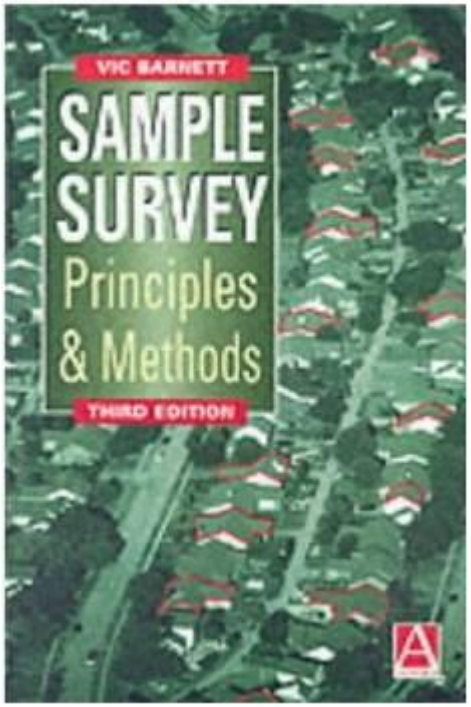
Sample survey



Barnett, Vic. 2002. Sample Survey: Principles & Methods (3rd ed.). New York: Oxford University Press

Inference for finite population

- Computation of population characteristics (census)
- Estimation of population characteristics (sample survey)

Census vs. sample survey

- Budget and time
- Coverage
- Accuracy
- Feasibility

Steps in sampling design

- What is the population?
- What are the parameters of interest?
- What is the sampling frame*?
- What size of the sample is needed?
- How much will it cost?

*the list of elements from which the sample is actually drawn

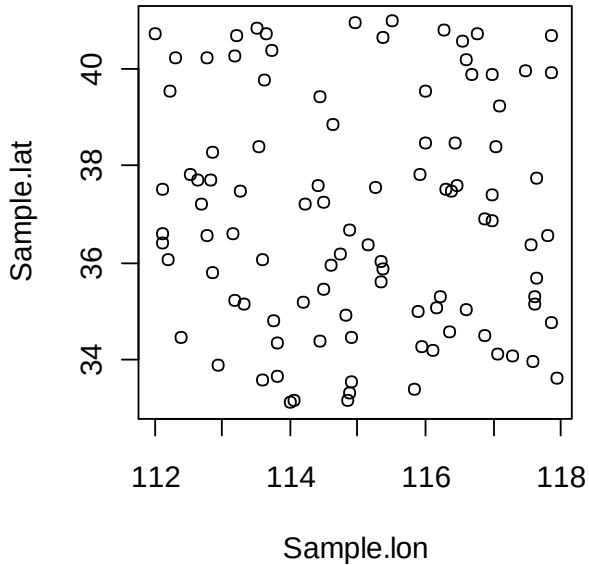
Fundamental Sampling Plans

- Simple Random Sampling (SRS)
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling
- Multistage Sampling

Fundamental Concept of Simple Random Sampling

- A population unit is randomly selected from the population until a set of sample of size “n” is achieved
- At each of the selection process, the remaining population units have an equal chance of being selected
- A set of samples occurs with an equal probability

```
# Simple Random Sampling for 100 locations
Sample.lat  <- runif(100,  min = 33,  max = 41)
Sample.lon <- runif(100, min = 112, max = 118)
plot (Sample.lon, Sample.lat)
```



Population Characteristics

- Population Total $Y = \sum_{i=1}^{N} y_i$
- Population Mean $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$
- Population Proportion $P = \frac{A}{N}$
- Ratio $R = \frac{Y}{X}$

Population characteristics and estimators under SRS

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$Y = \sum_{i=1}^{N} y_i \qquad \hat{Y} = N\overline{y}$$

$$P = \frac{A}{N} \qquad p = \frac{a}{n}$$

$$R = \frac{Y}{X} \qquad r = \frac{y}{x}$$

Quality of sampling

- Accuracy
  - Systematic variance
    - The variation in measures due to some known or unknown influences that “cause” the scores (results) to lean in one direction more than another
- Precision
  - Sampling error
    - The degree to which a given sample differs from the underlying population
    - Sampling error tends to be high with small sample sizes and will decrease as sample size increases

Properties of the estimator under SRS

$\bar{y}$ is an unbiased estimator for $\bar{Y}$ with the Variance

$$Var(\bar{y}) = (1 - f)\frac{S^2}{n};$$

where $f = \frac{n}{N}$ and $S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1}$

Proportion

- Basic Properties of the estimator

  Sample proportion $p = \frac{a}{n}$ is an unbiased estimator

  for the population proportion $P = \frac{A}{N}$ with the variance

  $$Var(p) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right) = \frac{PQ}{n}(1-f)$$

  where $Q = 1\text{-}P$

Sampling error and sample size

Sampling error e when estimating a proportion p with a sample of size n taken from an infinite population

$$\mathrm{var}(p) = \frac{p(1-p)}{n}(1-f)$$

$$\mathbf{e} = \sqrt{\frac{\mathbf{p(1-p)}}{\mathbf{n}}}$$

Confidence intervals

In a sample of 1,000 enterprises, 280 enterprises (28 percent) have been harassed by a predatory agency.

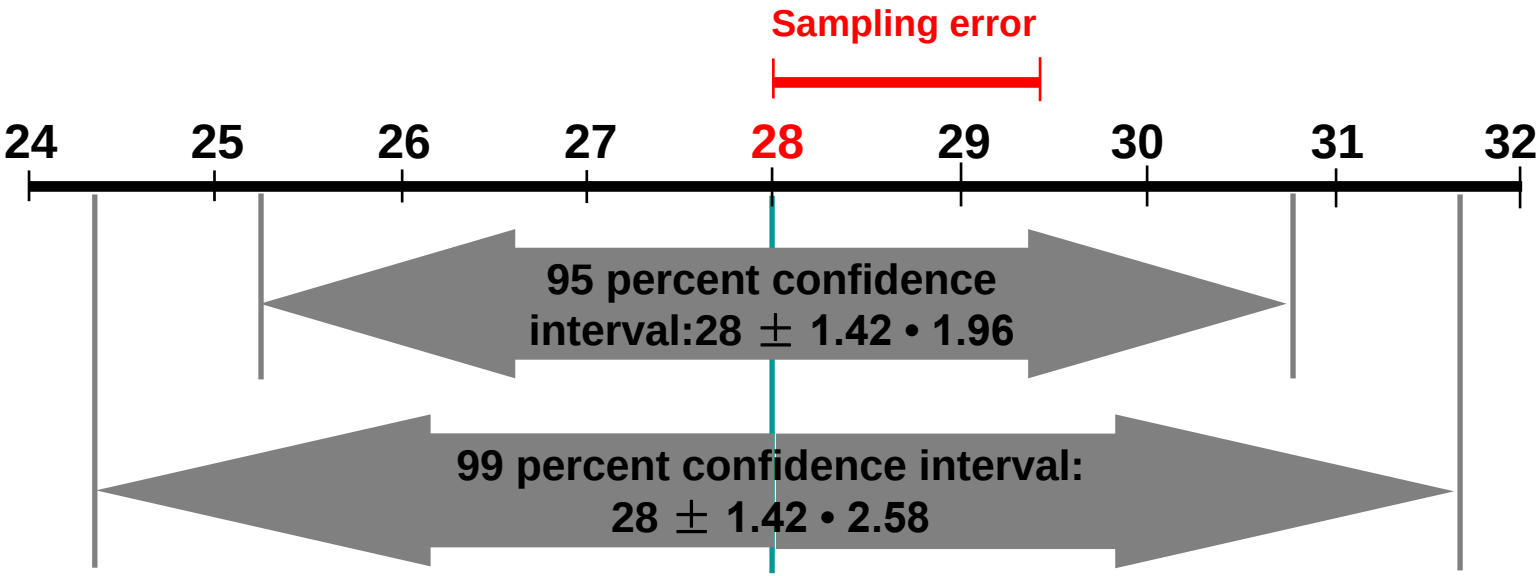
$$e = \sqrt{\frac{0.28 \times 0.72}{1{,}000}} = 0.0142$$

Sampling error is 1.42 percent.

Confidence intervals

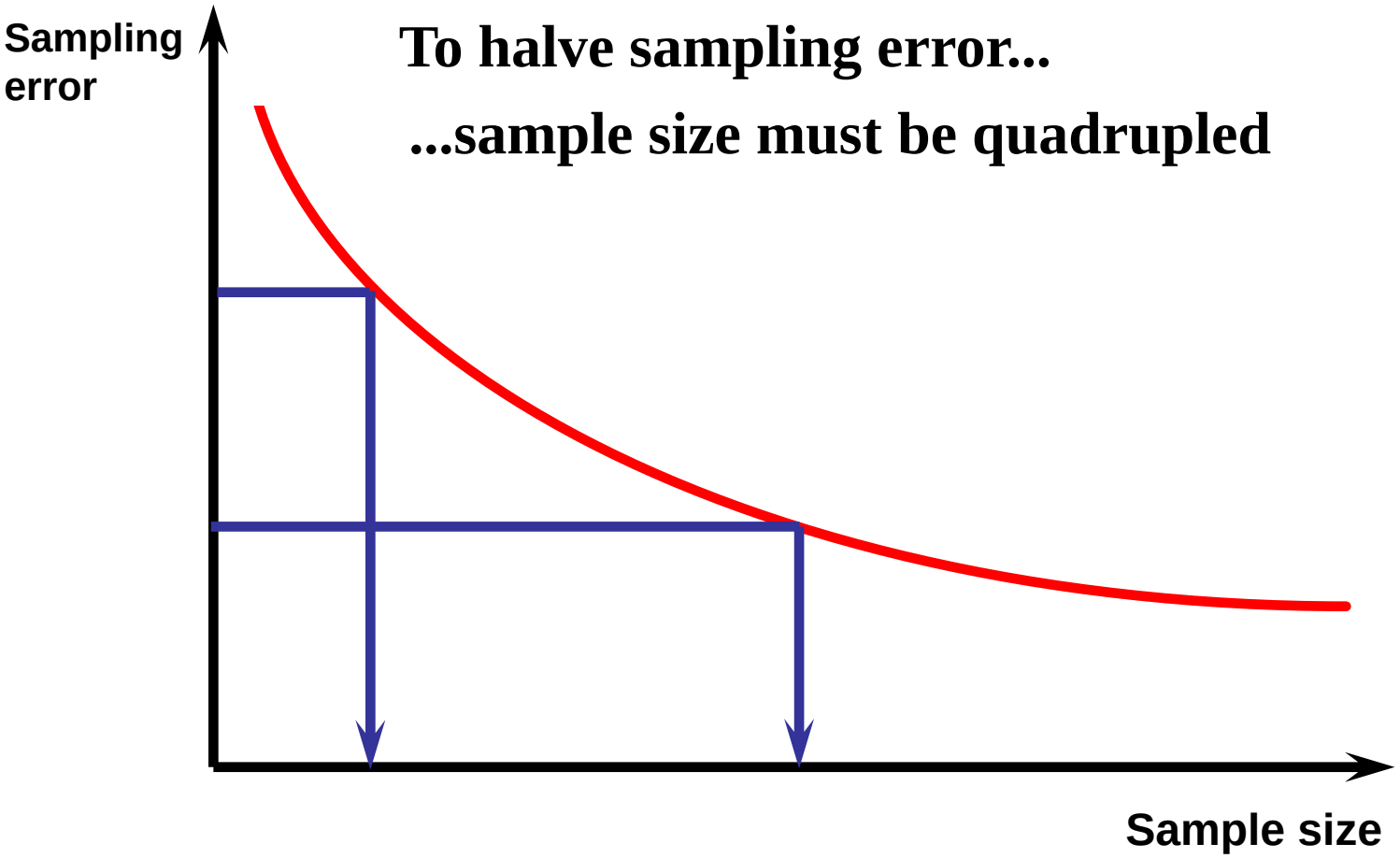
In a sample of 1,000 enterprises, 280 enterprises (28 percent) have been harassed by a predatory agency.

Sampling error is 1.42 percent.



Sampling error and sample size

$$e = \sqrt{\frac{p(1-p)}{n}}$$



Sample size and population size

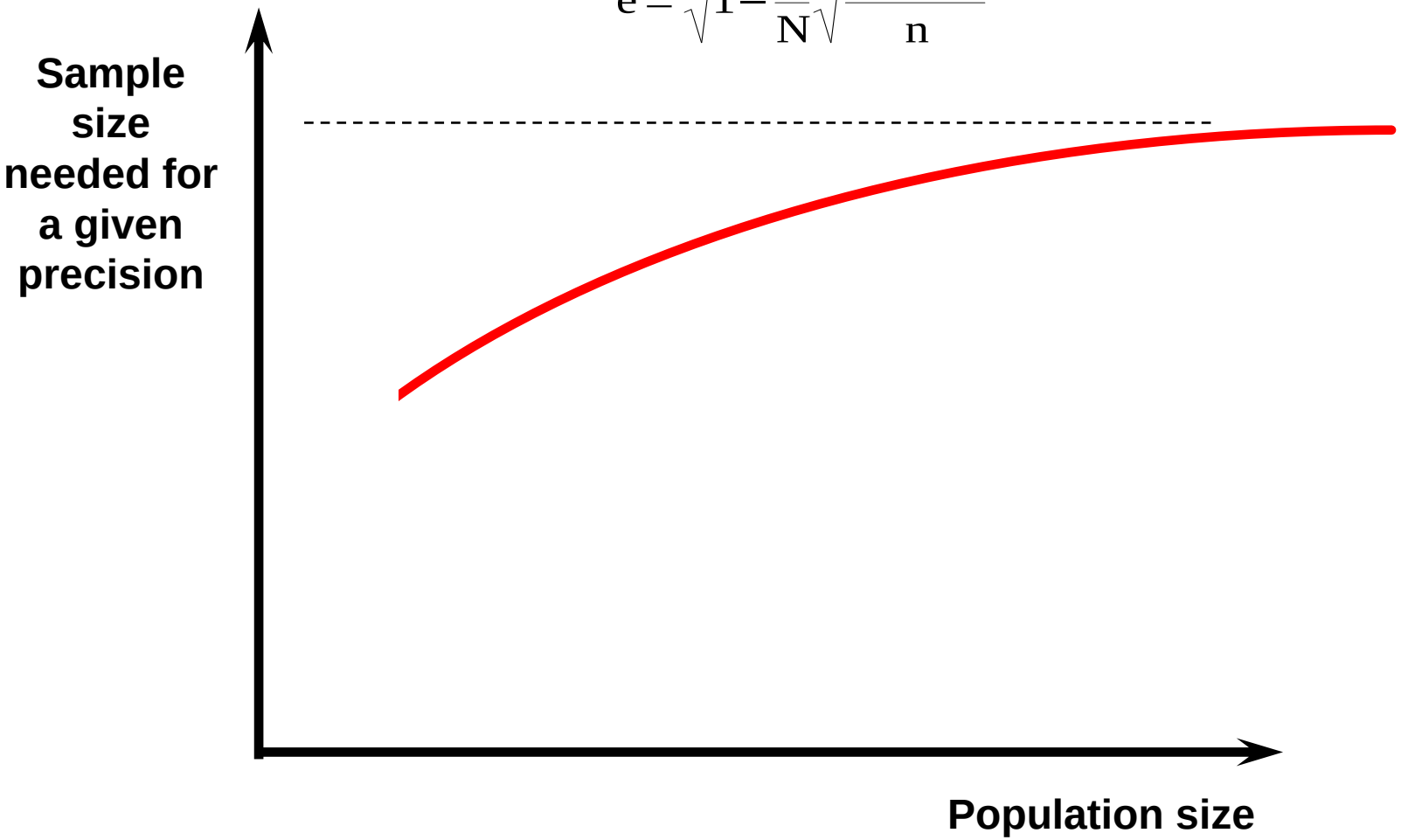
Sampling error e when estimating a proportion p with a sample of size n taken from a population of size N

