What is (are?) statistics?

Karl W. Broman

8 January 1998

Statistics as **objects**:

numbers used to describe/summarize data

Statistics as a **discipline**:

answering questions using data:

- exploratory data analysis
- inductive inference using probability
- quantifying uncertainty

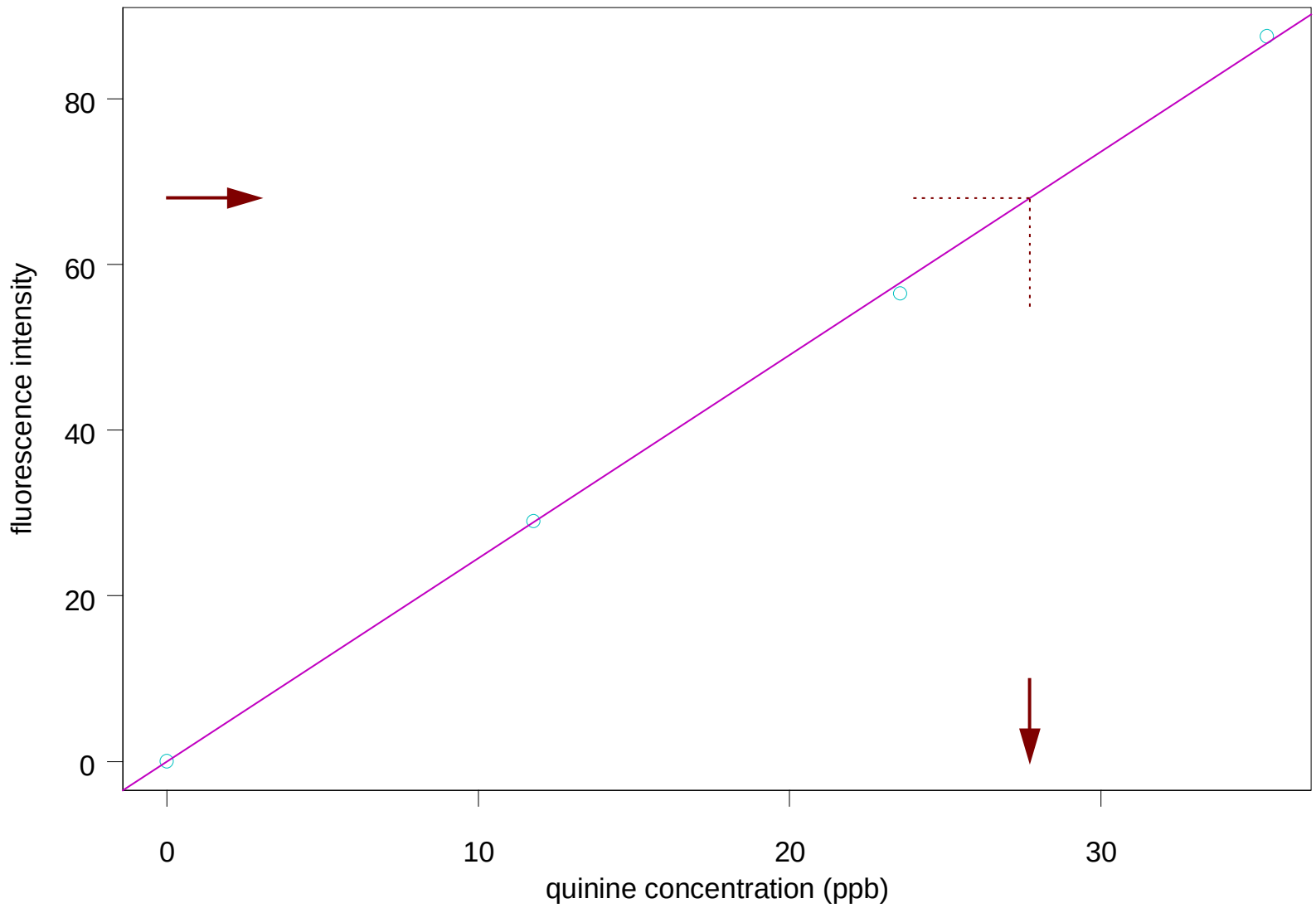
How I became interested in the subject:

Quantitative analytical chemistry:

determination of quinine concentration in tonic water

Experiment:

- prepare a set of solutions with known quinine conc.
- measure the fluorescence of each sample
- do the same for the unknown sample
- fit a line to the data on the standards


fluorescence intensity
0
20
40
60
80
0
10
20
30
quinine concentration (ppb)

Questions:

(1) estimated concentration in sample?

(2) estimated uncertainty?

