

Sampling

- Structured sampling and semi-structured sampling
 - stratified sampling
 - importance sampling
 - Latin hypercube sampling
 - others
- Random sampling and sampling of unknown structure
 - accidental sampling
 - convenience sampling
 - (pseudo)random sampling

Resampling and subsampling

- with replacement
 - bootstrap
- without replacement
 - sub-sampling
 - jackknife

Bootstrap

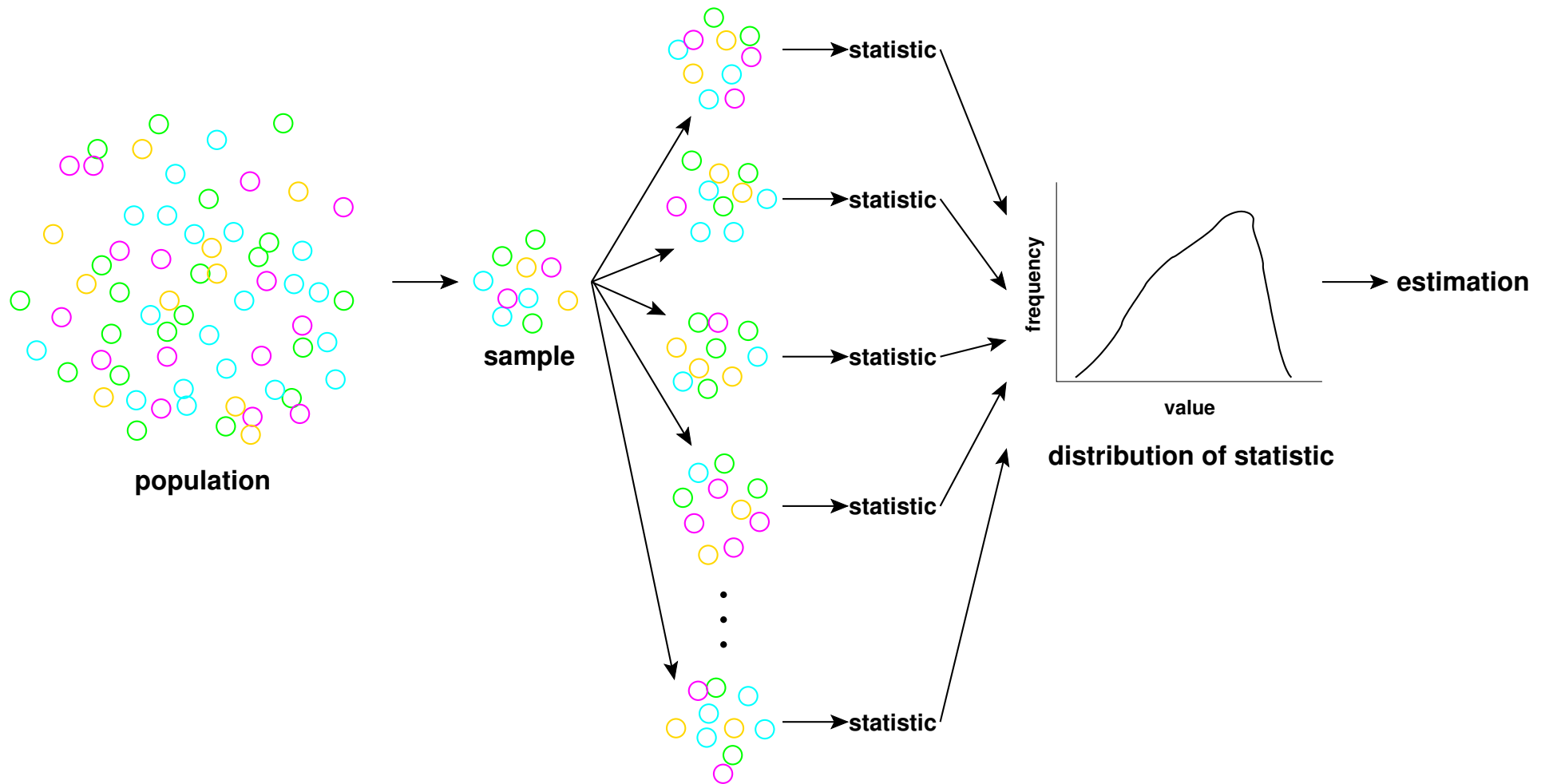
- a non-parametric technique for determining the sampling distribution of statistics based on re-sampling
- re-sampling is done with replacement to generate distributions similar to population sampling when re-sampling the population is not possible or practical
- an inference method used to estimate statistics when they cannot be calculated directly
- assumes that the initial sample is representative of the population, which in statistical terms means the sample is independently and identically distributed (i.i.d.)
- if the assumption is met, the bootstrap distribution accurately approximates the sampling distribution of the population