$$e = \sqrt{1 - \frac{n}{N}} \sqrt{\frac{p(1-p)}{n}}$$

finite population correction

Sample size and population size

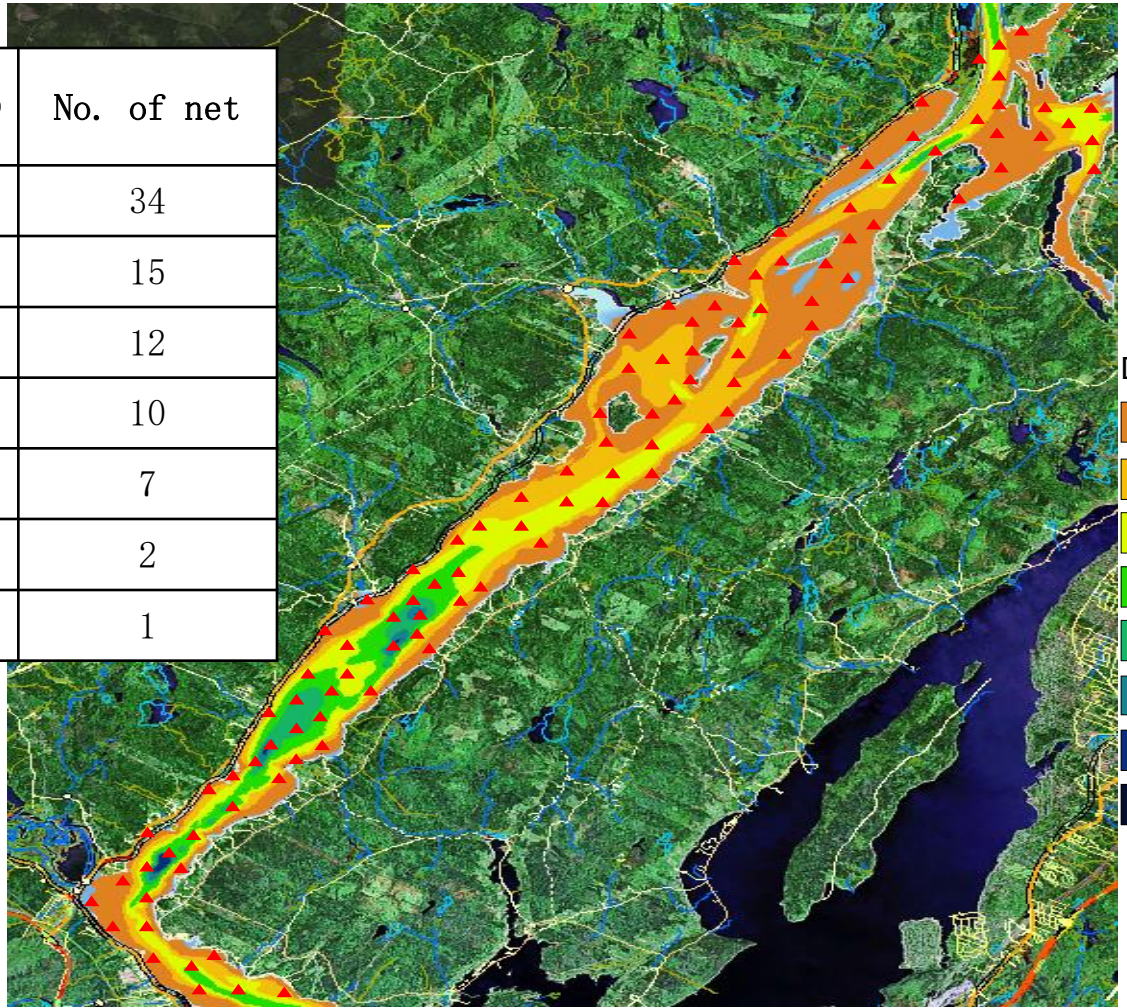
$$e = \sqrt{1 - \frac{n}{N}} \sqrt{\frac{p(1-p)}{n}}$$



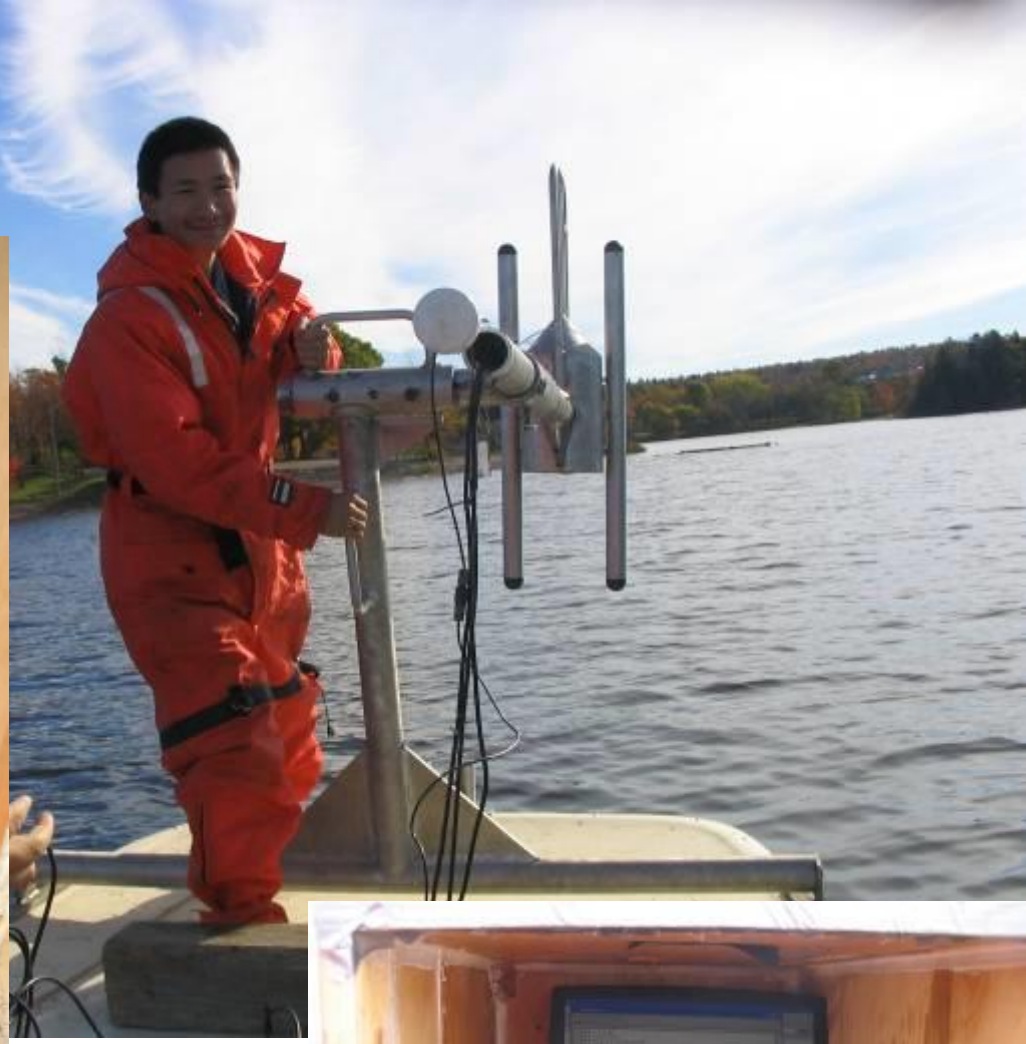
Stratified sampling

Sample the shortnose sturgeon using gillnet following a stratified sample design

Strata	Depth (m)	Area (km)	No. of net
Stratum1	0-5	68	34
Stratum2	5-10	31	15
Stratum3	10-15	25	12
Stratum4	15-20	20	10
Stratum5	20-25	15	7
Stratum6	25-30	5	2
Stratum7	>30	2	1

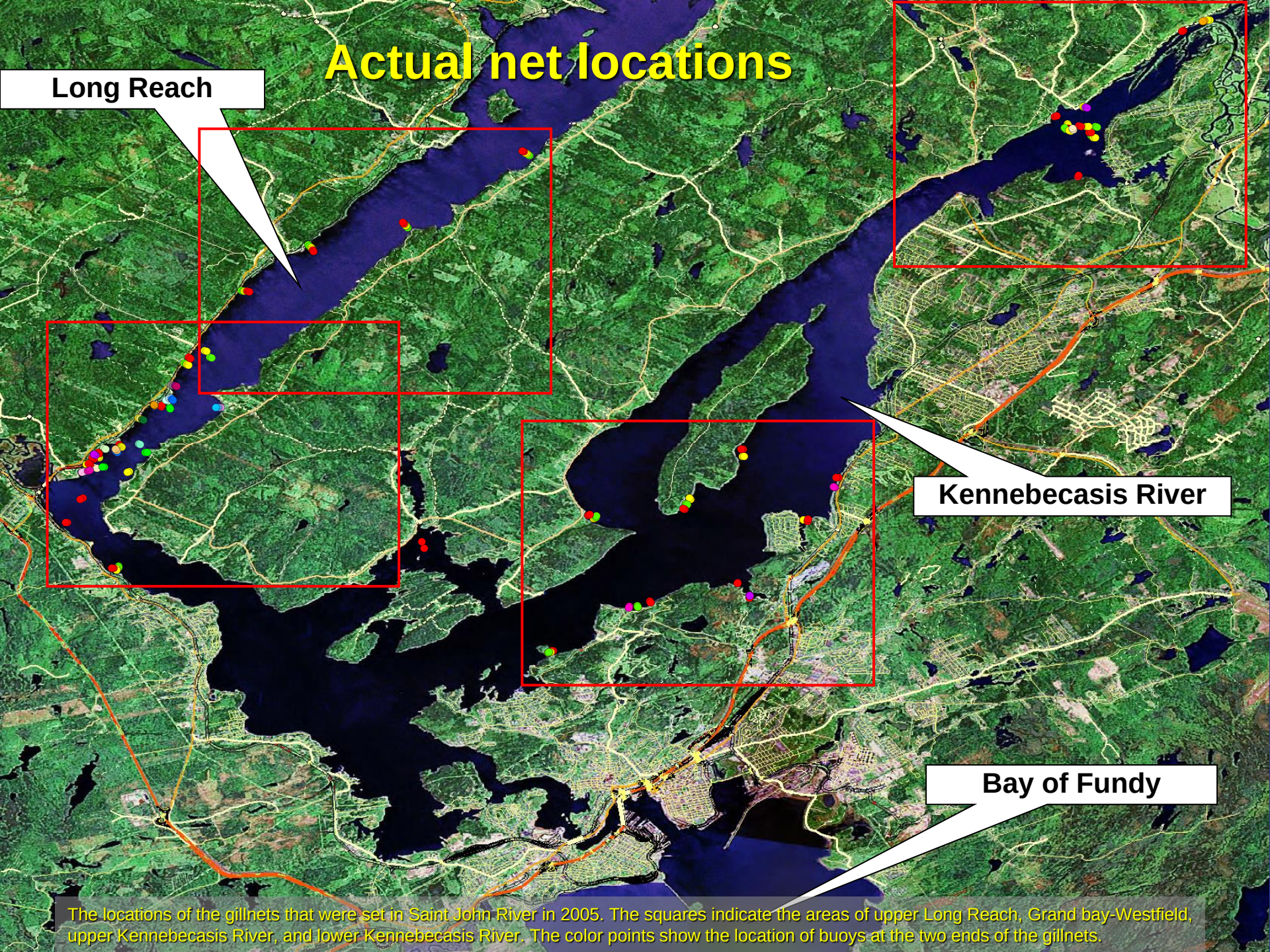


Sidescan sonar recording river bottom substrate and depth

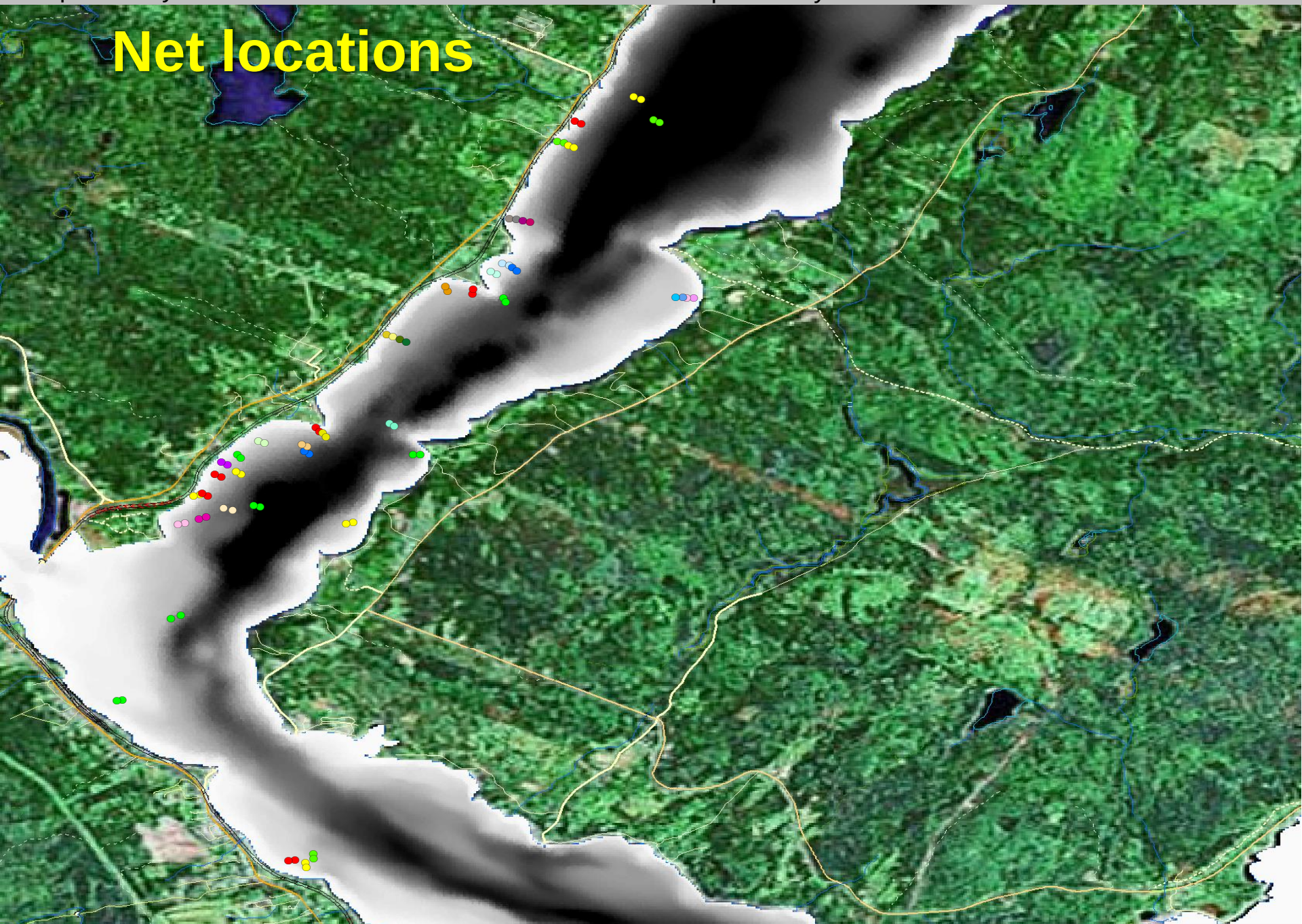


Detect depth

What we got from deep water



The locations of the gillnets that were set in Saint John River in 2005. The squares indicate the areas of upper Long Reach, Grand bay-Westfield, upper Kennebecasis River, and lower Kennebecasis River. The color points show the location of buoys at the two ends of the gillnets.