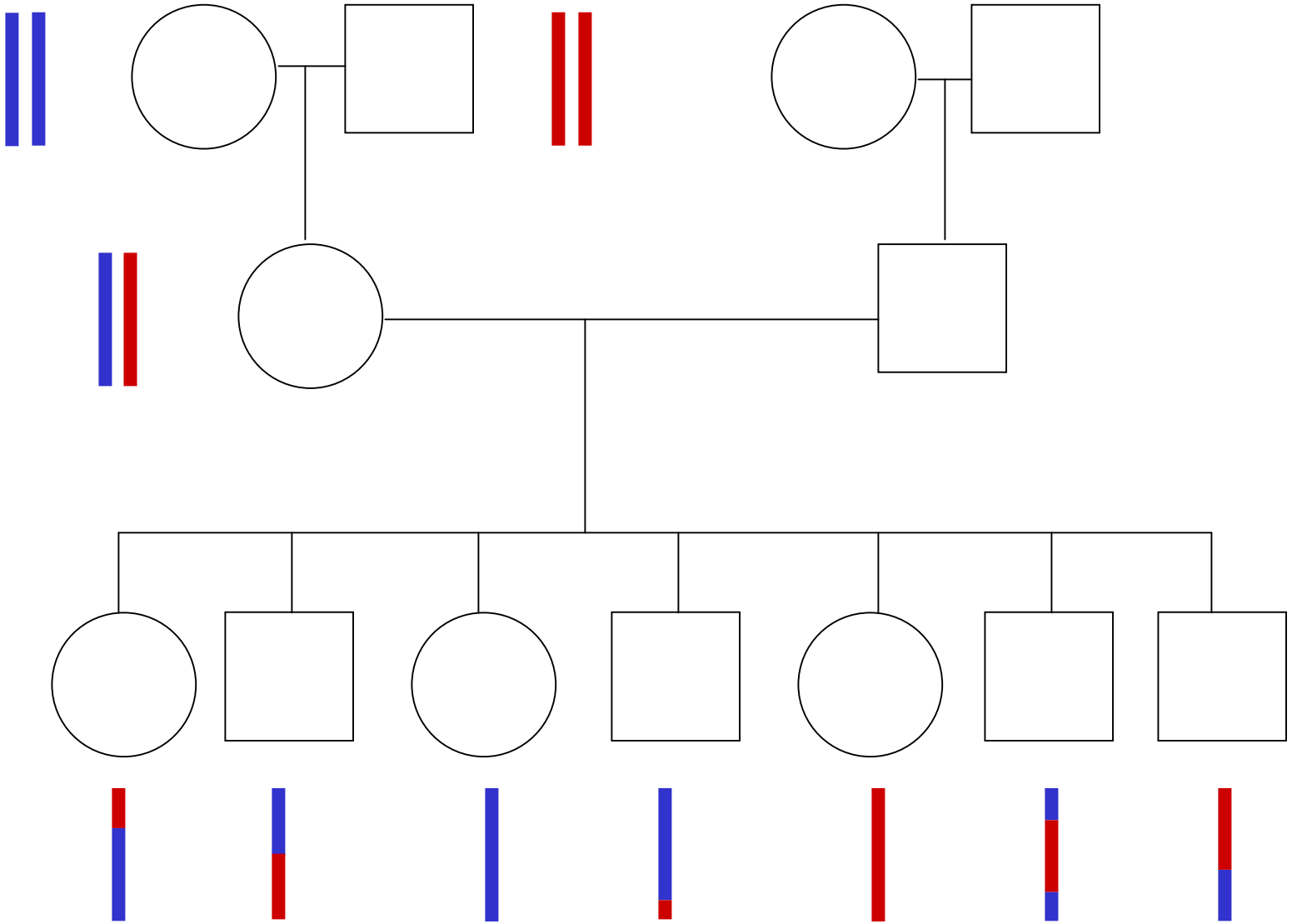
27.7 +/– 0.1 ppb?
27.7 +/– 1.0 ppb?
27.7 +/– 10.0 ppb?

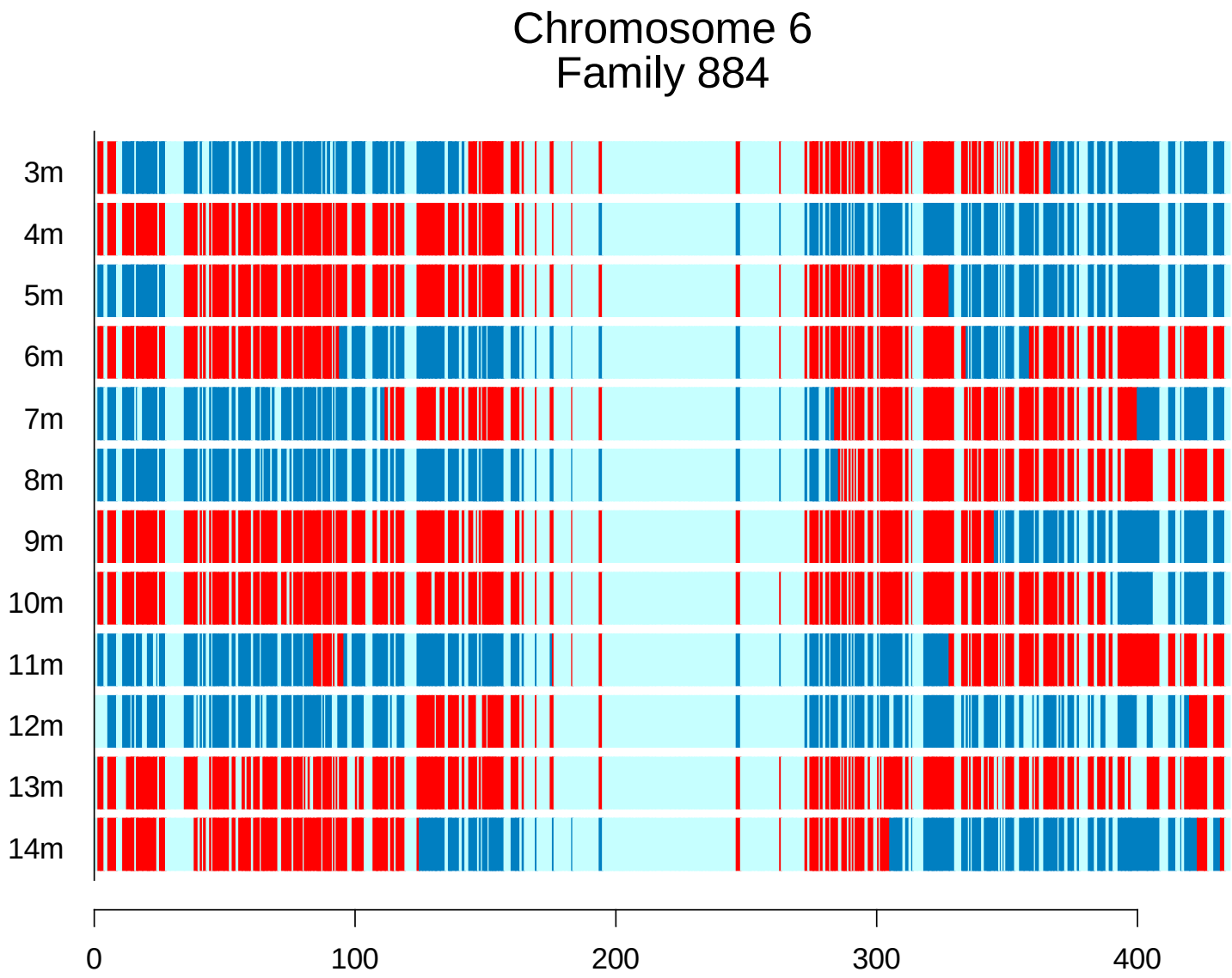
Dilution error and measurement error cause
variability in the fitted line which thus causes
variability in the estimated quinine concentration.