Bootstrap — Origin of the Term in Statistics

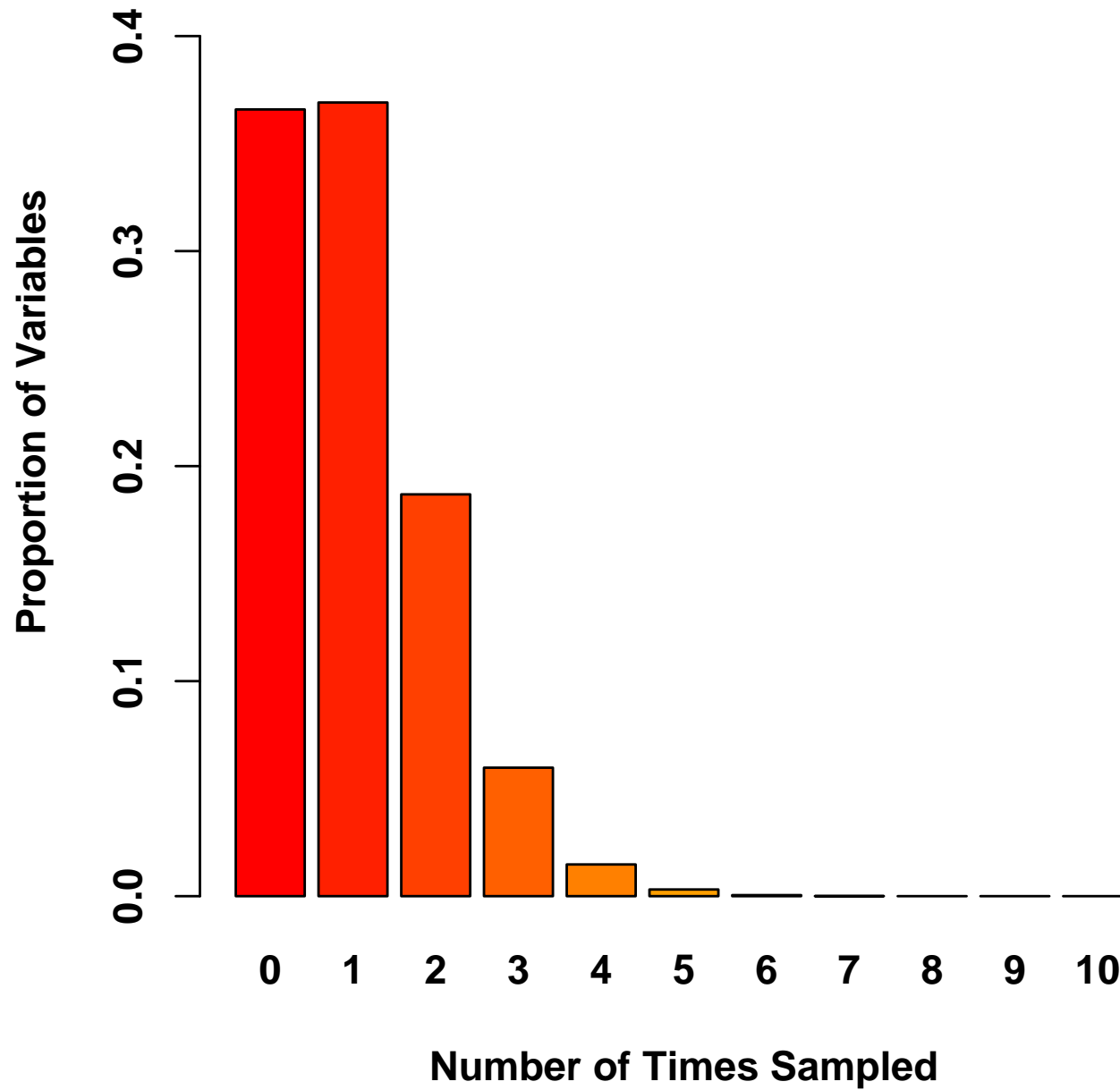
“The use of the term bootstrap derives from the phrase *to pull oneself up by one’s bootstrap*, widely thought to be based on one of the eighteenth century Adventures of Baron Munchausen, by Rudolph Erich Raspe. (The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.) It is not the same as the term ‘bootstrap’ used in computer science meaning to ‘boot’ a computer from a set of core instructions, though the derivation is similar.”

— Bradley Efron and Robert J. Tibshirani (1993)

Bootstrap — Graphical Explanation



Bootstrap Sampling of Variables



Bootstrap: a numerical value to remember

0.632

Generalized Permutation Test Steps¹

1. Analyze the problem.
2. Choose a test statistic.
3. Compute the test statistic for the original labeling of the observations.
4. Rearrange (permute) the labels and recompute the test statistic for the rearranged labels. Repeat until you obtain the distribution of the test statistic for all possible permutations.
5. Accept or reject the hypothesis using this permutation distribution as a guide.

¹Good, P. 1993. Permutation tests; a practical guide to resampling methods and testing hypotheses. Springer-Verlag, New York.

Permutation Test: Some Nuances

- clear specification of the null hypothesis
- design experiment to test null hypothesis accounting for associations not of interest
- test statistic: the obvious in terms of the null hypothesis versus the statistically equivalent, but more computationally convenient and efficient
- permutation algorithm: complete enumeration versus randomization

Permutation Test: Attributes

- applicable to a wide range of problems
- based on the empirical observations at hand
- subsumes any idiosyncrasies of the data
- requires no assumptions about distribution of the data or statistics
- have statistical power is usually equal to the most powerful parametric alternatives when these can be applied
- are exact in that the estimated P -values are accurate (unbiased) with precision determined by the number of permutations evaluated

Empirical Data

- Are the realization of biological processes, which are poorly understood.
- Are often idiosyncratic.
- Have high dimensionality and mixture of characteristics that can provide increased realism in experimental designs.

Simulated Data

- Are the realization of a much simplified model designed to emulate particular characteristics of real data.
- Are often moderately general.
- Have decreased dimensionality, and the control of characteristics can be advantageous in experimental designs.