Net locations

Advantages of stratified sampling

1. Ensures that each strata (subpopulation) is well weighted

2. Can result in estimates with smaller standard errors if sampling is well allocated

3. Small Stratum will not be missed

Disadvantages of stratified sampling

1. More complicated than SRS

2. Need to identify strata ahead of time.

   Hence, more information needed prior to sampling than than for SRS

Systematic Sampling

A type of probability sampling in which every kth member of the population is selected

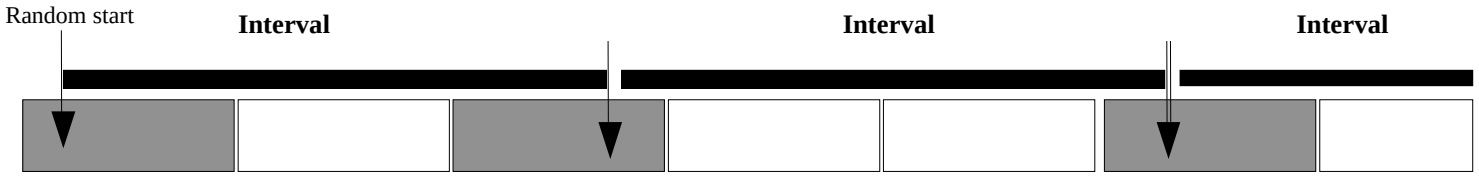
k=N/n

N = size of the population

n = sample size

Systematic Sampling

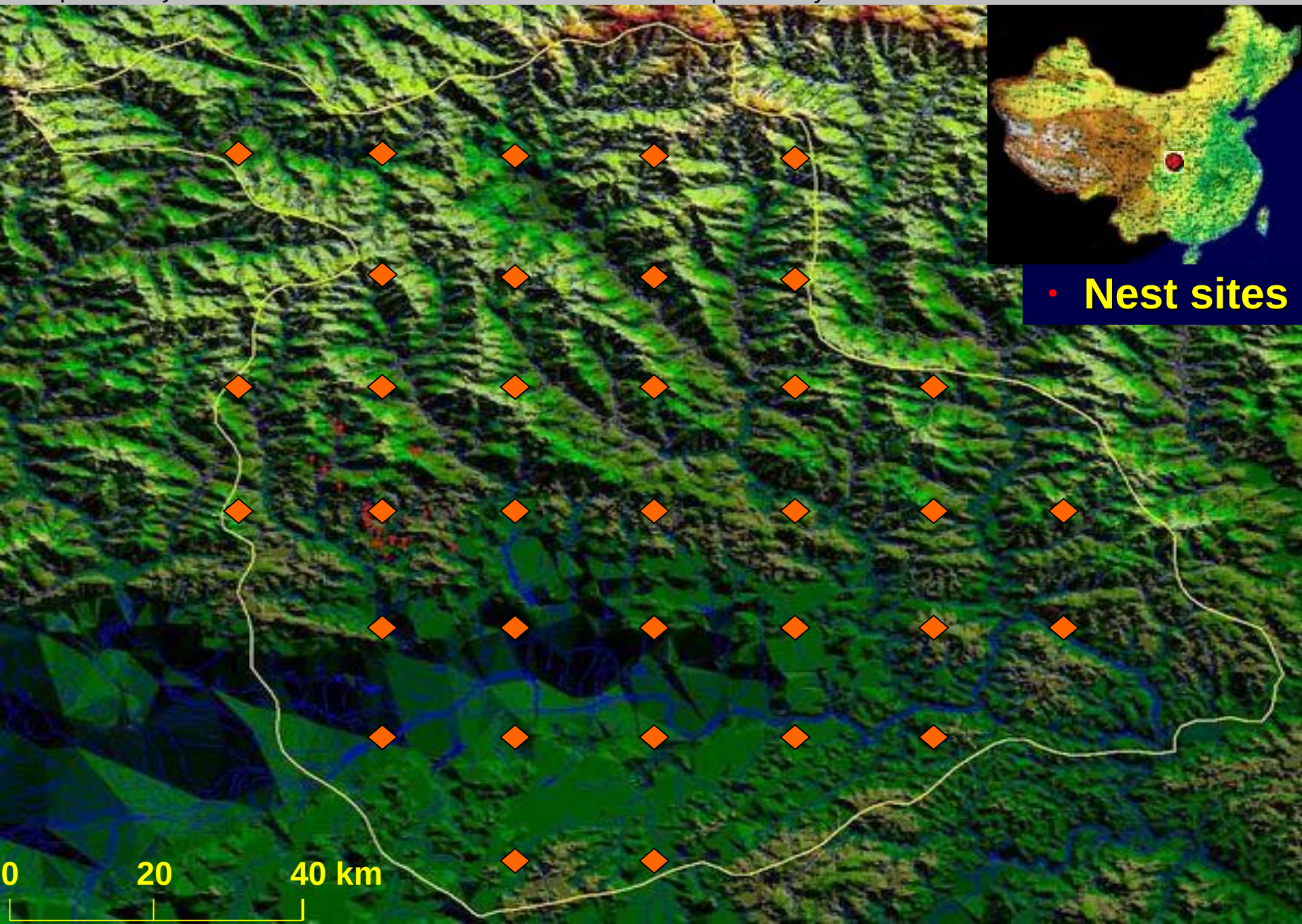
Illustration of systematic sampling procedure.



EQUAL PROBALITY SELECTION

PPS (probability proportional to size**) SELECTION**

Selected unit **Not selected**


Nest sites
0
20
40 km

Systematic Sampling

- Advantages
  - Simplicity. It allows the researcher to add a degree of system or process into the random selection of subjects.
  - The assurance that the population will be evenly sampled. There exists a chance in simple random sampling that allows a clustered selection of subjects. This is systematically eliminated in systematic sampling
- Disadvantage
  - The process of selection can interact with a hidden periodic trait within the population. If the sampling technique coincides with the periodicity of the trait, the sampling technique will no longer be random and representativeness of the sample is compromised.

Cluster Sampling

- The sampling unit contains more than one population element.
- For simple cluster sampling, each cluster contain the same number of elements; clusters are chosen randomly; all selected elements are included in the sample.

Cluster Sampling

- Suppose there are A clusters in the population; a clusters are selected.
- Each cluster contains B elements.
- Thus, the sample size is: n=aB
- Population size is N=AB

Cluster Sampling

- Sample mean is also the mean of the a cluster means.

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{1}{aB}\sum_{\alpha=1}^{a}\sum_{\beta=1}^{B} y_{\alpha\beta} = \frac{1}{a}\sum_{\alpha=1}^{a}\left(\frac{1}{B}\sum_{\beta=1}^{B} y_{\alpha\beta}\right)$$

$$= \frac{1}{a}\sum_{\alpha=1}^{a} \bar{y}_{\alpha}$$

Cluster Sampling

In terms of variance of the estimator, the situation is exactly the same as in SRS.

Property: An unbiased estimator of sample variance is:

$$Var(\bar{y}) = Var(\frac{1}{a}\sum_{\alpha=1}^{a}\bar{y}_\alpha)$$

$$= \frac{1-f}{a}\frac{1}{A-1}\sum_{\alpha=1}^{A}(\bar{y}_\alpha - \bar{y})^2$$

$$= \frac{1-f}{a}\frac{S_a^2}{B}$$

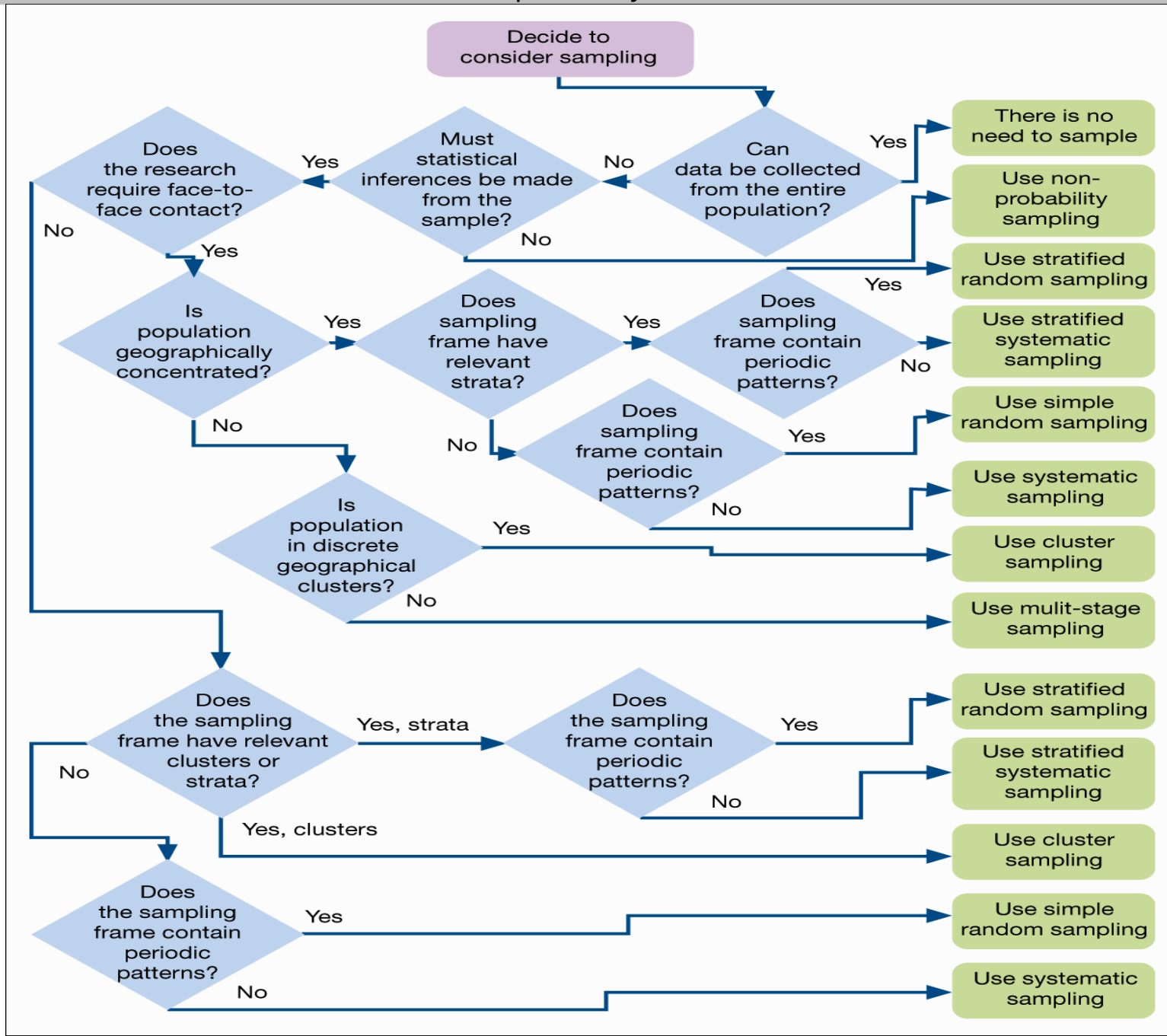
Key points

- This simple case of cluster sampling is very similar to srs of elements.
- The precision of the estimator depends on between cluster variance only. Thus, when selecting clusters, we want to **“minimize” between variance, or equivalently, “maximize” within variance**.
- Unfortunately in many cases, clusters are naturally formed. For example, county, classes, etc.

Multistage Sampling

- Stage 1
  - randomly sample clusters (or apply other sampling methods)
- Stage 2
  - randomly sample individuals from the cluster selected