Exploratory data analysis

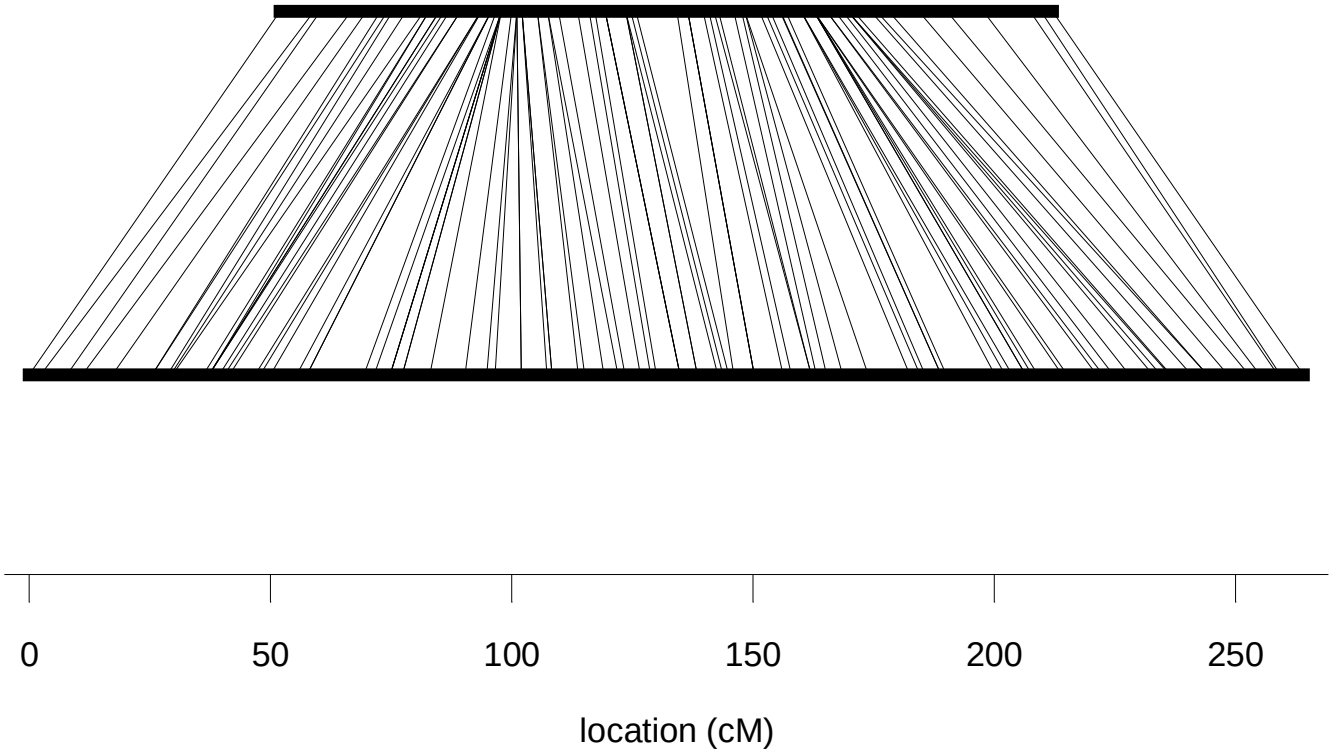
- detective work
  find new relationships in data

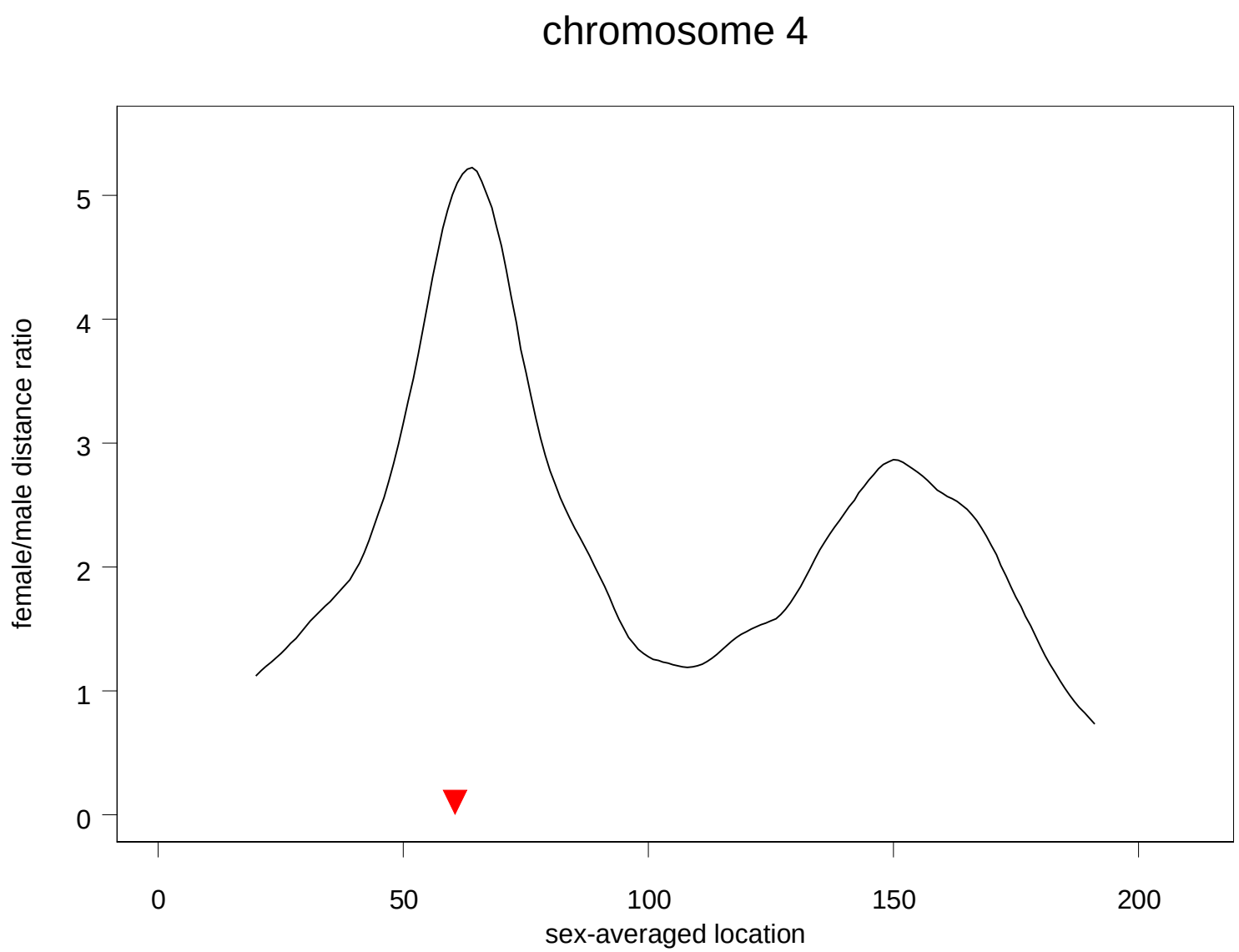
- visual design
  find better ways to display information