Two General Uses for Simulation

1. To conduct experiments, including both descriptive and hypothesis testing experiments.
2. To overcome mathematical intractability in quantitative estimation problems (e.g., bootstrap, MCMC).

Monte Carlo Test

1. Analyze the problem.
2. Choose a test statistic and an assumed model.
3. Compute the test statistic for the original set of the observations.
4. Generate samples under the assumed model and calculate the test statistic for each sample. Repeat until you obtain the distribution of the test statistic for an appropriately sized sample.
5. Accept or reject the hypothesis using this distribution as as guide.

Randomization Test

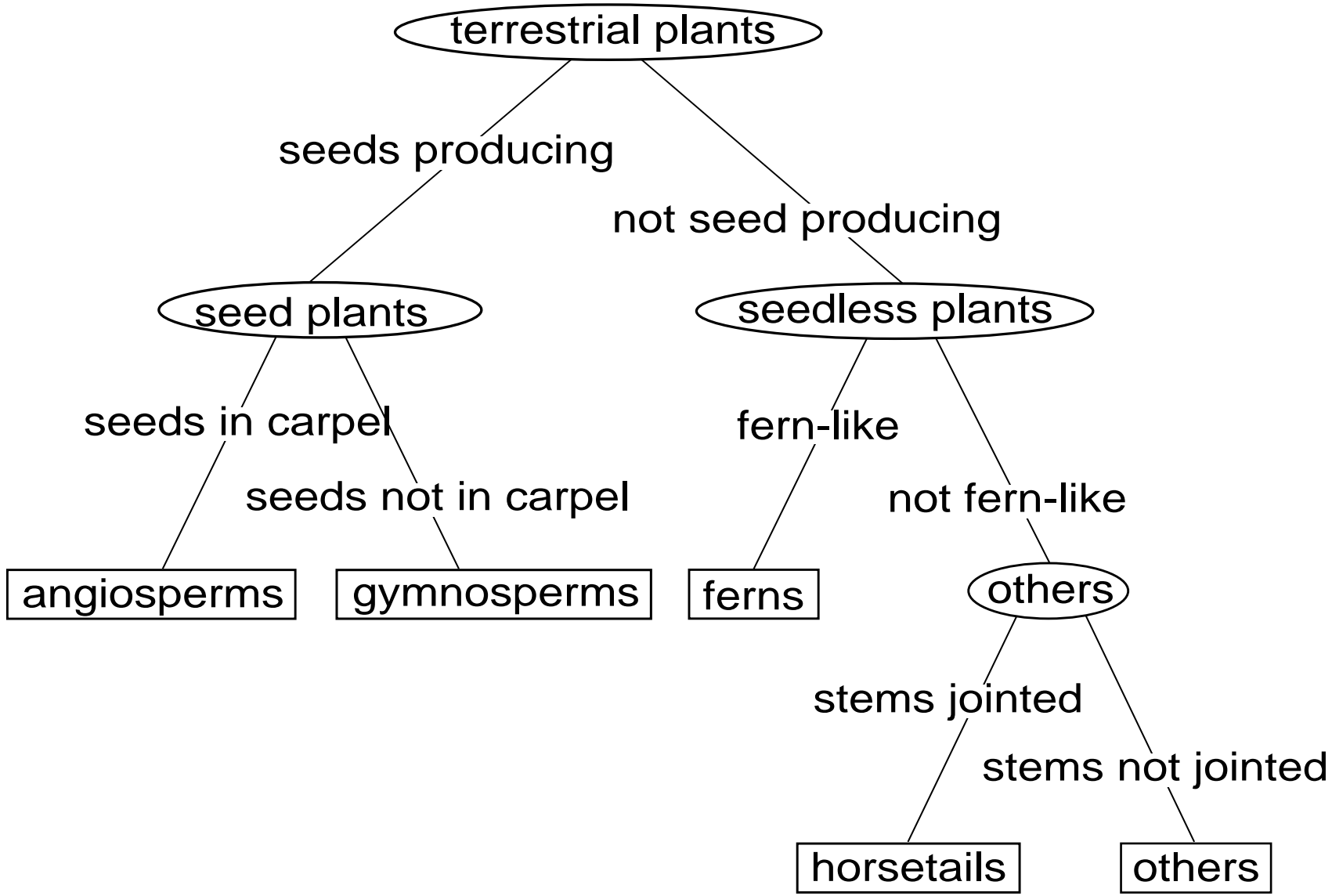
A restricted case of the Monte Carlo Test where the assumed model is random.

Tree-Based Statistical Models

- Recursively partition a data set in two (binary split) based on the value of a single predictor variable
 - to best achieve homogeneous subsets of a categorical response variable (classification)
 - to best separate low and high values of a continuous response variable (regression)

Simple Key to Terrestrial Plants (Embryophyta)

1. Seed producing (Spermatophyta) → 2
 2. Seeds in carpel (Magnoliophyta)
 - 2'. Seeds not in carpel (“Gymnosperms”)
- 1'. Not seed producing → 3
 3. Fern-like (Filicophyta)
 - 3'. Not fern-like → 4
 4. Stems jointed (Equisetophyta)
 - 4'. Stems not jointed → 5



Elements of Initial Tree Growing Procedure¹

- Set of binary questions
 - Question: Is observation $x_i \in A$?
Where A is a region of variable space, X
 - Answer: yes, $x_i \in A$ or no, $x_i \notin A$
- Set of binary questions
- Goodness of fit criterion that can be evaluated for any split
 - Formally, $\Phi(s, t)$, for any split s of any node t
- Stop-splitting rule
- Rule for assigning a value to every node

¹Breiman *et al.* 1984. *Classification and Regression Trees*. Wadsworth & Brooks/Cole.