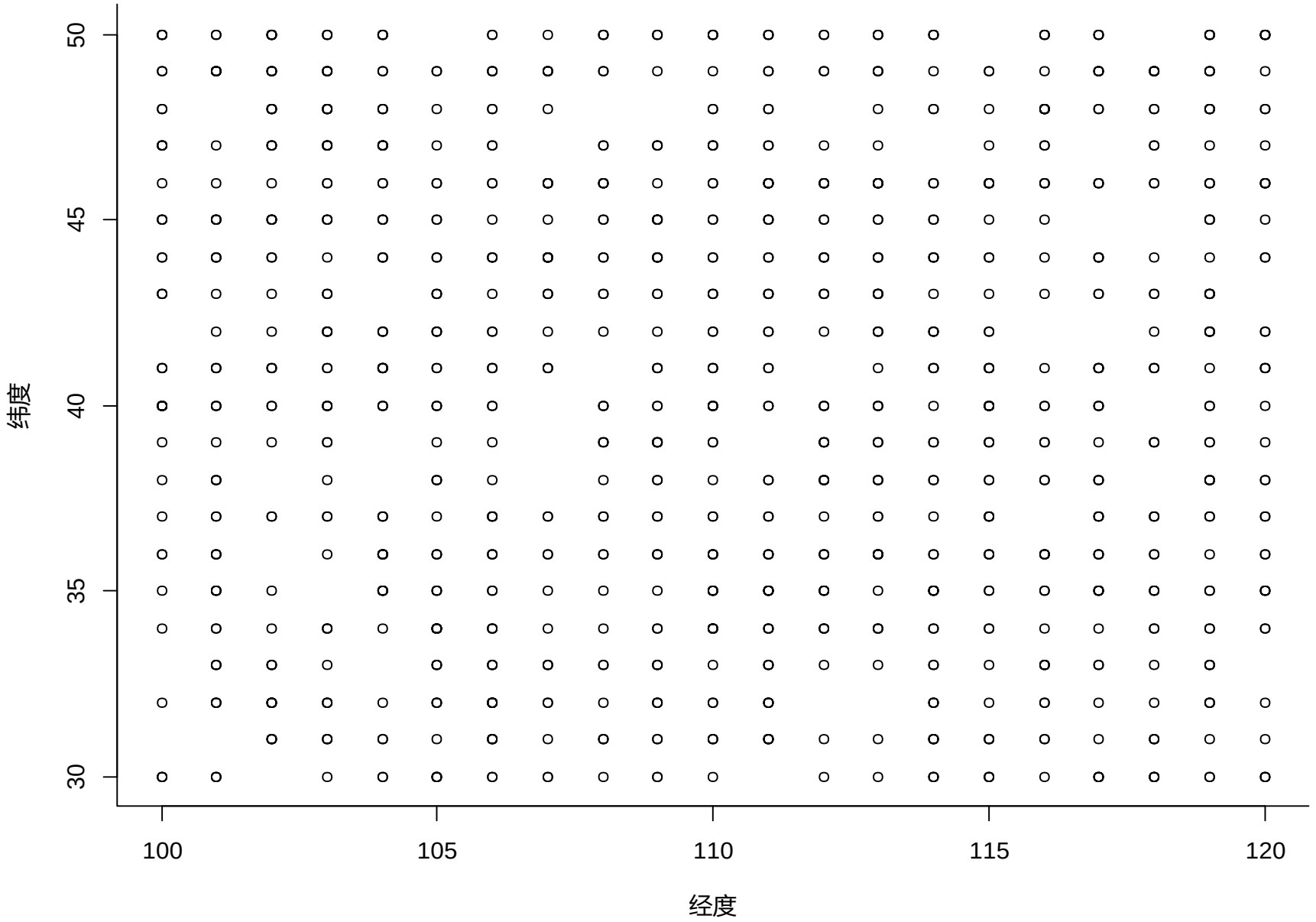
Overview of probability sampling



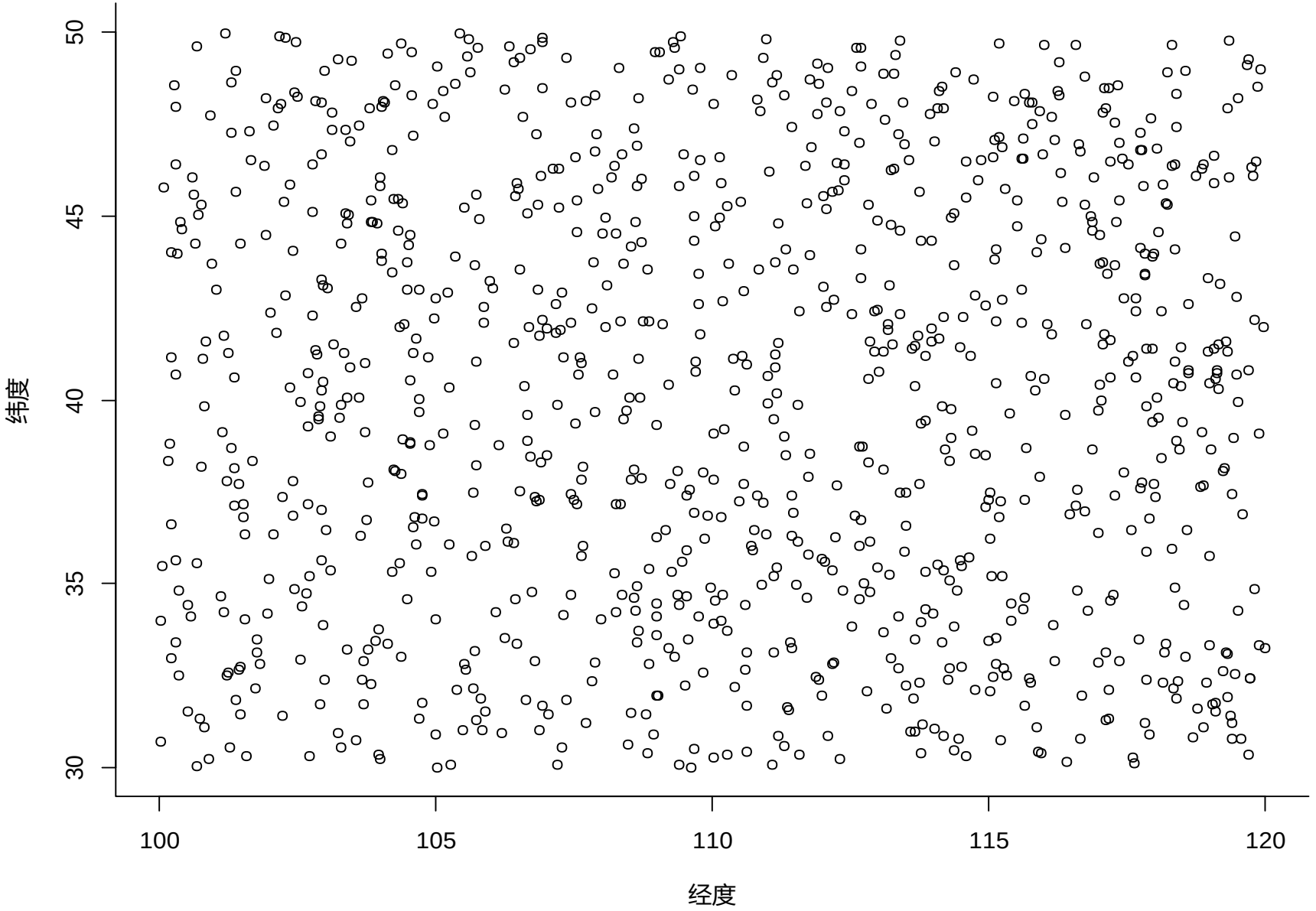
Case study: wildlife survey

- Uniform distribution (based on binomial distribution)
- Random distribution (based on Poisson distribution)
- Cluster distribution (based on negative binomial distribution)

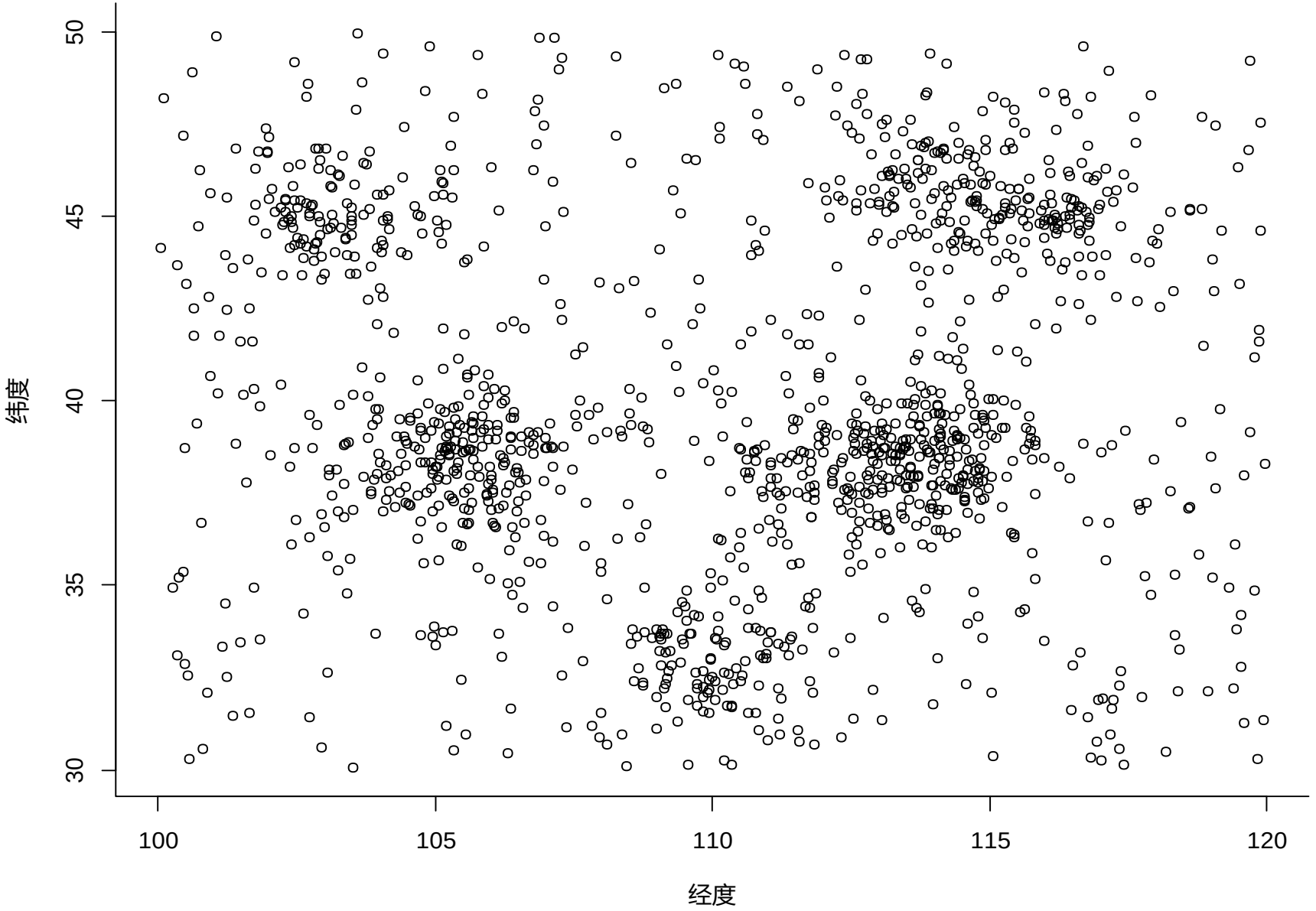
Uniform distribution



Random distribution

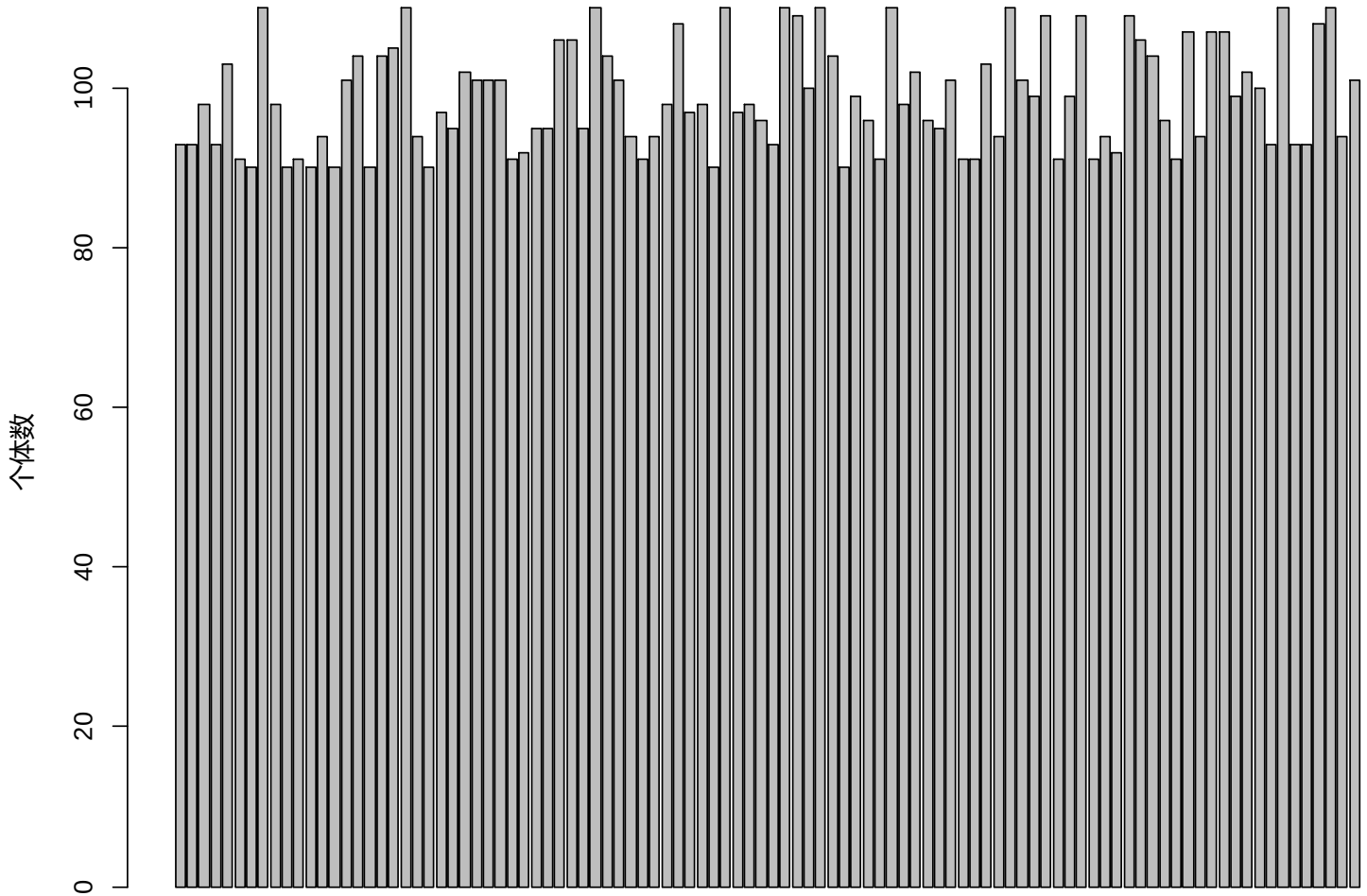


Cluster distribution

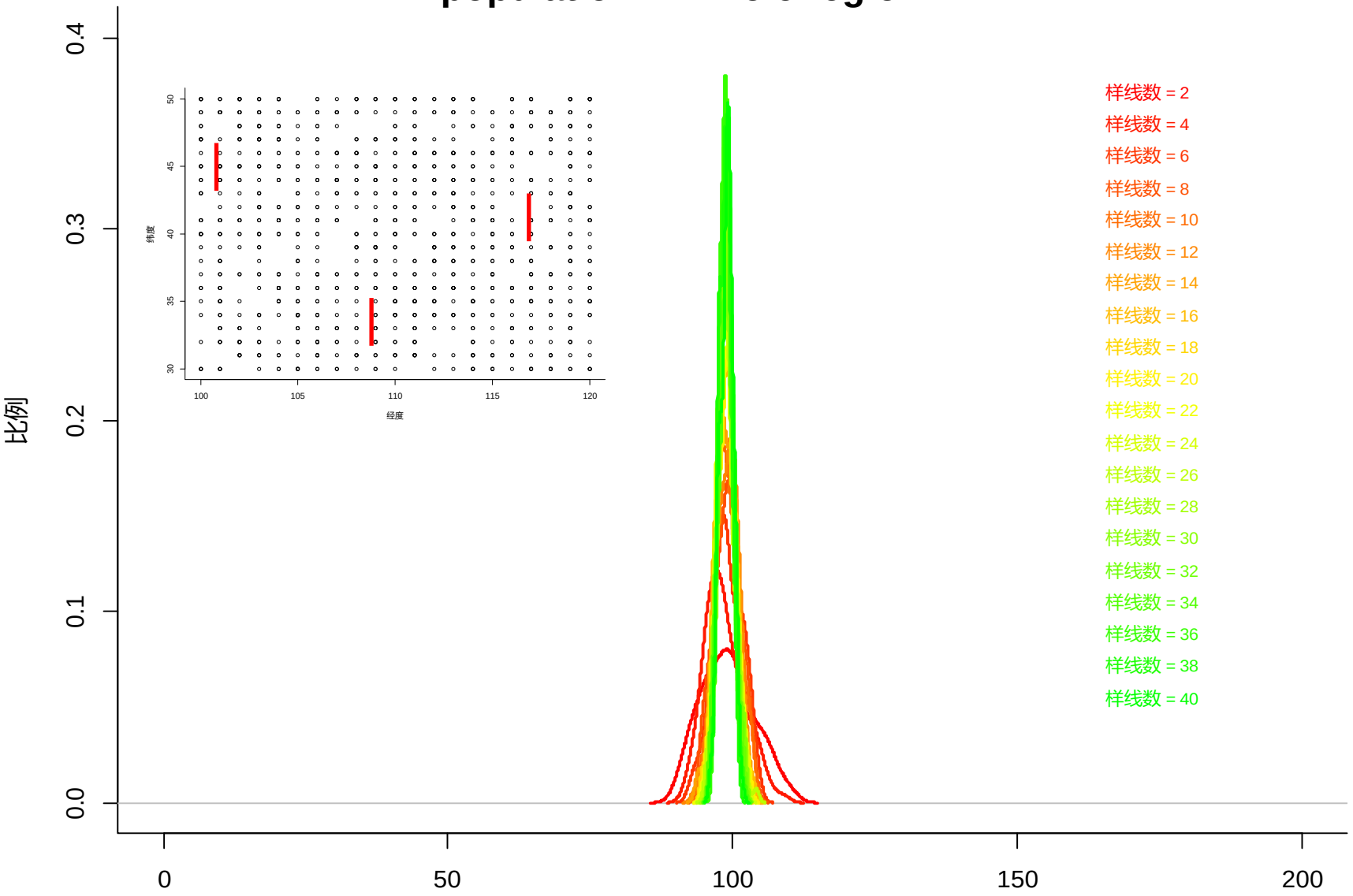


One question:
how many samples (survey routes) needed for estimating population size