Chromosome 6
Family 884
3m
4m
5m
6m
7m
8m
9m
10m
11m
12m
13m
14m
0
100
200
300
400

Chromosome 4




chromosome 4
female/male distance ratio
0
1
2
3
4
5
0
50
100
150
200
sex-averaged location

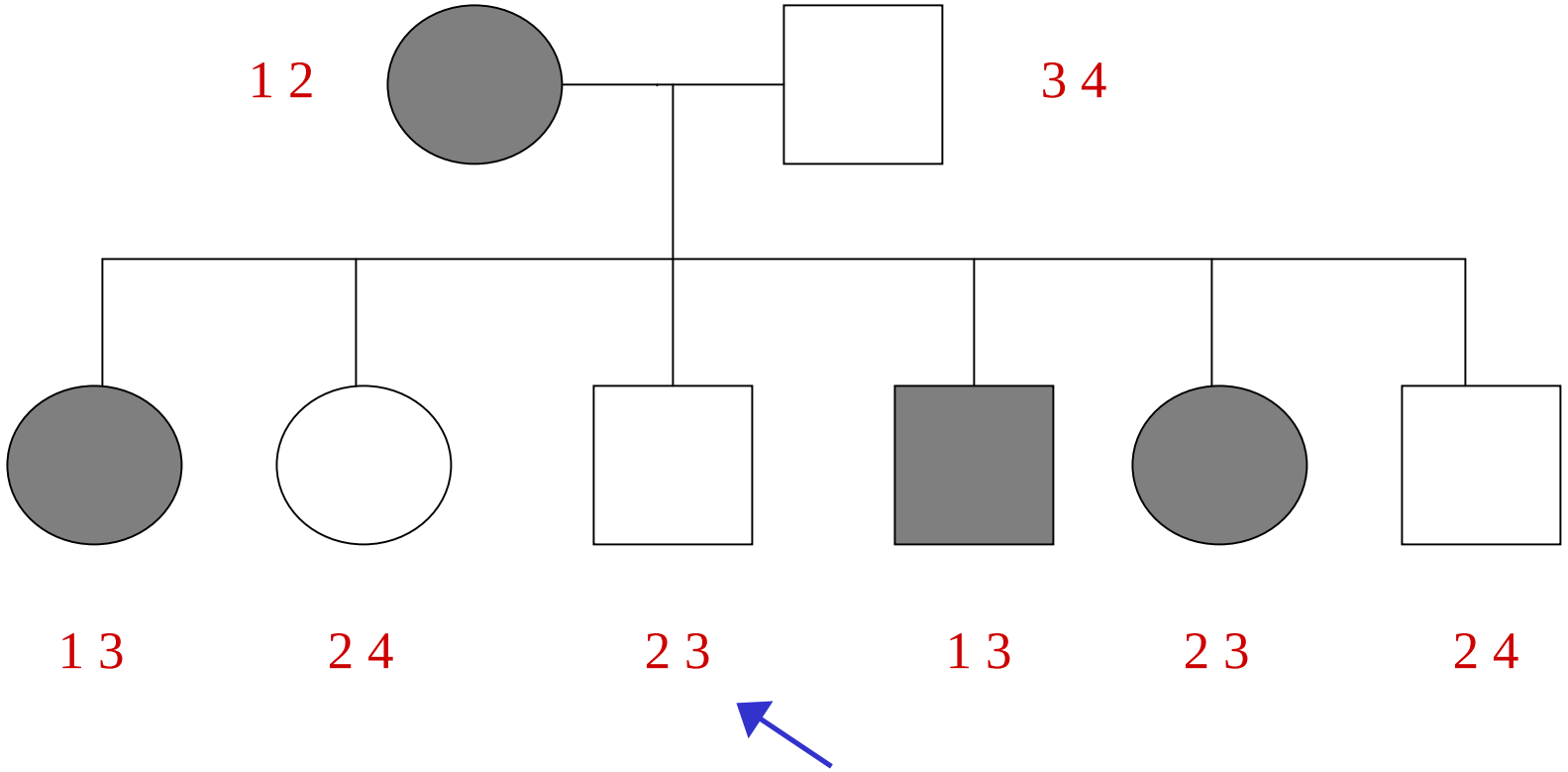
Statistical inference

- making general statements using limited data
- quantifying the uncertainty in such statements