Random Forests

If a single tree model is good, then an ensemble of tree models should be better.

Random forest attempts to improve upon a simple tree-based statistical model by generating a collection of such models and using them in aggregate.

Random Forest Algorithm (pseudocode)

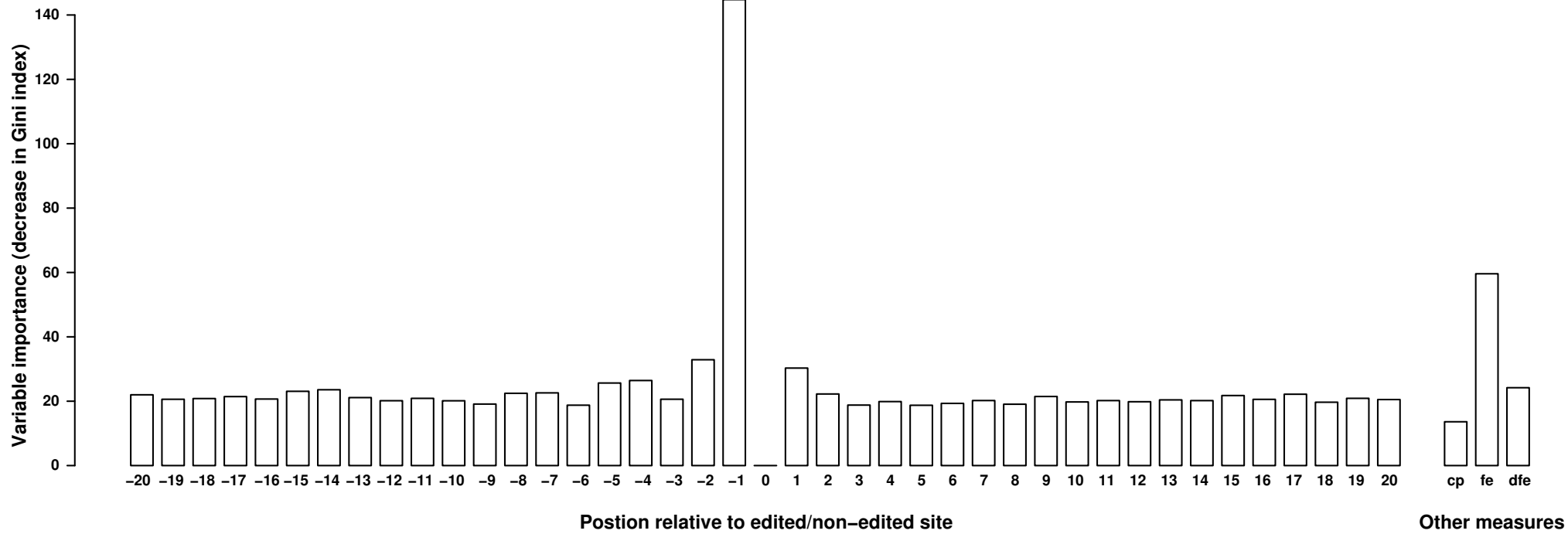
```
for (i= 0; i < N; i++)  
  bootstrap sample observations  
  foreach sample calculate tree model (unpruned)  
    foreach node  
      sample  $k$  predictor variables  
      choose best partition  
    end foreach  
  end foreach  
end for  
combine models  
calculate summary statistics, variable importance, . . .
```

Random Forest: error rate estimation

1. For each bootstrap sample predict the out-of-bag (OOB) observations (i.e., those not within the bootstrap sample).
2. Aggregate out-of-bag predictions and calculate error rate.

Random Forest: variable importance

- How much prediction error increases when out-of-bag data for that variable are permuted (data for other variables is not permuted).
- Decrease in the Gini index, a measure of impurity (classification), or residual sum of squares (regression), induced by splitting the data on a particular variable averaged over all trees.



Advantages of Random Forest

- Relatively low error (perhaps the lowest of any method)
- No over-fitting
- Elegant handling of missing values
- Only partially black-box (e.g., results include variable importance, outlier detection)
- Can be used for supervised and unsupervised learning problems

Flowering Time in Maize — Background

- Quantitative trait locus (QTL) and mutagenesis studies suggested that the *Dwarf8* gene might affect flowering time and plant height
- Thornsberry et al. conducted a field and laboratory study to examine these suggested relationships in maize

Flowering Time in Maize — Principal Questions

- What features of *Dwarf8* are important in flowering time?
- How much of the variance in flowering time can be explained by sequence differences in *Dwarf8*?
- How much of the variance in flowering time can be explained by environment?

