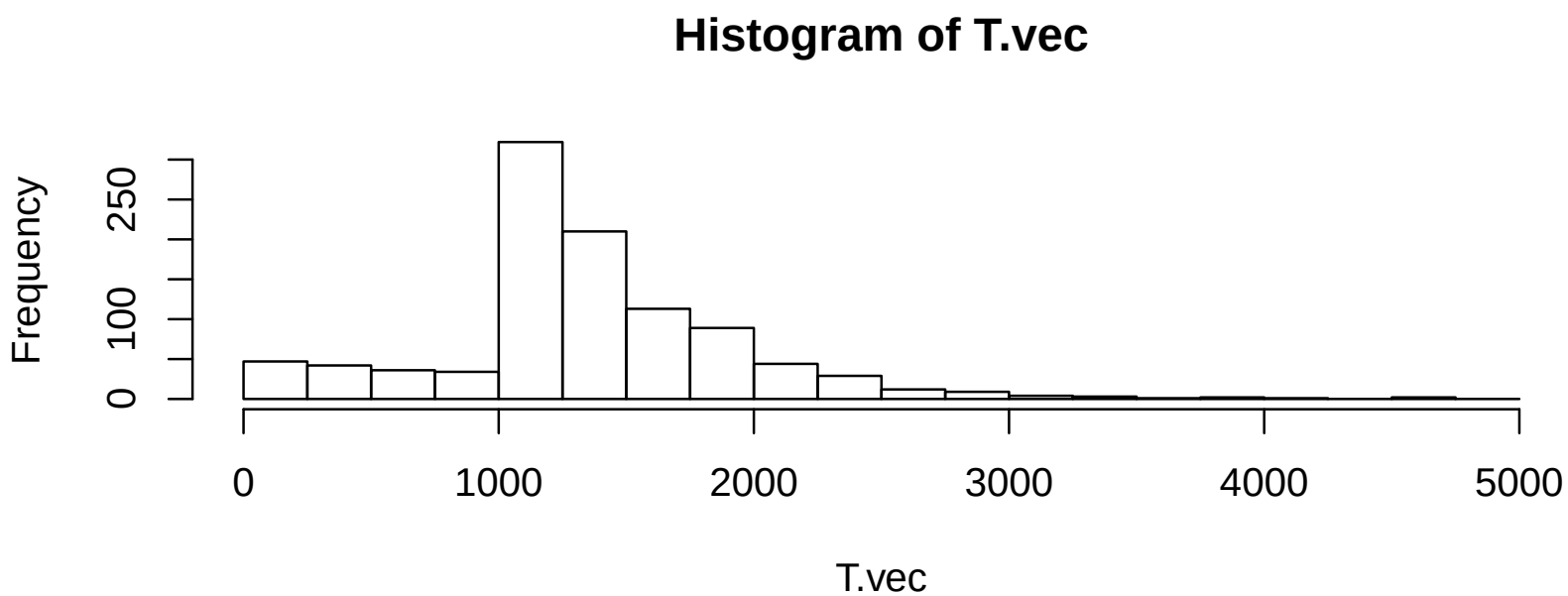
- Last time we said hazard of coalescence was $h = c = 1/N$.
- What is hazard of coalesence in one generation for two different populations: $N = 1000$? $N = 2000$?
- What if within the same population $N(t) = 1000$ and $N(t+1) = 2000$? (Hint: we still follow FW in allowing children to choose their parents.)

Coalesence when population is changing

- Last time we said hazard of coalescence was $h = c = 1/N$.
- What is hazard of coalesence in one generation for two different populations: $N = 1000$? $N = 2000$?
- What if within the same population $N(t) = 1000$ and $N(t+1) = 2000$? (Hint: we still follow FW in allowing children to choose their parents.)
- If the population size changes over time $N(t)$, then hazards of coalescence in will change too: $h(t) = 1/N(t)$.

A small simulation

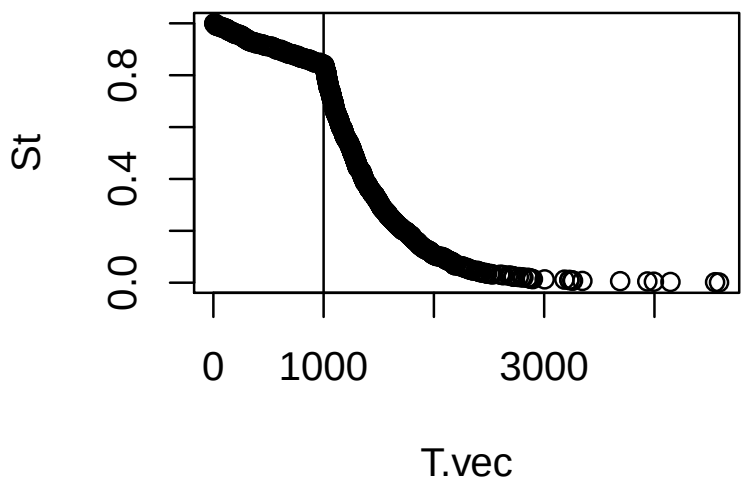
```
N_recent = 5000 ## population last T_thresh years
T_thresh = 1000
N_ancient = 500 ## earlier population
n = 1000 ## sampled individuals
set.seed(0.4886)
T1 <- rexp(n, rate = 1/N_recent) ## give everyone a chance to coalesce
T1[T1 > T_thresh] <- NA ## if they don't in 1st 1000 years, resample them
n2 <- sum(is.na(T1))
T2 <- T_thresh + rexp(n2, rate = 1/N_ancient) ## at ancient rate
T.vec <- c(T1, T2)
hist(T.vec, breaks = seq(0, 5000, 250))
```