Uniform distribution

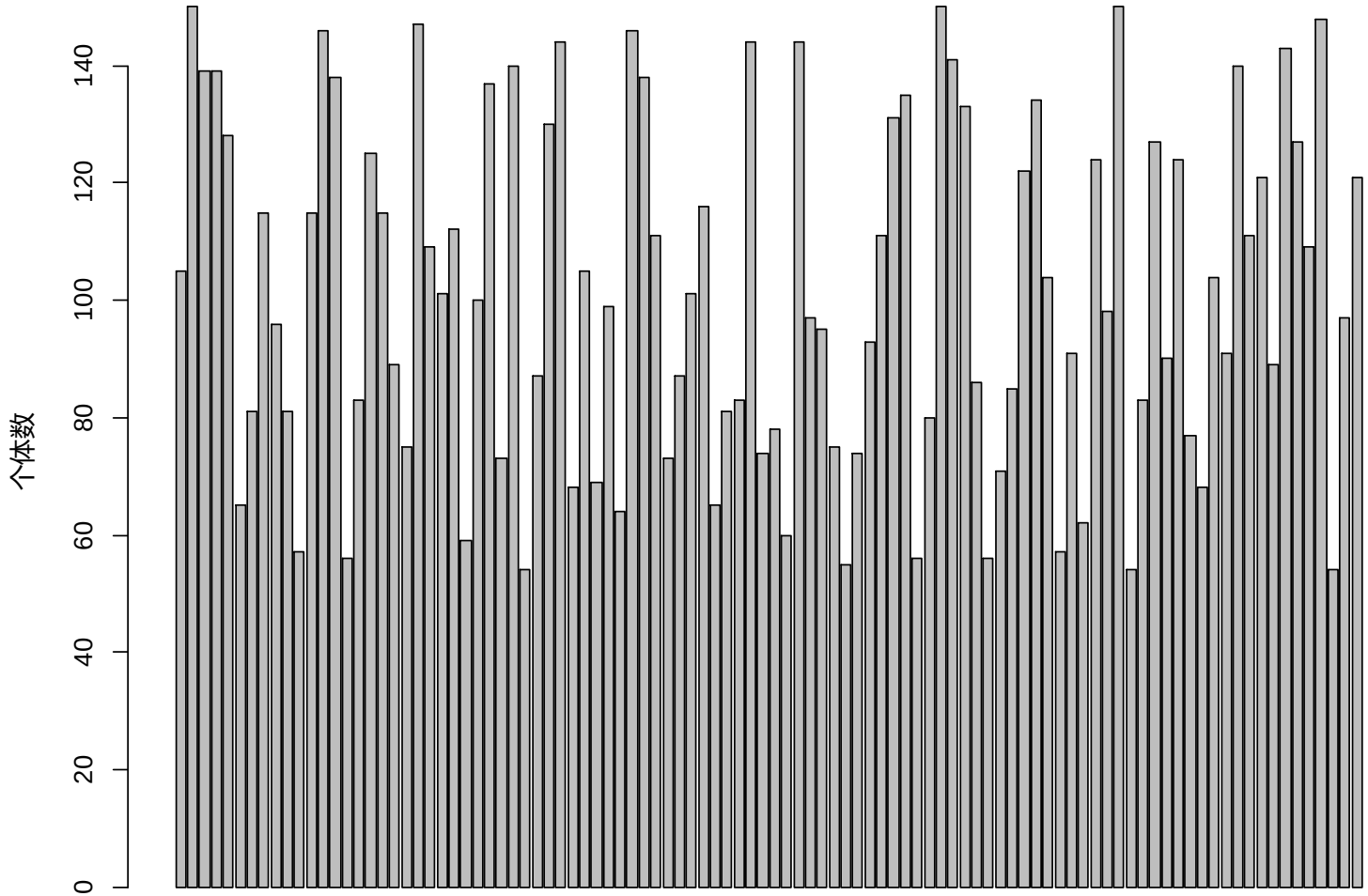


For uniform distribution, three survey routes are enough for estimating the population in whole region

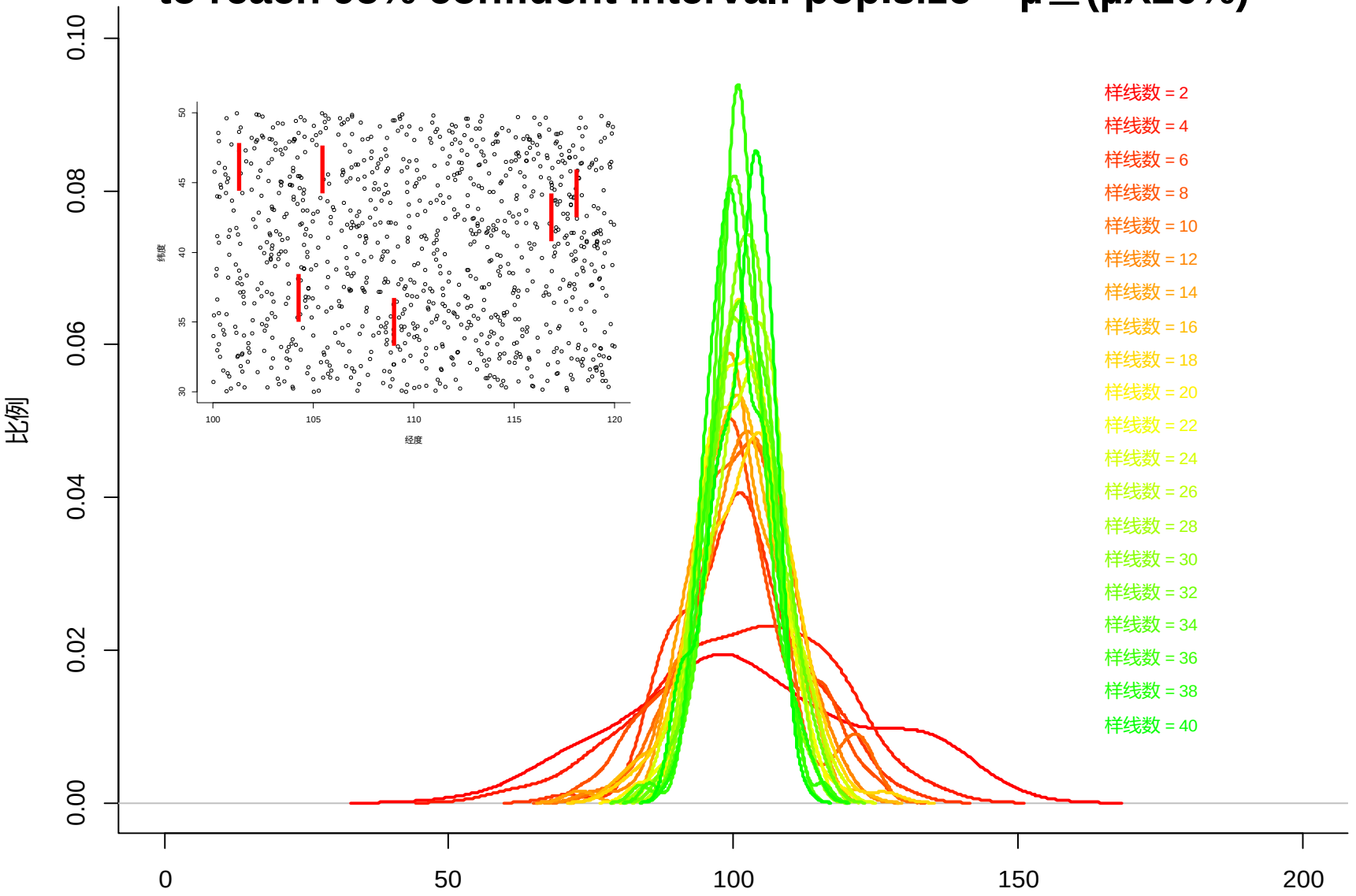


动物数量（根据不同样线数量估计动物的密度和数量）

Random distribution (variance = 10μ)

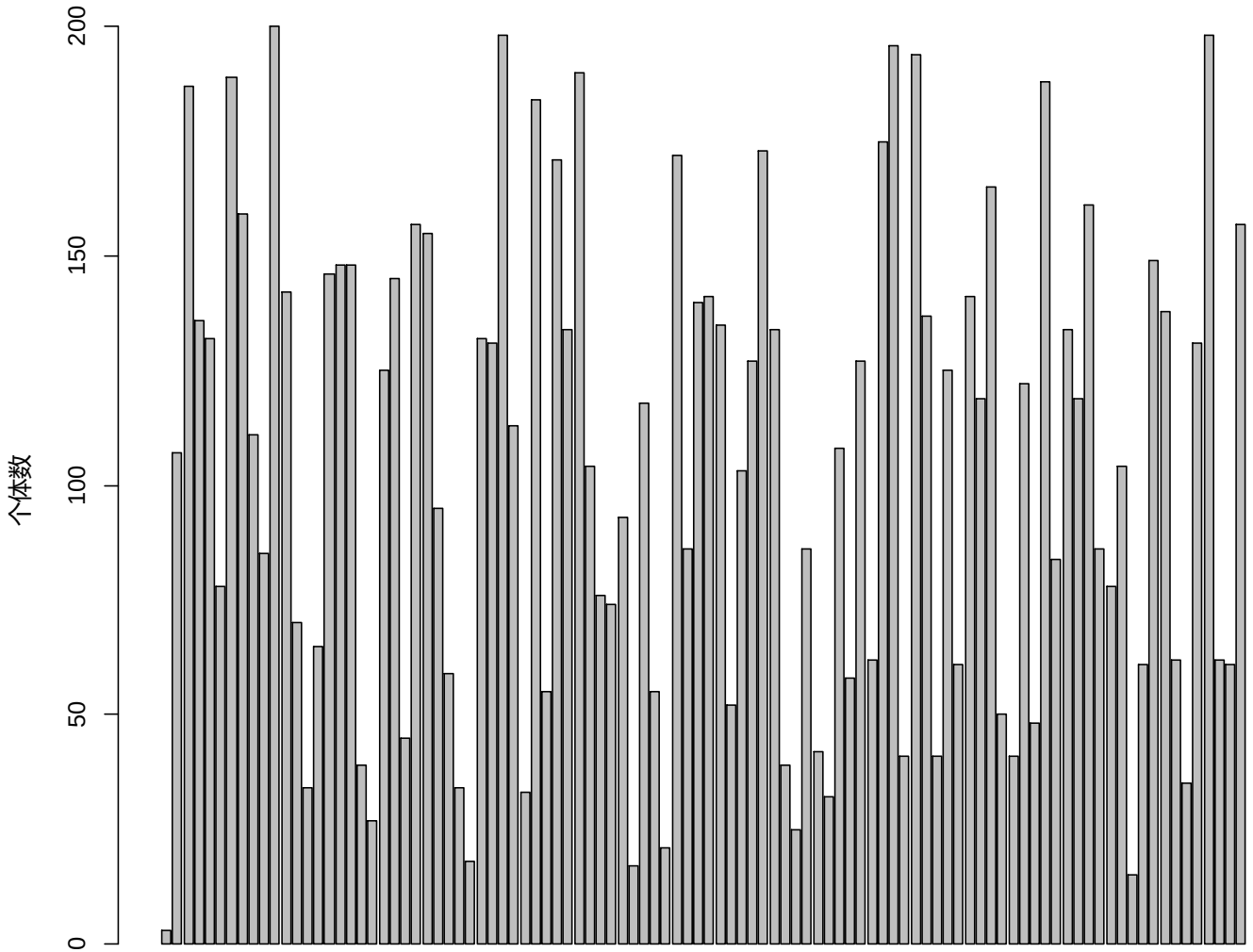


Random distribution (variance = 10μ), 10-16 survey routes needed to reach 95% confident interval: pop.size = μ±(μX20%)

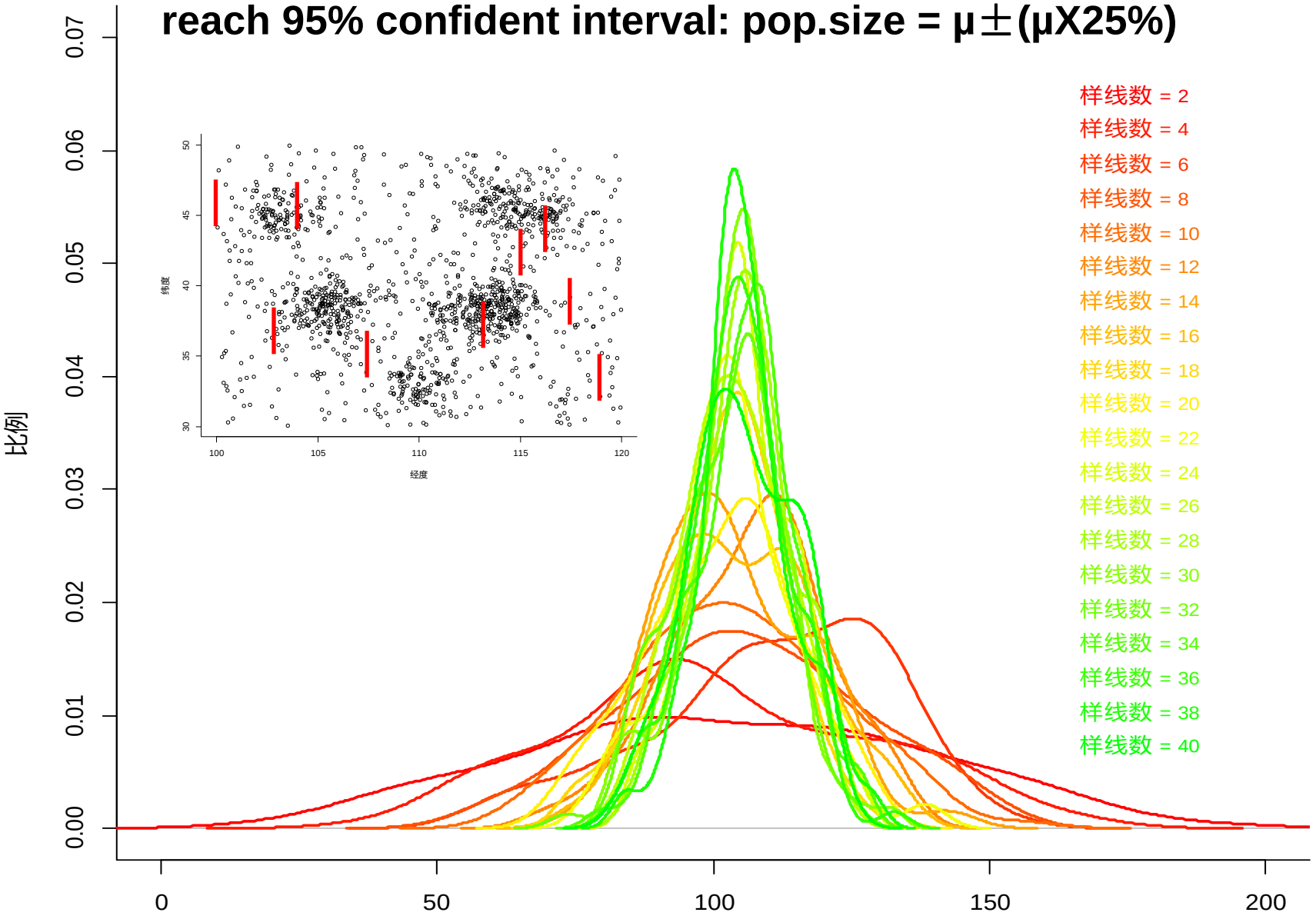


（根据不同样线数量估计动物的密度和数量）

Cluster distribution (variance = 30μ)