Tools

mathematical models of random variation; probability

Example: linkage between gene for simple recessive trait and a genetic marker



θ = recombination fraction between marker and gene

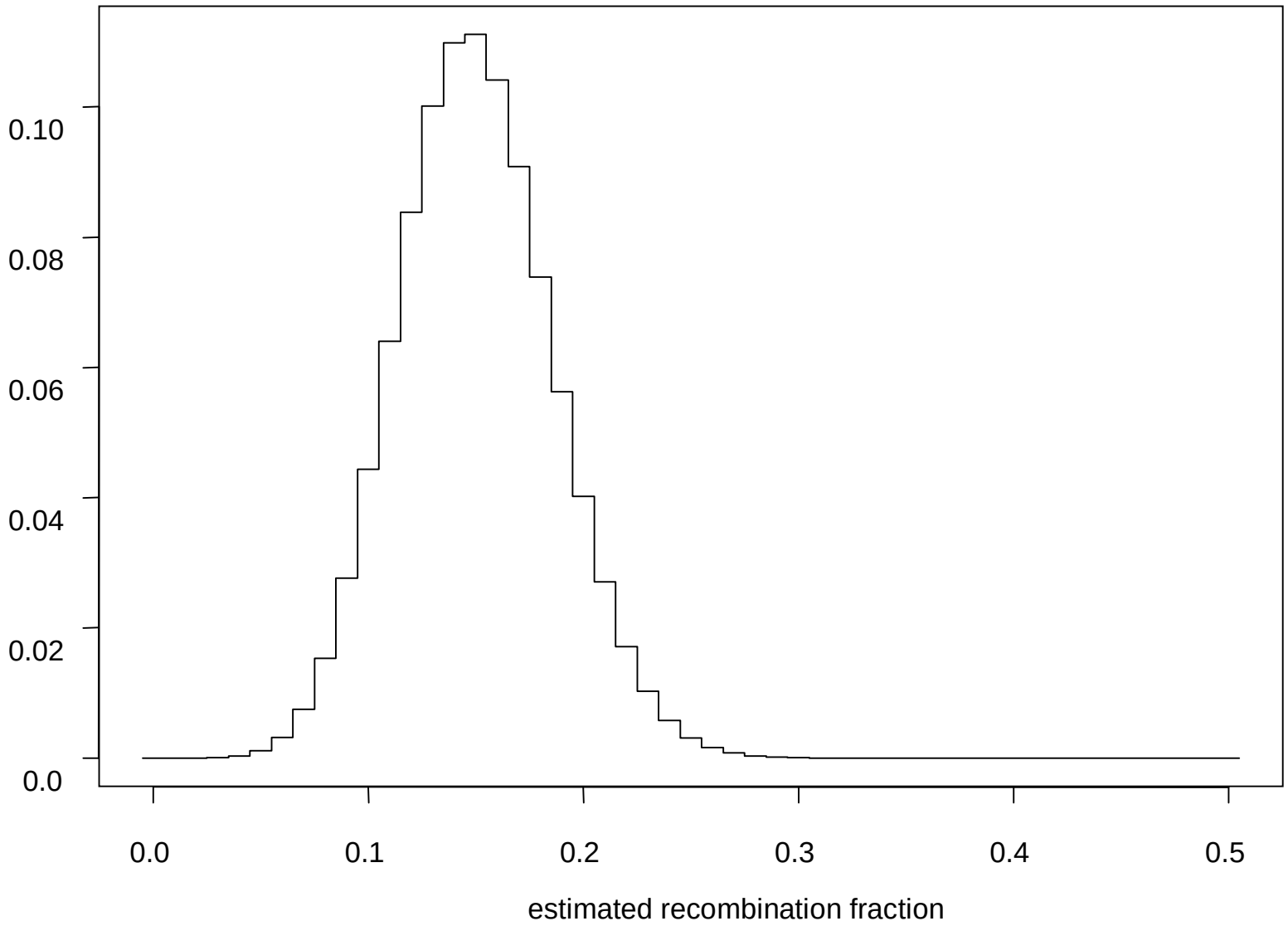
$L(\theta) = \Pr(\text{data}|\theta) \propto \theta^k (1-\theta)^{n-k} + \theta^{n-k}(1-\theta)^k$

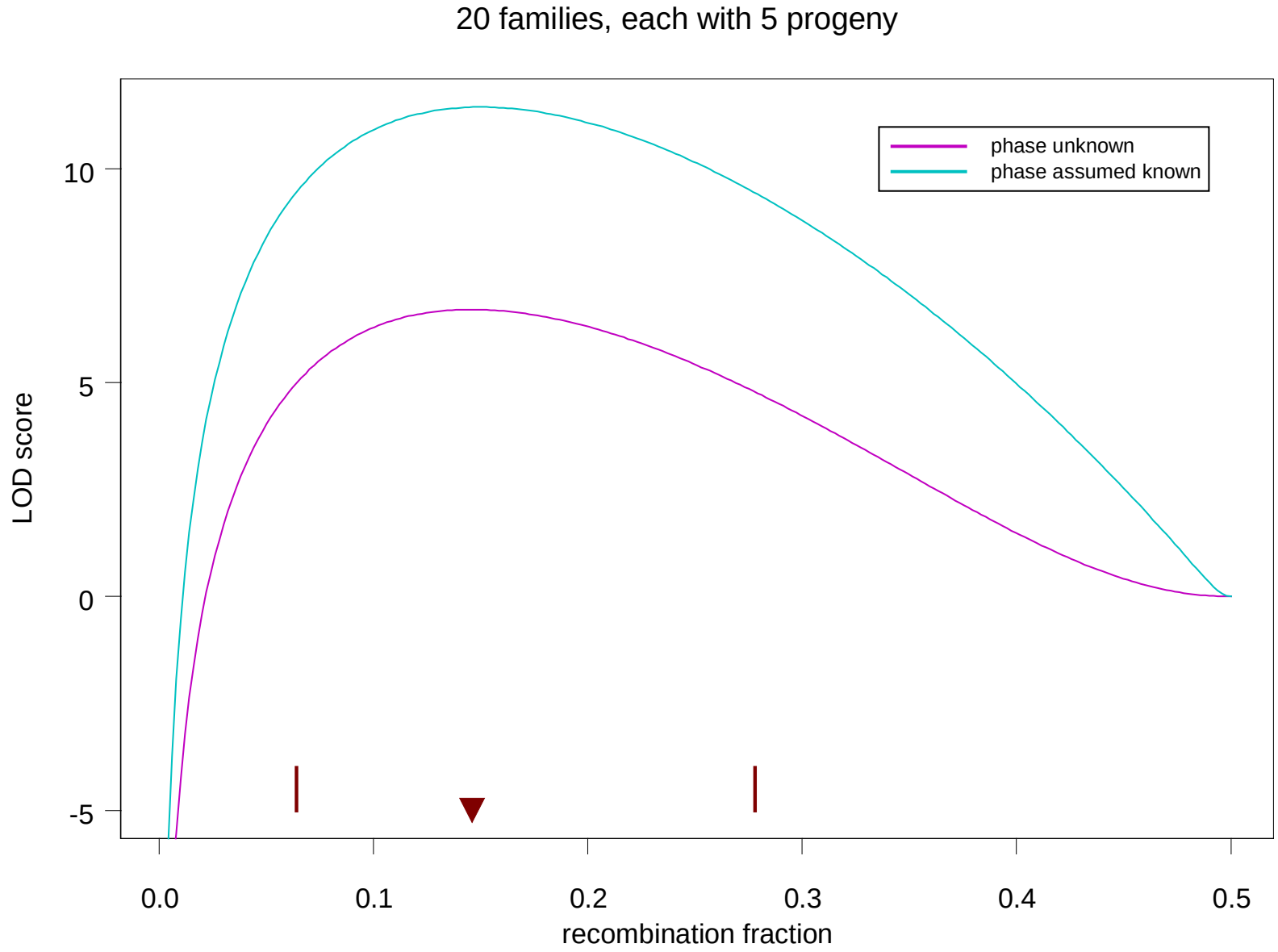
$\hat{\theta}$ = value of θ maximizing $L(\theta) \approx k / n$