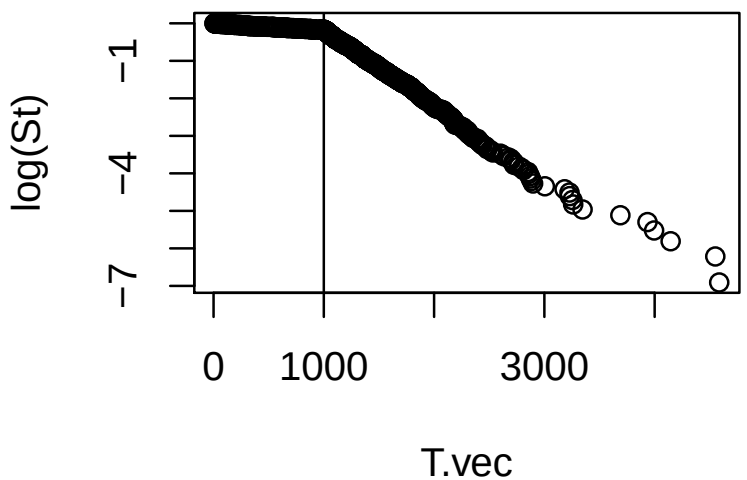


Q: How could we estimate population sizes from this histogram?

A small simulation, Demography Returns!

```
T.vec <- sort(T.vec)
St = (n:1)/n
par(mfrow = c(1,2))
plot(T.vec, St); abline(v = T_thresh); plot(T.vec, log(St)); abl
```





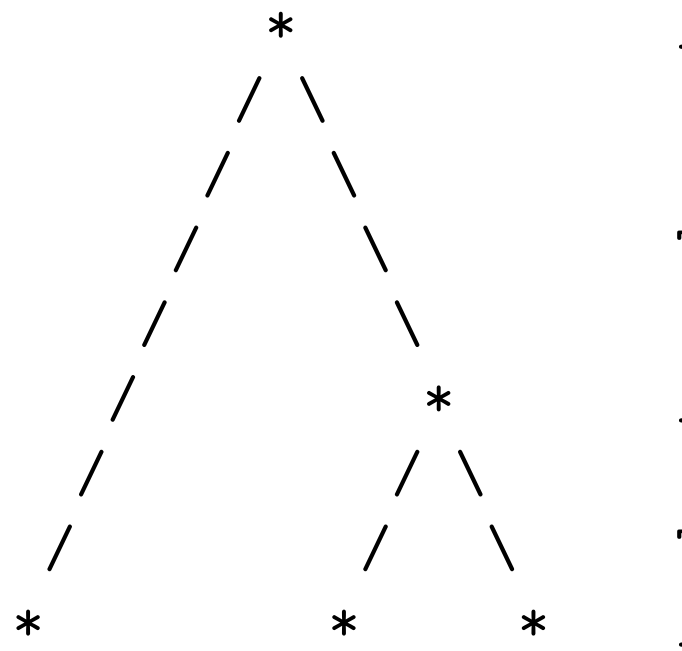
Q: How can we estimate hazards from this histogram?

Our approach

- Say we have i pairs of haploids
- We then compute how many pairwise differences there are, but instead of computing $\bar{k}$, we keep the distributional information k_i.
- Each k_i implies a T_i
- We then have a set of "death times" (coalescence times), can build a life table, estimate the hazards, and infer $N(t)$.

Detour to length of branches

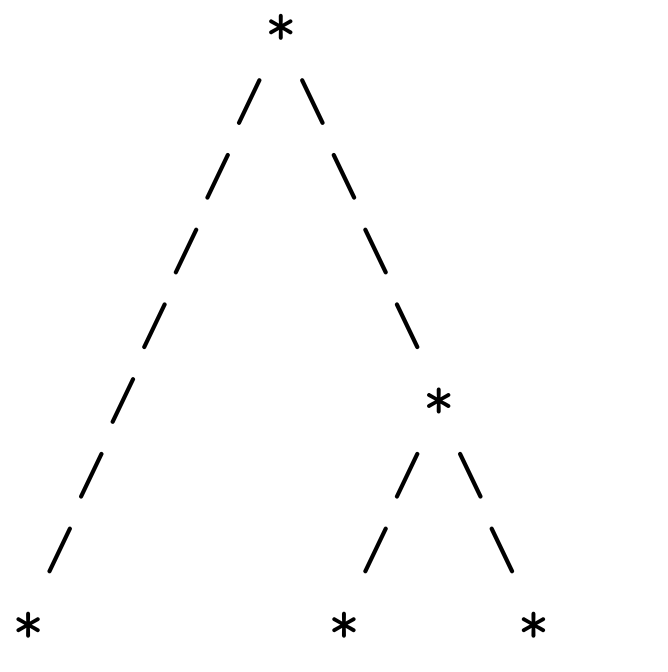
A sample of 3: Note we're using N (not $2N$)



```
-------

T(2) : E(T(2)) = N

-------

T(3) : E(T(3)) = N * 2/[3*2]
-------

 ...

-------
T(n) : E(T(n)) = N * 2/[n * (n-1)]
```

Intuition: when we have more individuals, there's more chance that some pair of them will coalesce.

A sample of 3: Note we're using N (not $2N$)



```
-------

T(2) : E(T(2)) = N

-------

T(3) : E(T(3)) = N * 2/[3*2]
-------

 ...

-------
T(n) : E(T(n)) = N * 2/[n * (n-1)]
```

Question: If we sample 4, how much of time to TMRCA do we have 4 branches, 3 branches, and 2 branches?

Last part: estimating actual historical pop sizes

Let's go to .Rmd in RStudio