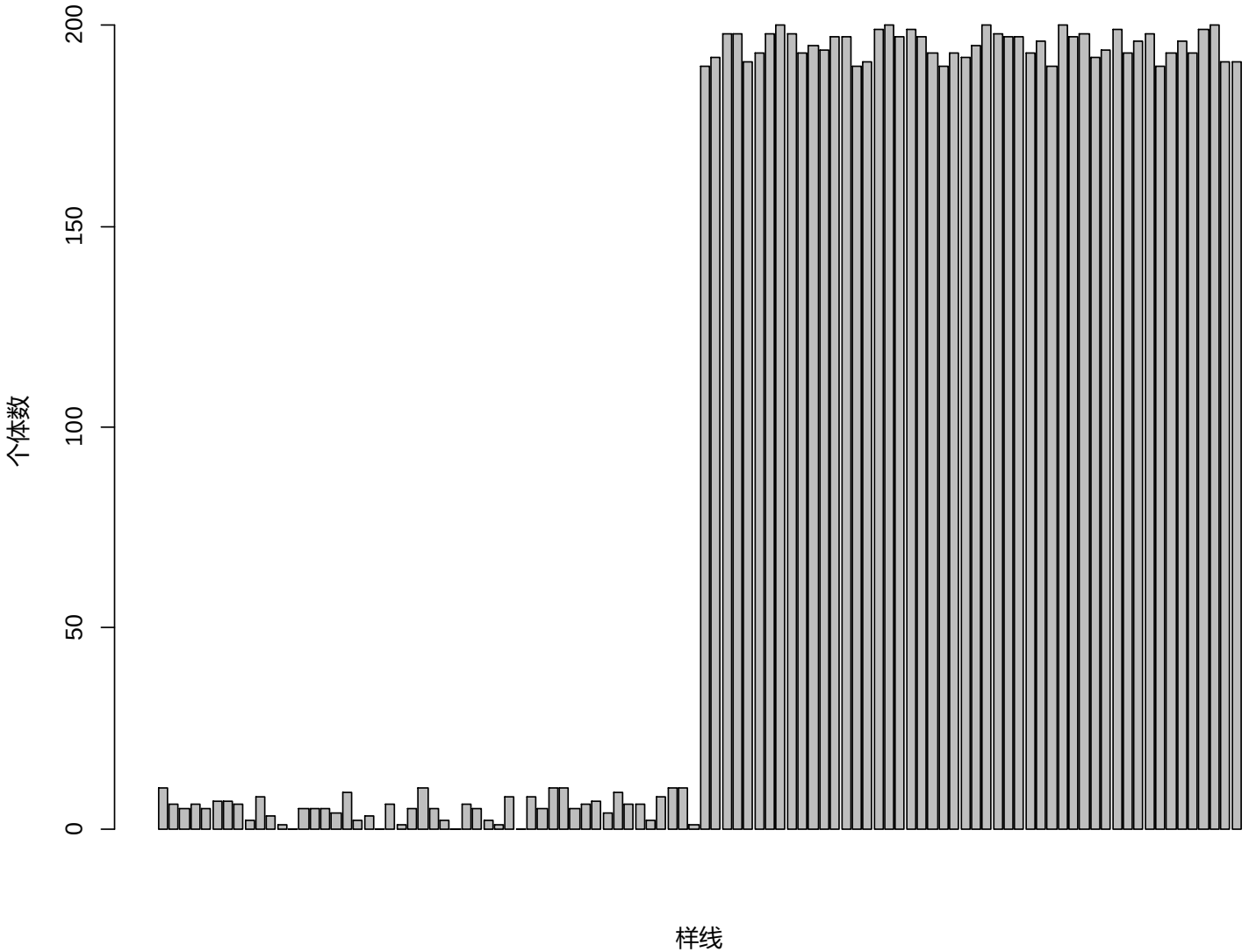


Cluster distribution (variance = 30μ), 24-32 survey routes needed to reach 95% confident interval: pop.size = μ±(μX25%)

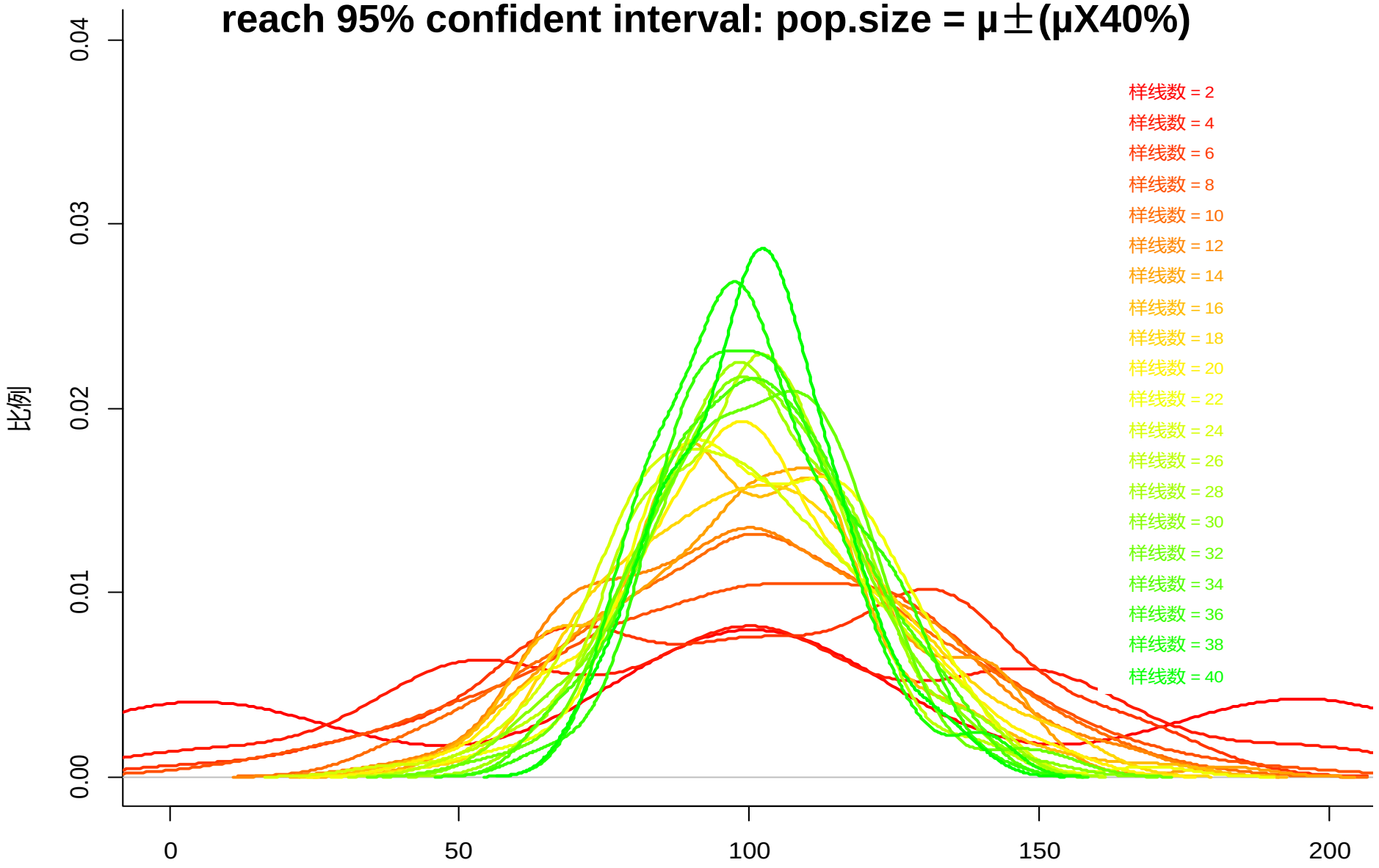


动物数量（根据不同样线数量估计动物的密度和数量）

Cluster distribution (variance = 90μ)

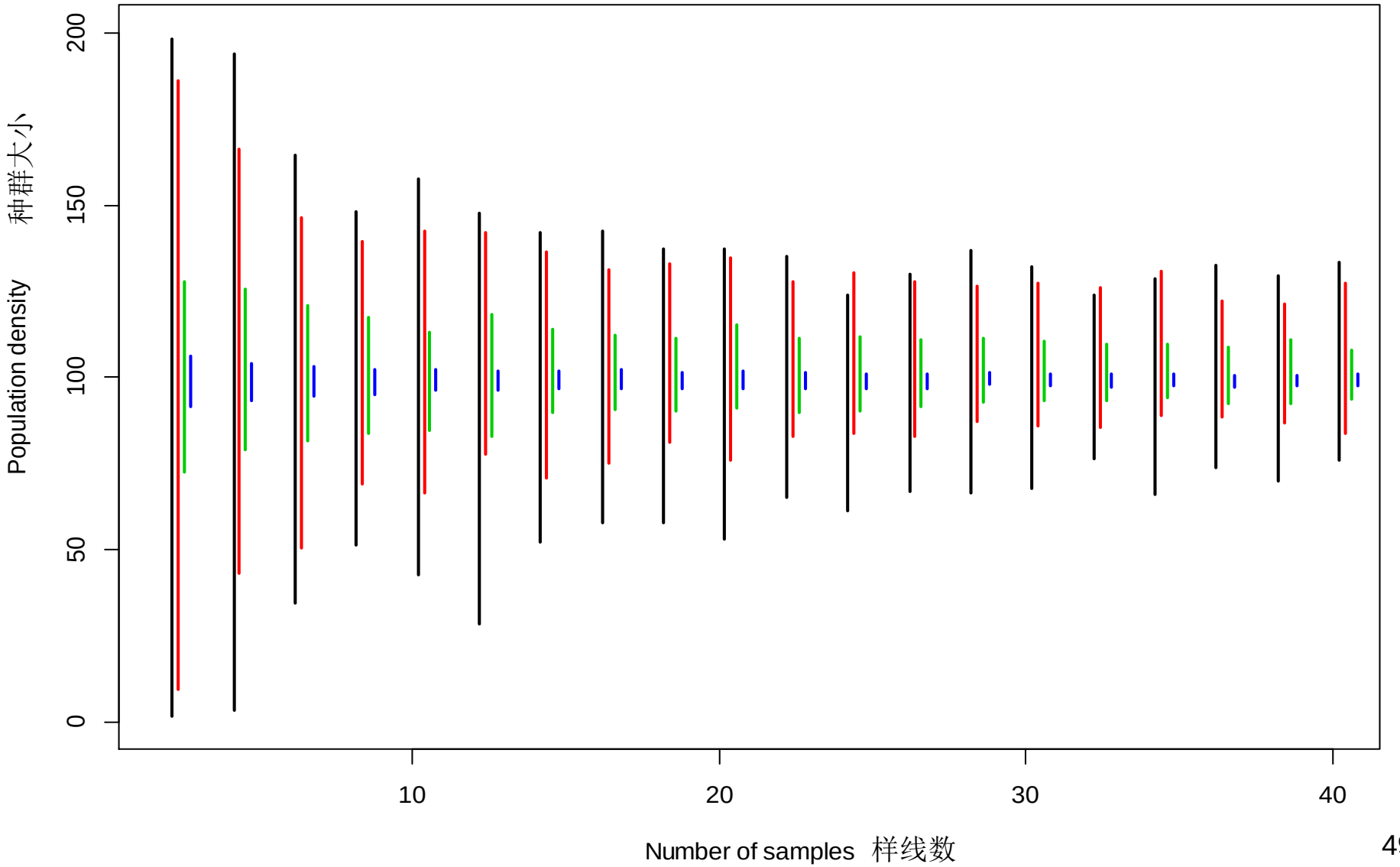


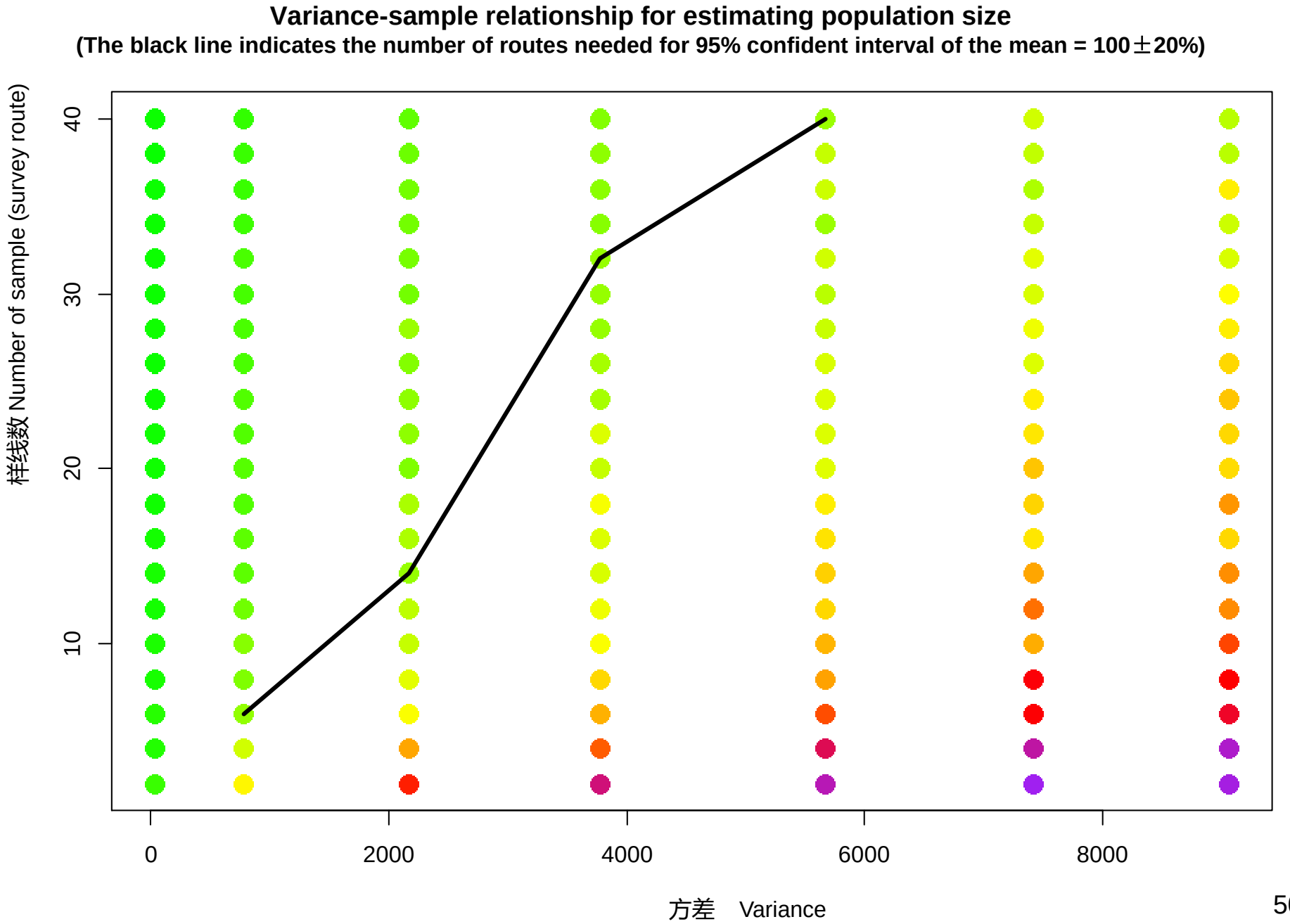
Cluster distribution (variance = 90μ), 36-40 survey routes needed to reach 95% confident interval: pop.size = μ±(μX40%)