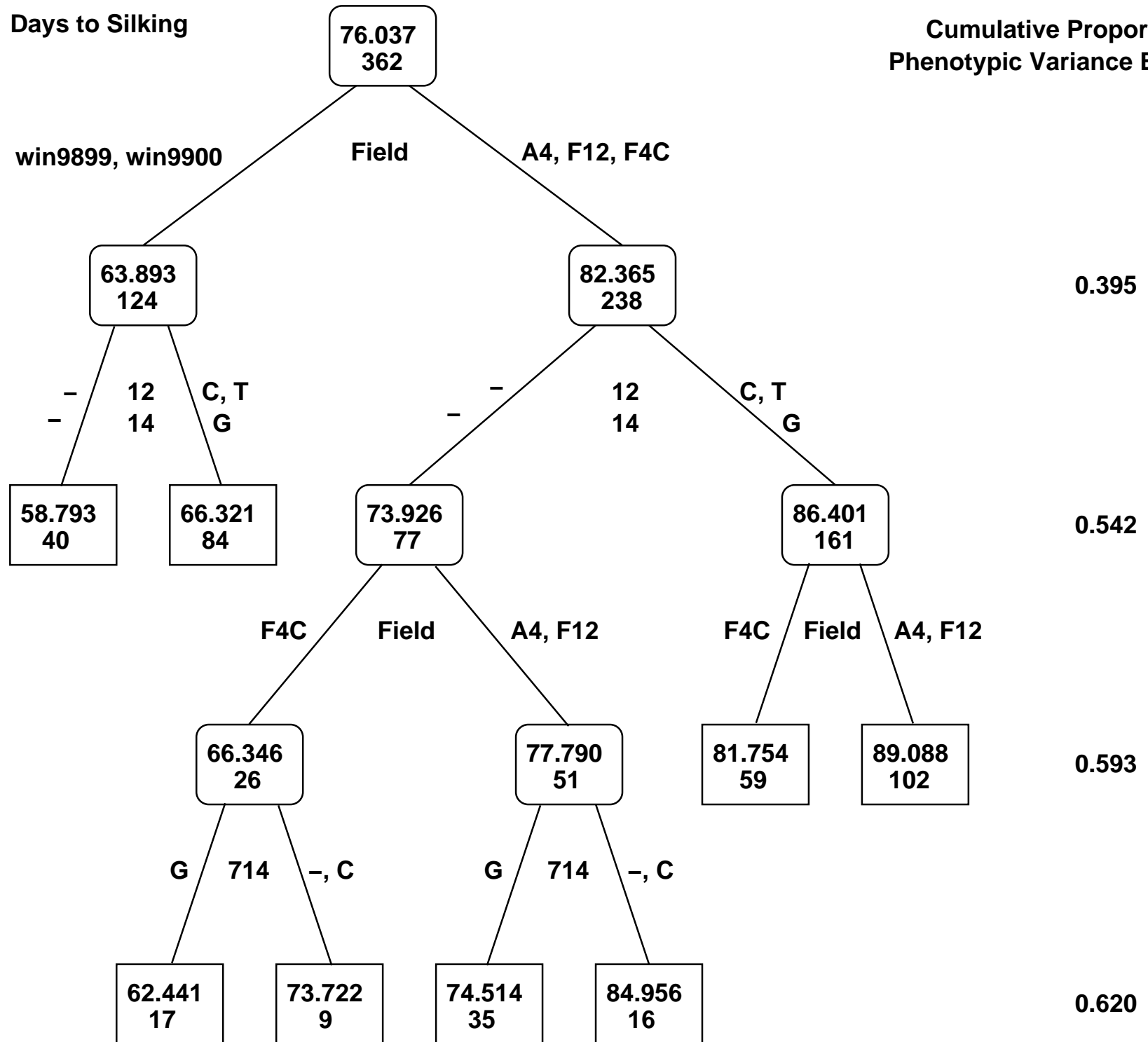
Dwarf8 and Flowering Time — Data⁴

- Genotype: *Dwarf8* and promoter region sequence from inbred lines.
- Phenotype: days to silking, days to pollen (and ear height, plant height)
- Environment: field sites in North Carolina and Florida, USA
- Sample size: 92 inbred lines, 41 haplotypes, 1440 polymorphic sites, 5 environments (field, year combinations)