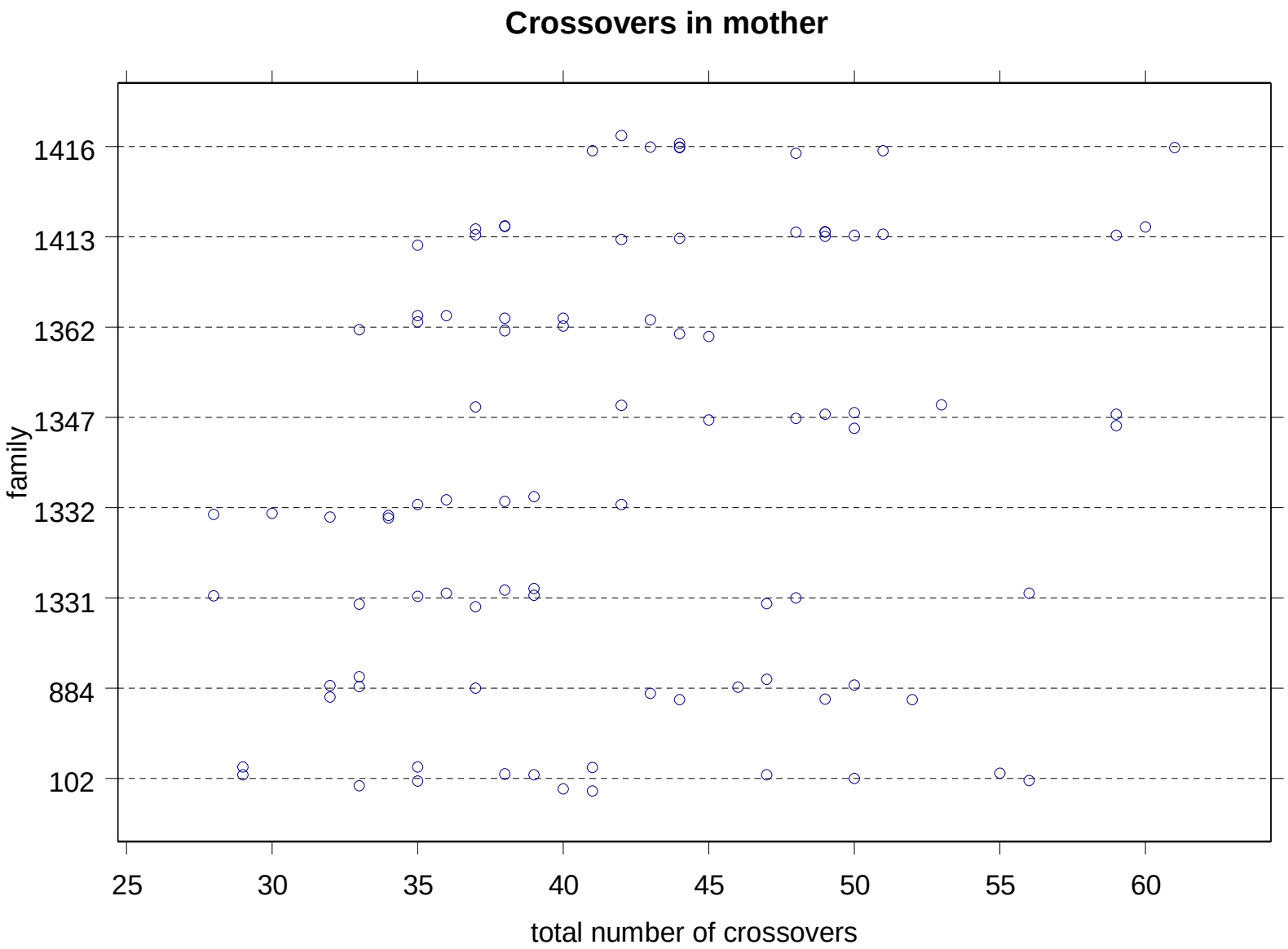
$\text{LOD}(\theta) = \log_{10}[L(\theta) / L(½)]$