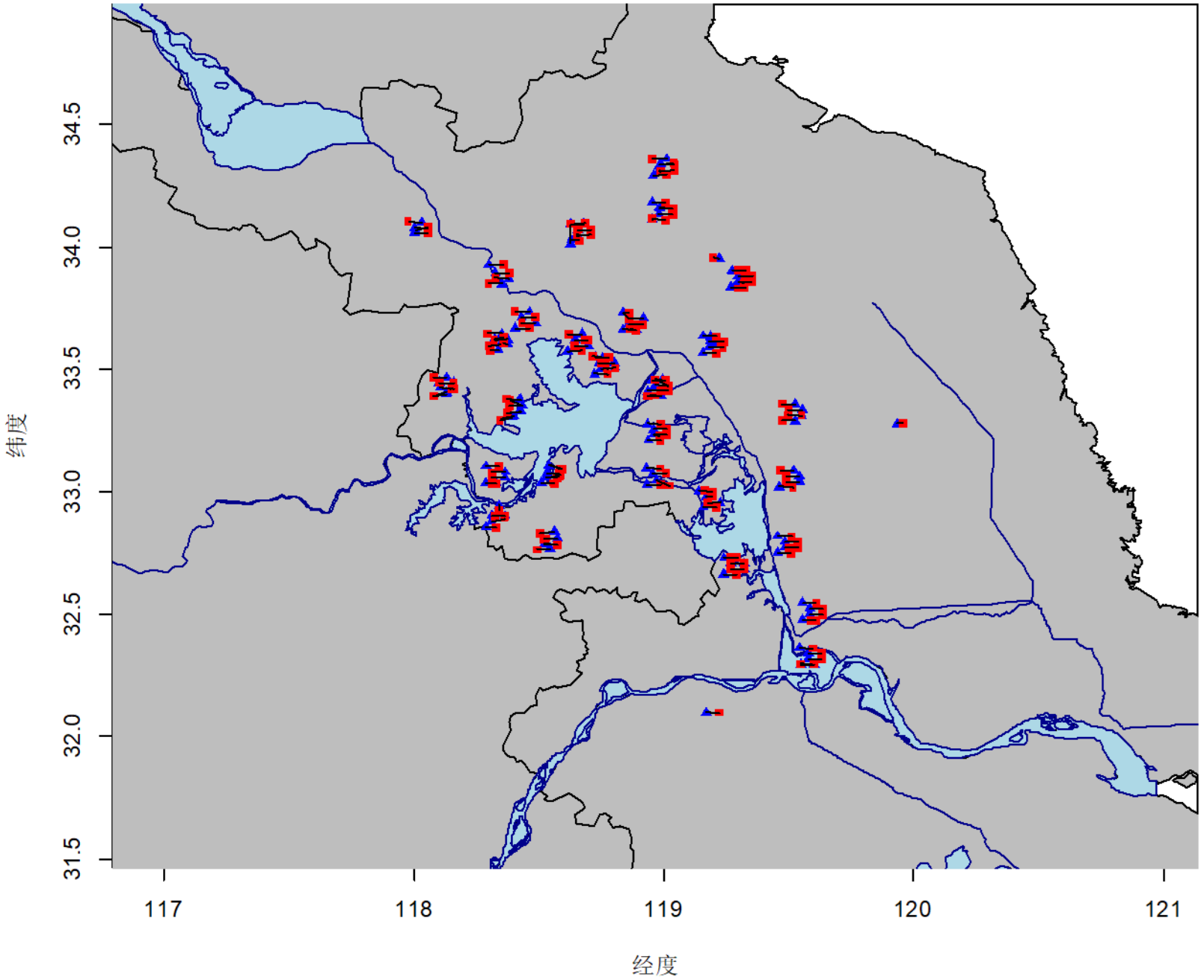
动物数量（根据不同样线数量估计动物的密度和数量）

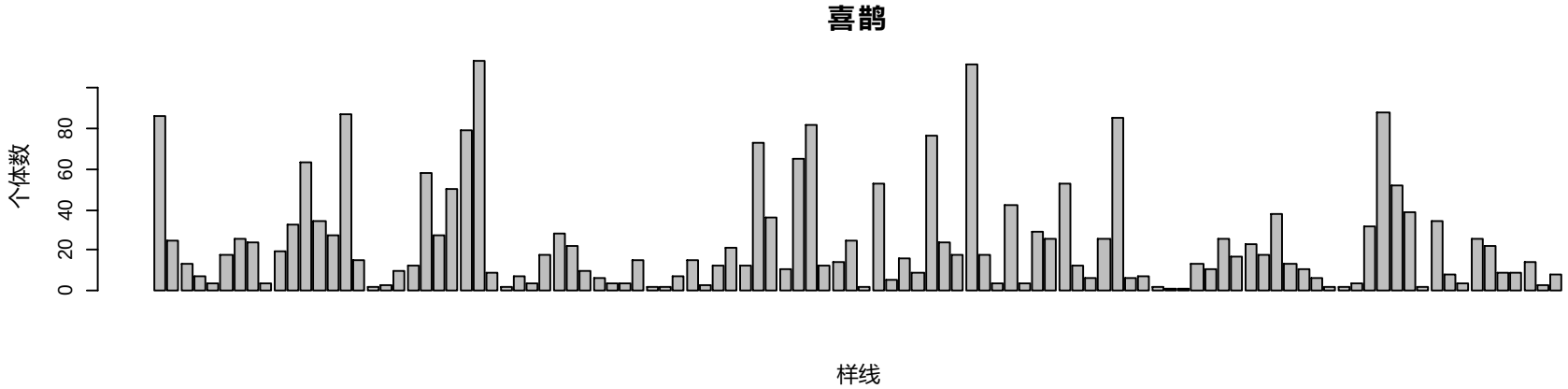
Confident intervals (95%) for estimating population size when the variance of animal count at each sample (survey route) is 90 (black), 30 (red), 10 (green) and 0.3 (blue) times as same as the mean, respectively

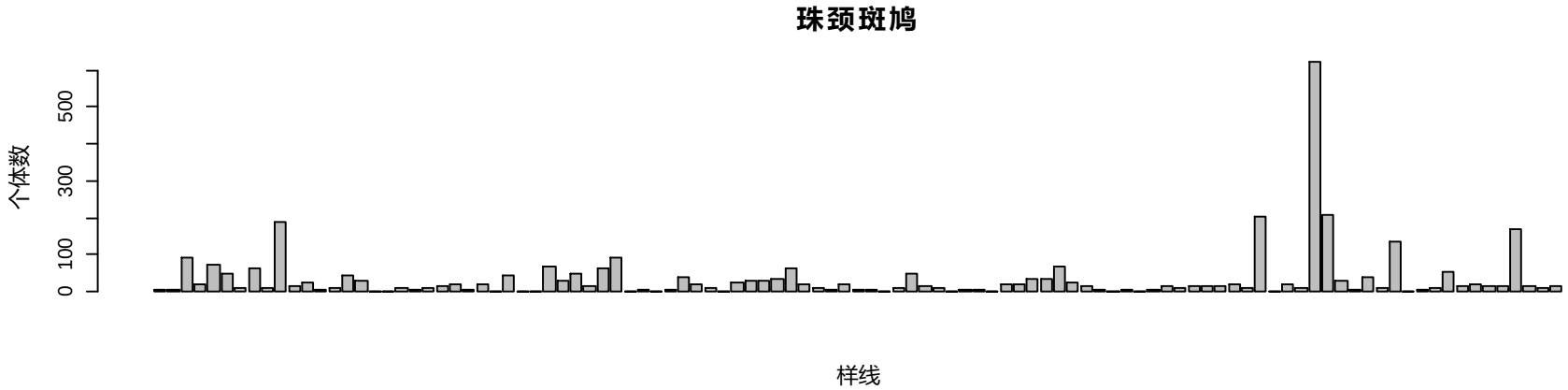


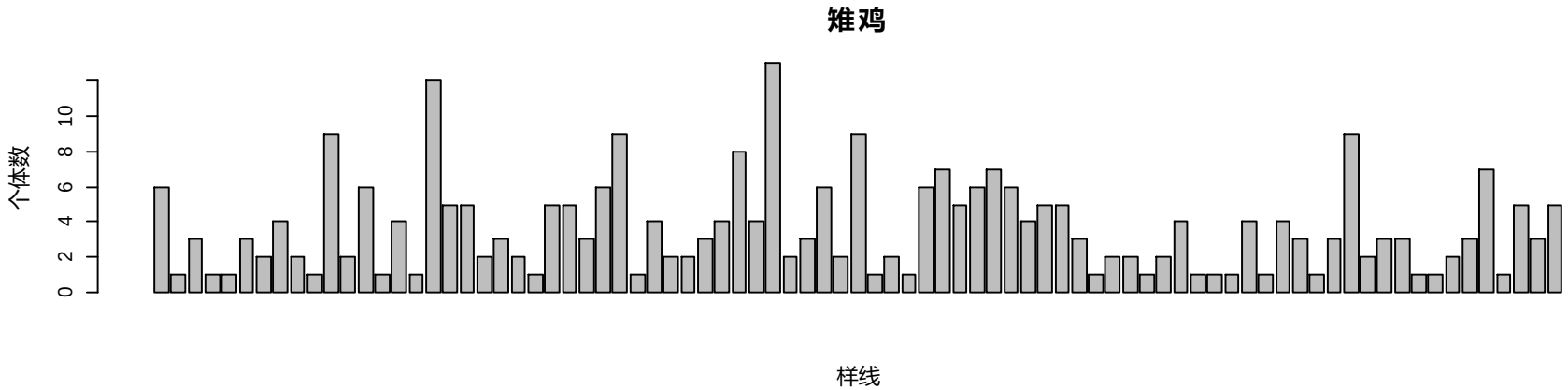

Variance-sample relationship for estimating population size
(The black line indicates the number of routes needed for 95% confident interval of the mean = 100±20%)
样线数 Number of sample (survey route)
10
20
30
40
0
2000
4000
6000
8000
方差 Variance