⁴Thornsberry *et al.* 2001 *Nature Genetics*

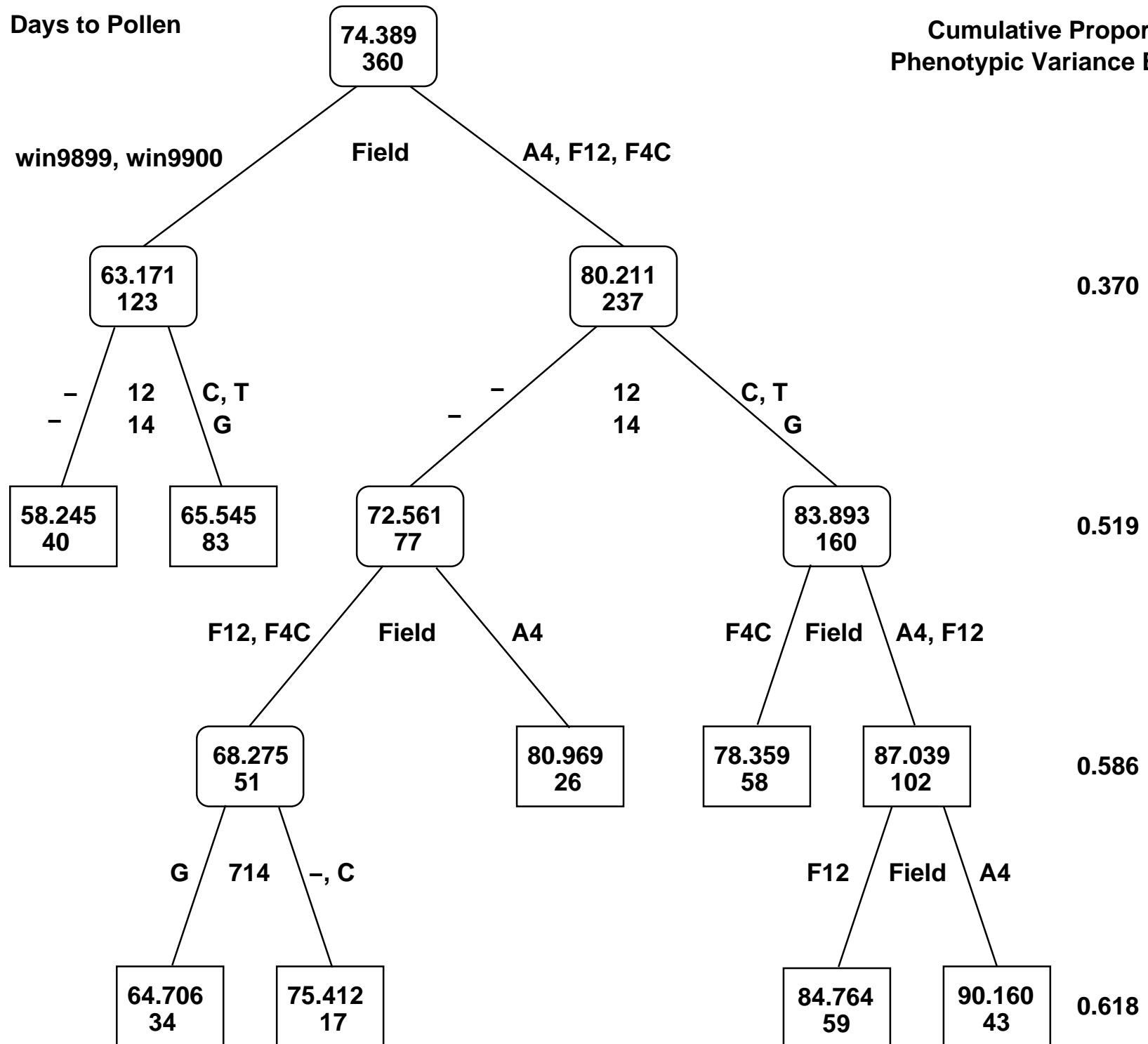
Days to Silking

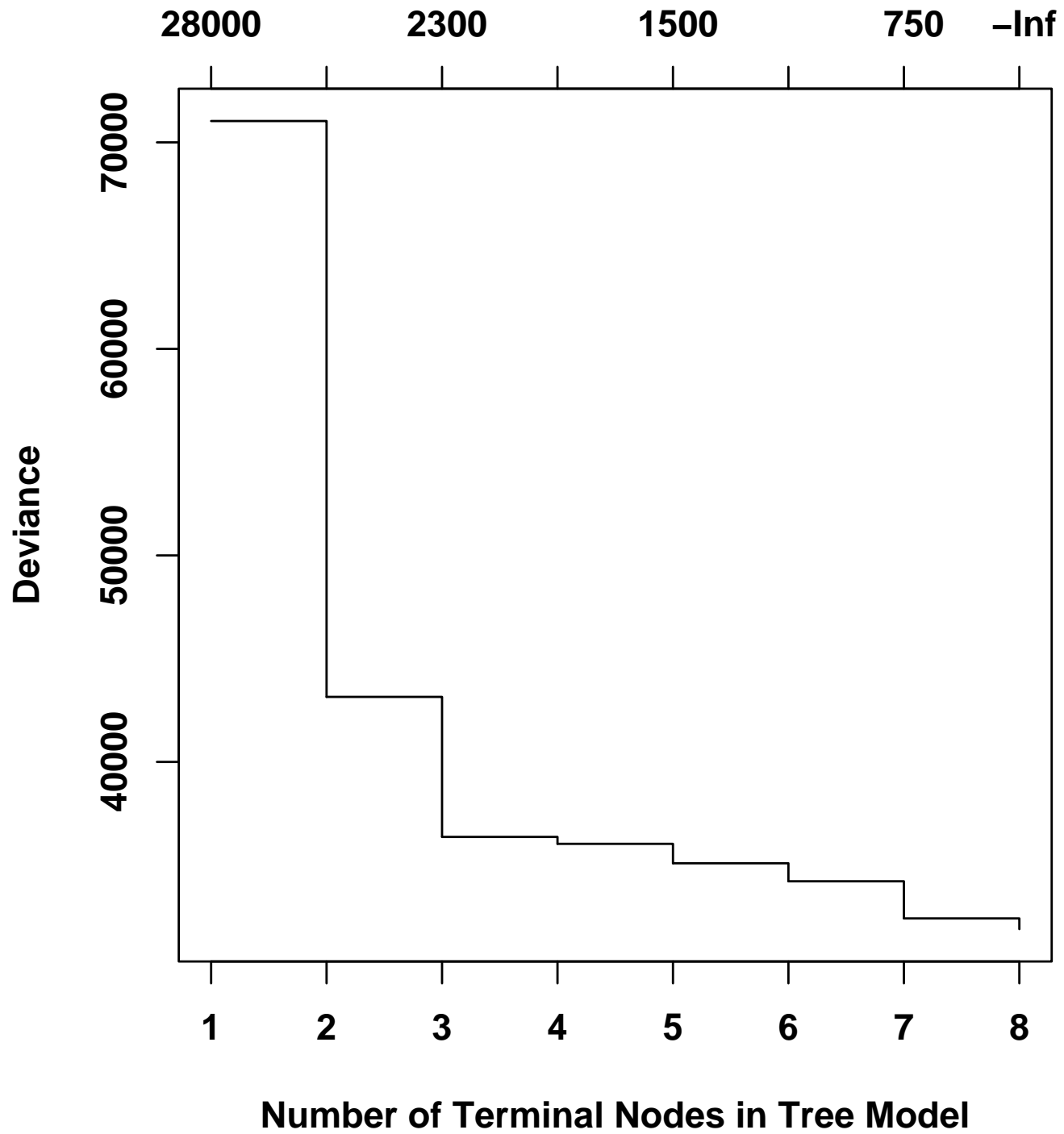
Cumulative Proportion of Phenotypic Variance Explained