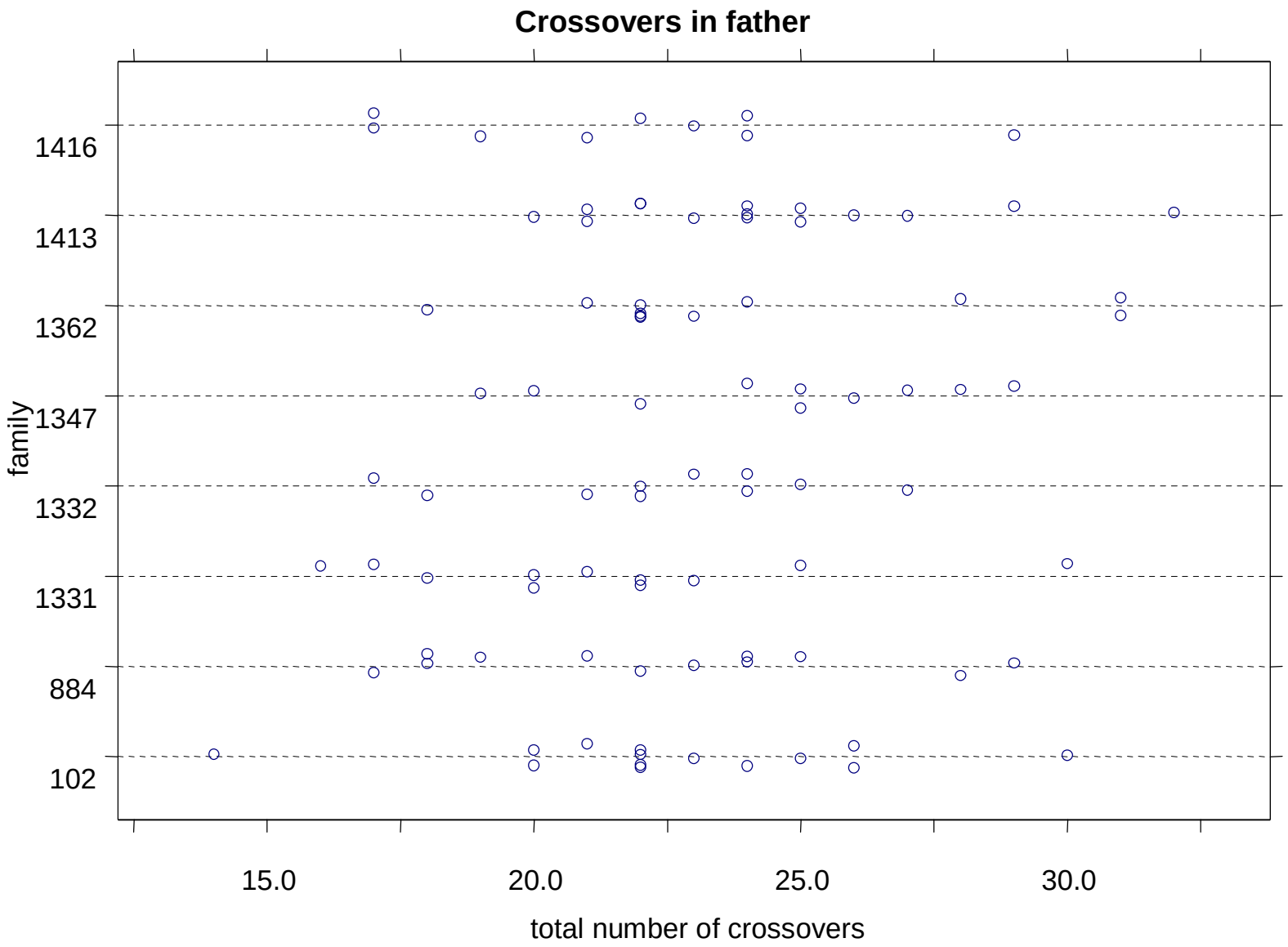
sampling distribution

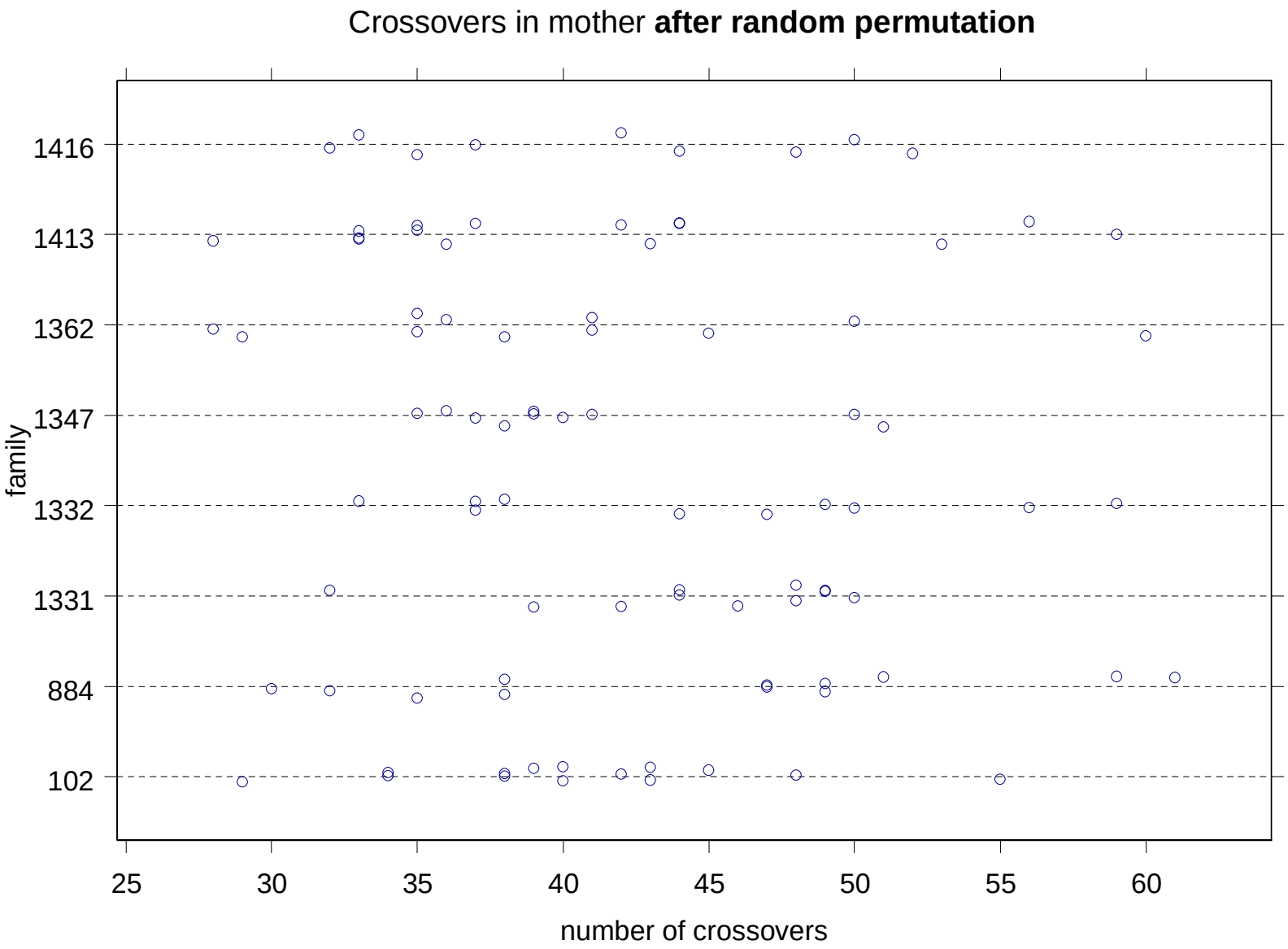
20 families with 5 progeny each; assume $\theta = 0.15$



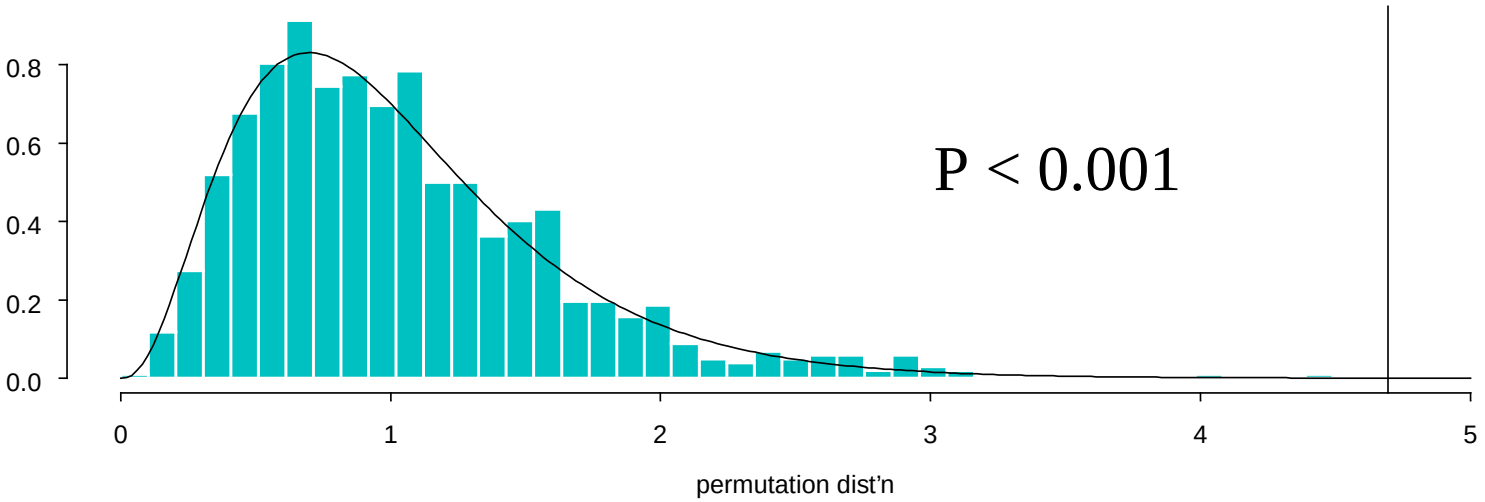

20 families, each with 5 progeny
phase unknown
phase assumed known
LOD score
10
5
0
-5
0.0
0.1
0.2
0.3
0.4
0.5
recombination fraction


Crossovers in mother
family
1416
1413
1362
1347
1332
1331
884
102
25
30
35
40
45
50
55
60
total number of crossovers


Crossovers in father
family
1416
1413
1362
1347
1332
1331
884
102
15.0
20.0
25.0
30.0
total number of crossovers

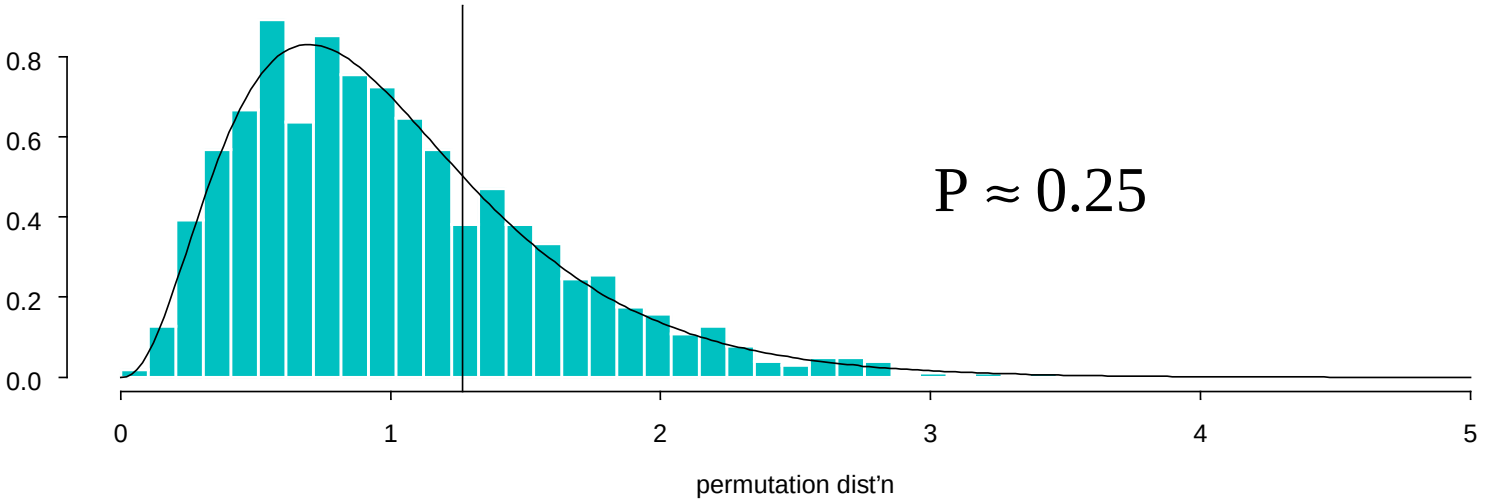

Crossovers in mother **after random permutation**
family
1416
1413
1362
1347
1332
1331
884
102
25
30
35
40
45
50
55
60
number of crossovers

mothers



$P < 0.001$

fathers



$P \approx 0.25$

Summary

- exploratory data analysis
  - find new ways to look at data
  - follow up strange observations

- statistical inference
  - probability model for data
  - sampling distribution of estimator
  - confidence interval, hypothesis test