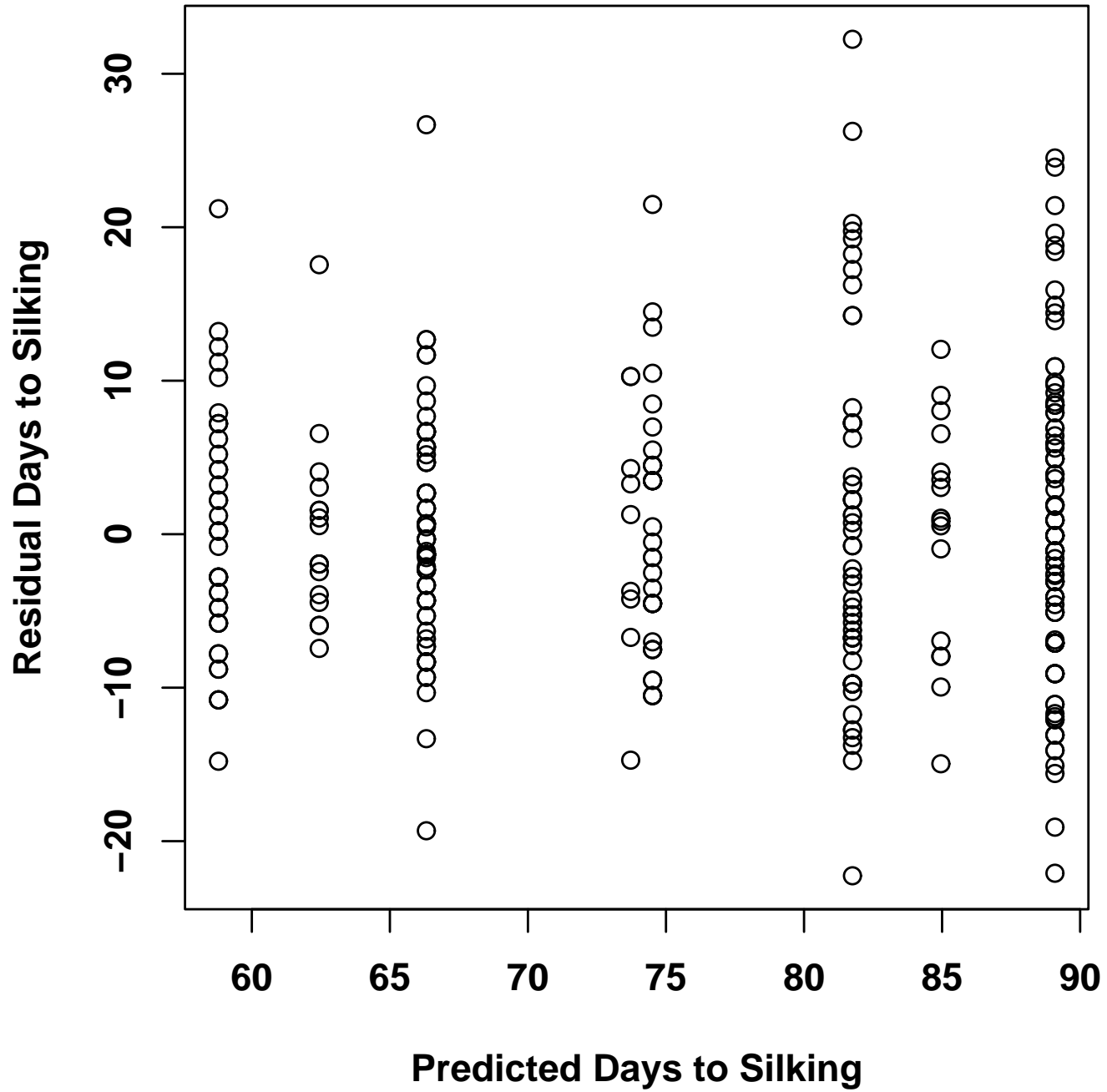
Days to Pollen

Cumulative Proportion of Phenotypic Variance Explained

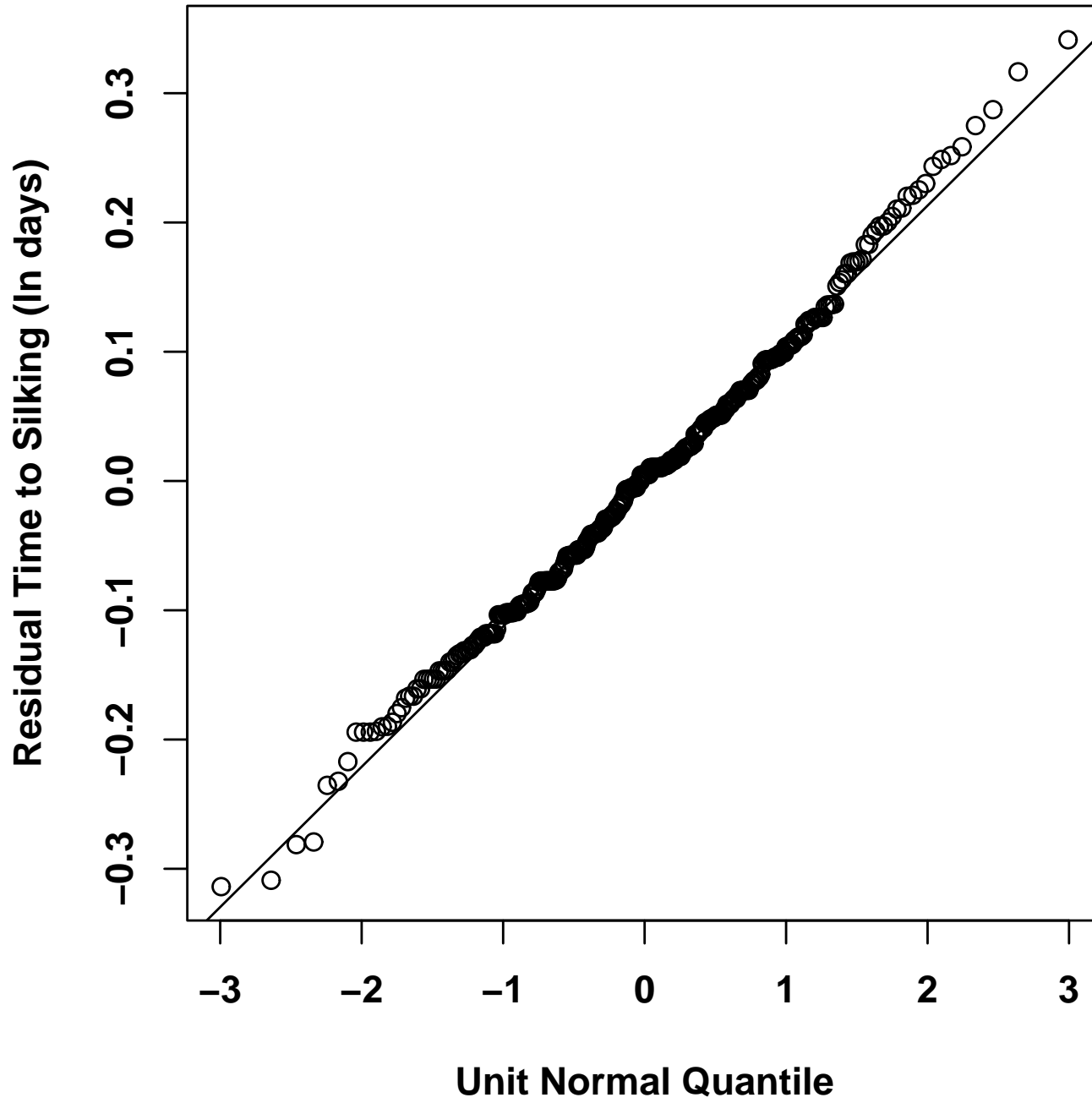




Residuals Plot for Days to Silking



Normal Q-Q Plot for Days to Silking



Proportion of Variance Explained

| | Environment | Genotype | Total |
|-----------------|-------------|----------|-------|
| Days to Silking | 0.45 | 0.17 | 0.62 |
| Days to Pollen | 0.44 | 0.18 | 0.62 |

Flowering Time in Maize — Summary

- Environment has a large effect on flowering time
- Variation in *Dwarf8* region is associated with variation in flowering time
- Very few of the 1440 polymorphic sites in *Dwarf8* region appear to affect flowering time

Random Forest Compared to Single Tree-based Model

| <i>Problem</i> | <i>Single Tree</i> | <i>Random Forest</i> |
|------------------------------|--------------------|----------------------|
| Tuberculosis, regression | 0.679 | 0.861 |
| Tuberculosis, classification | 0.884 | 0.942 |
| RNA editing, classification | 0.705 | 0.848 |

Why do random forests work so well?

- variance reduction from averaging over models
- randomization steps decrease correlation between individual models in the ensemble

Tools

- General programming languages
 - FORTRAN
 - C/C++
 - Perl
- Higher level tools
 - commercial tools: MATLAB, Mathematica, SPlus, SAS, . . .
 - R, a statistical computing system

Suggested Readings

- Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone. 1984. Classification and regression trees. Wadsworth & Brooks/Cole, Pacific Grove.
- Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Monographs in Statistics and Applied Probability 57. Chapman & Hall, London.
- Good, P. 1993. Permutation tests; a practical guide to resampling methods and testing hypotheses. Springer-Verlag, New York.
- Hastie, T., R. Tibshirani and J. Friedman. 2001. The elements of statistical learning; data mining, inference and prediction. Springer, New York.
- Manly, B. F. J. 1991. Randomization and Monte Carlo methods in biology. Chapman & Hall, London.

Suggested Readings, continued

Maritz, J. S. 1995. Distribution-free statistical methods, second edition. Monographs in Statistics and Applied Probability 17. Chapman & Hall, London.

Ripley, B. D. 1987. Stochastic simulation. John Wiley and Sons, New York.