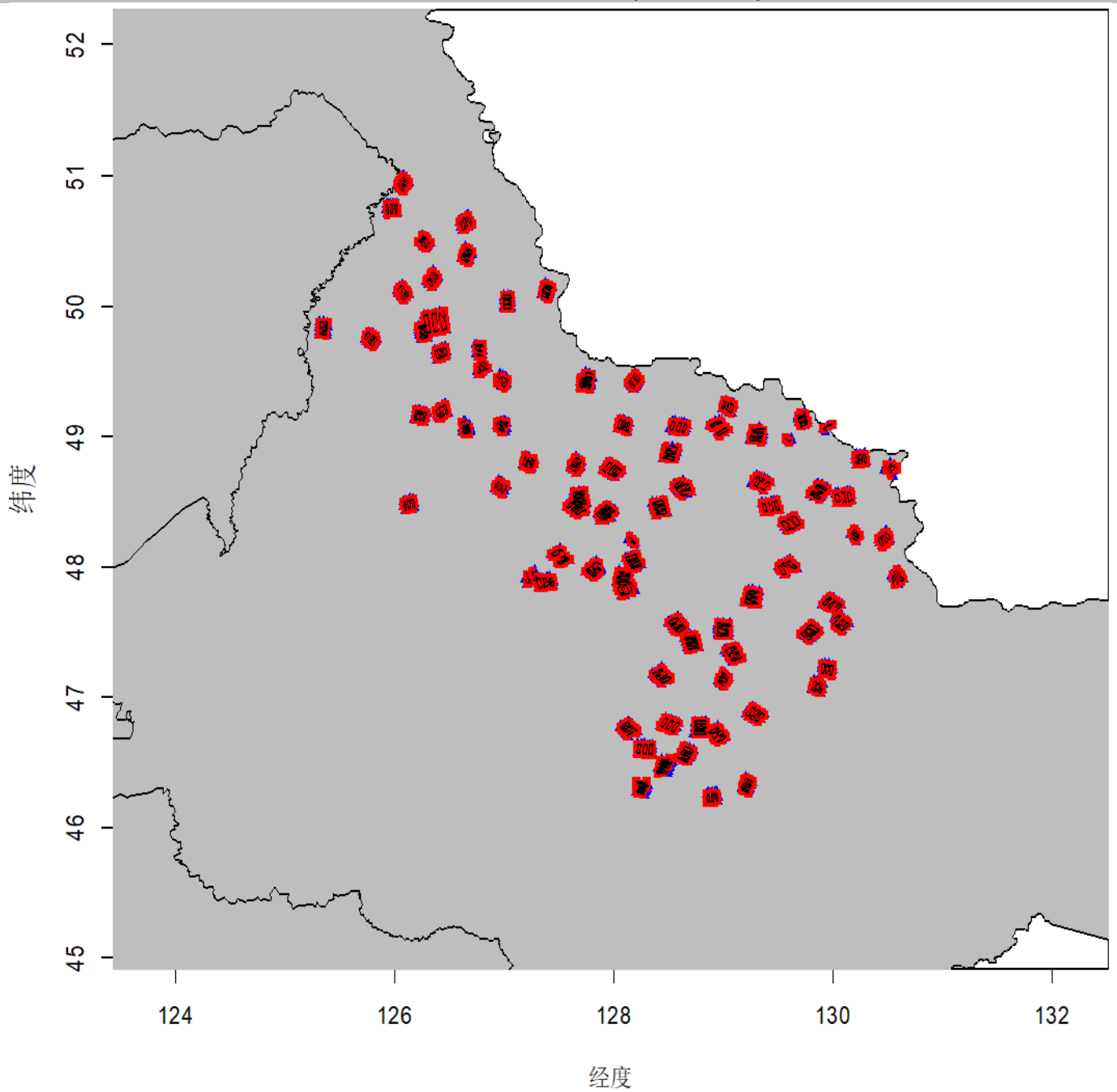
Actual survey results

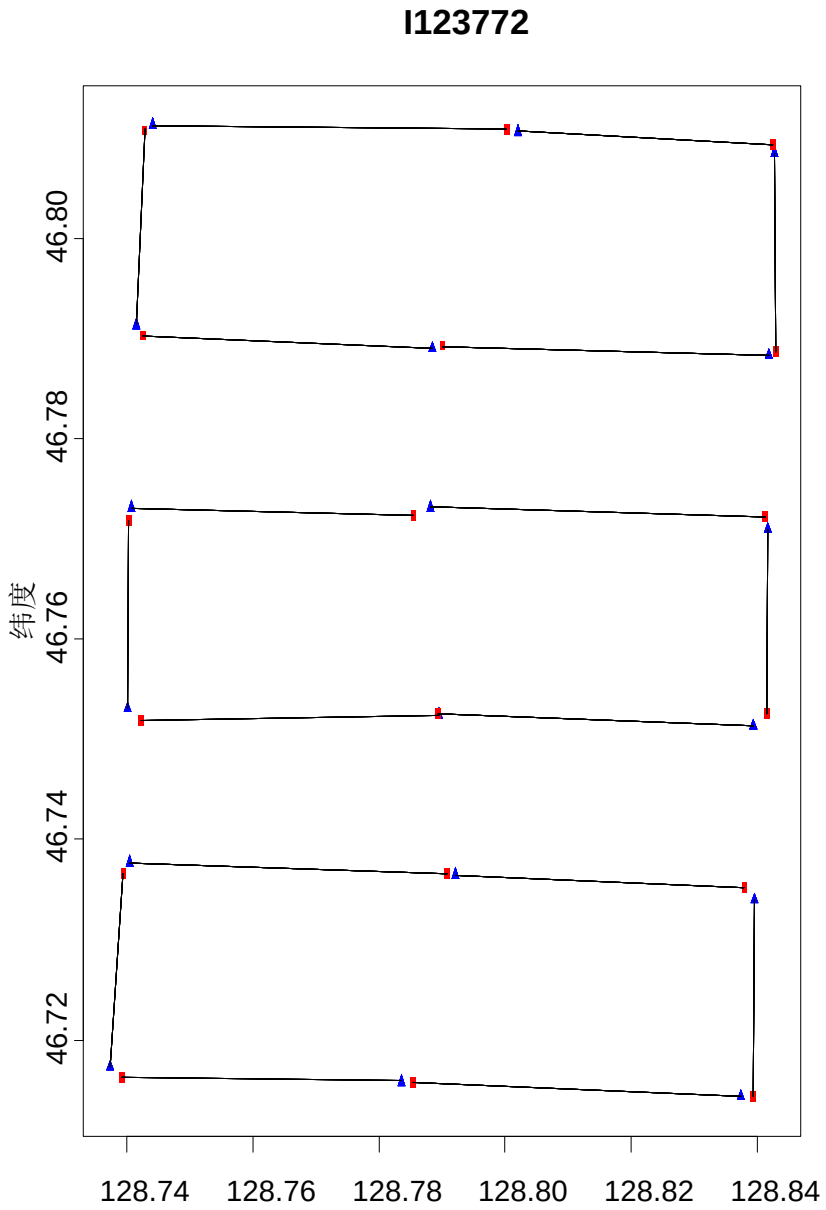

34.5
34.0
33.5
33.0
32.5
32.0
31.5
纬度
117
118
119
120
121
经度

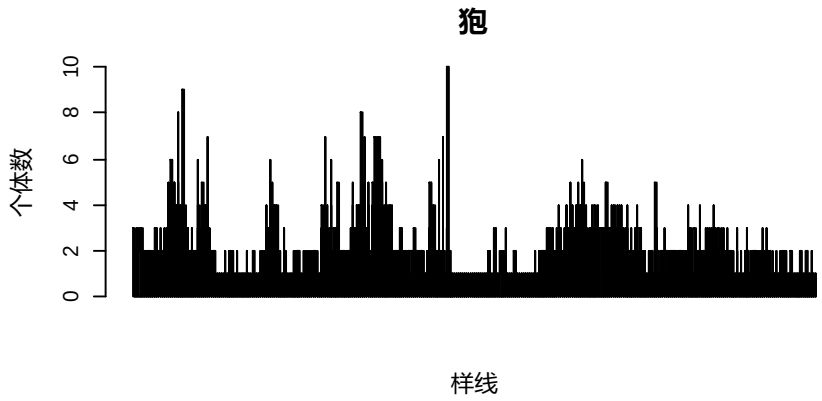

喜鹊
个体数
0
20
40
60
80
样线

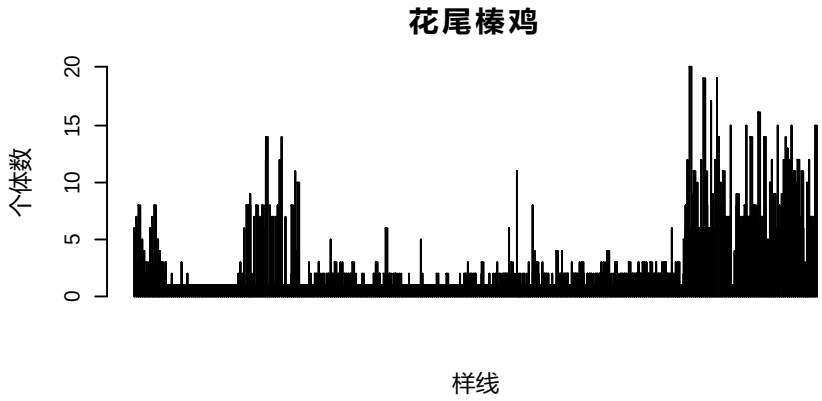

珠颈斑鸠
个体数
0
100
300
500
样线

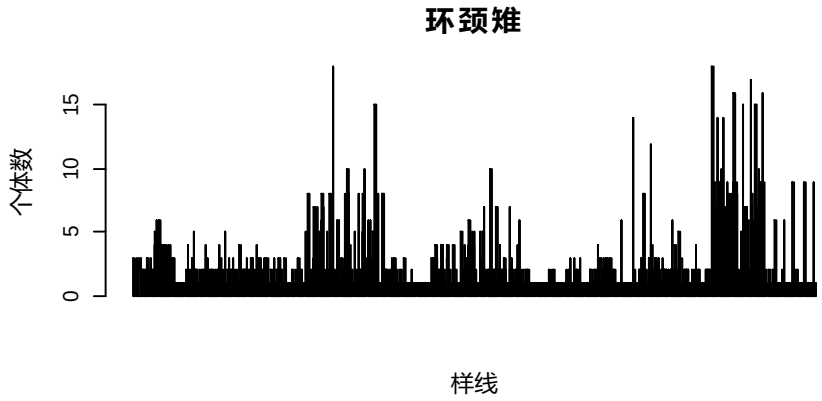

雉鸡
个体数
0
2
4
6
8
10
样线

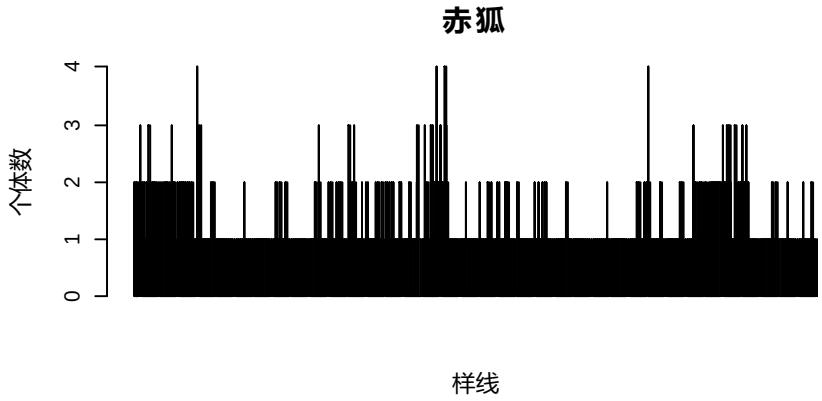

52
51
50
49
48
47
46
45
纬度
124
126
128
130
132
经度

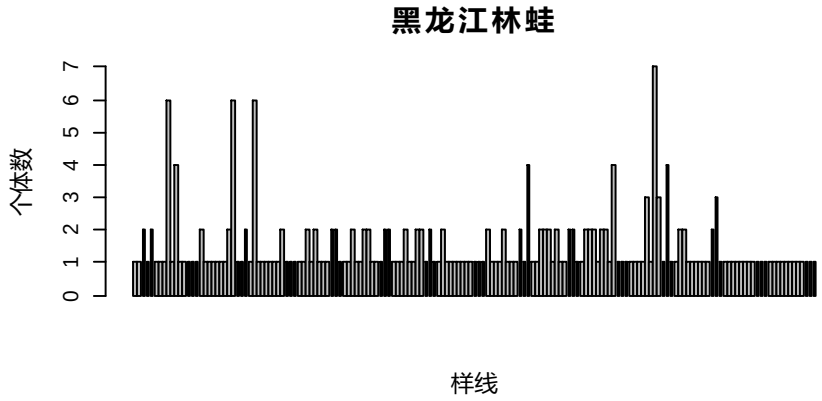

I123772
46.80
46.78
46.76
46.74
46.72
纬度
128.74
128.76
128.78
128.80
128.82
128.84
经度

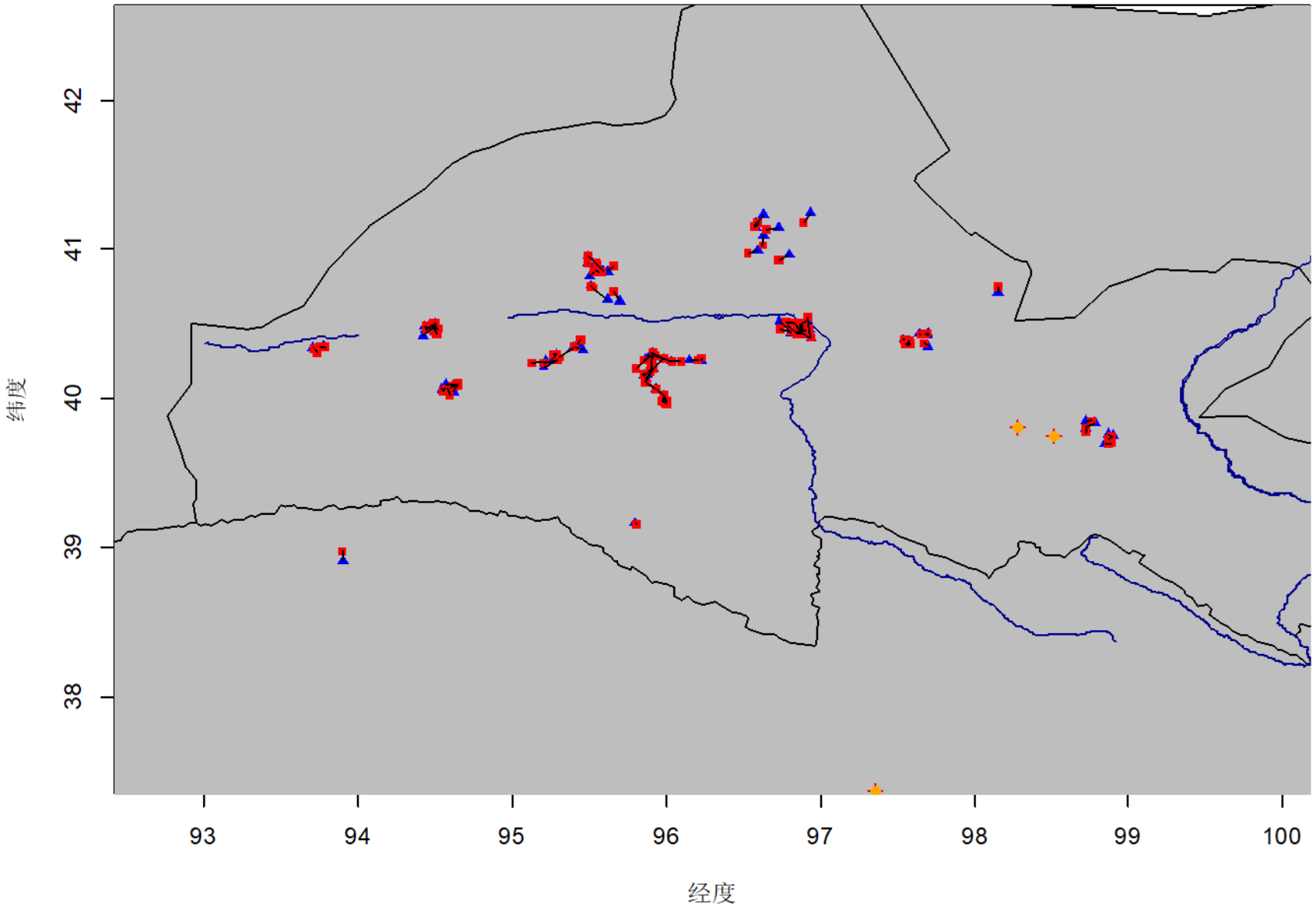

狍
个体数
0
2
4
6
8
10
样线

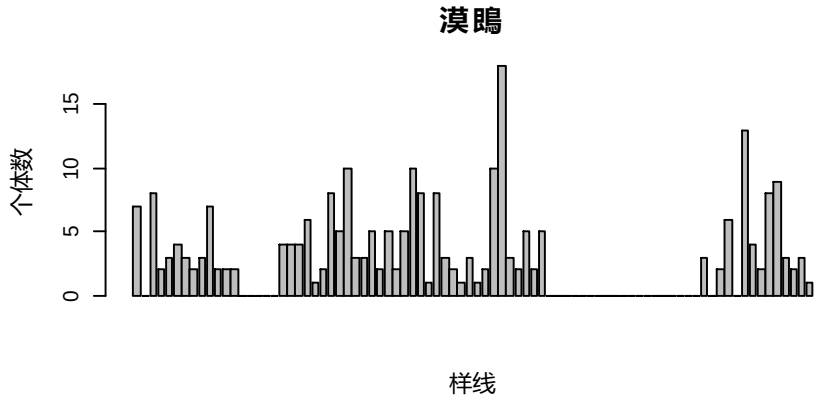

花尾榛鸡
个体数
0
5
10
15
20
样线

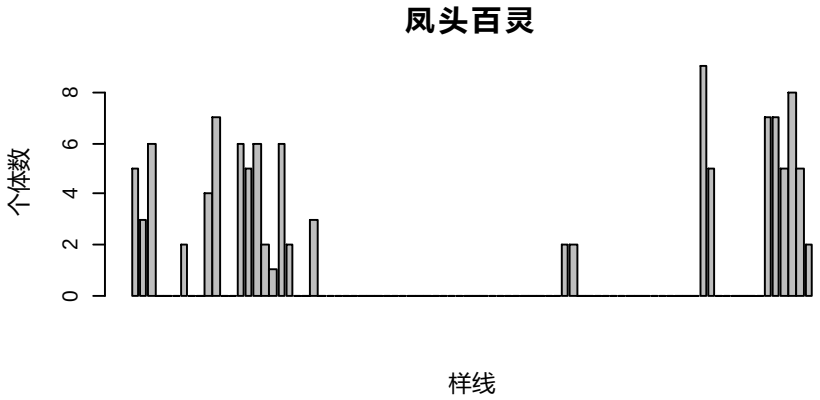

环颈雉
个体数
0
5
10
15
样线

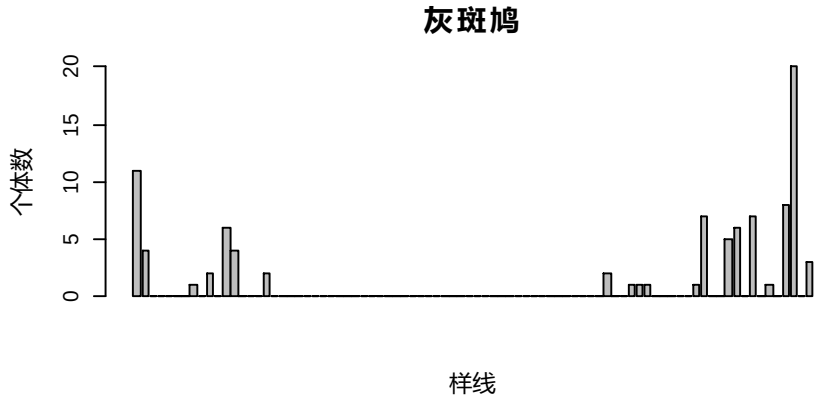

赤狐
个体数
0
1
2
3
4
样线

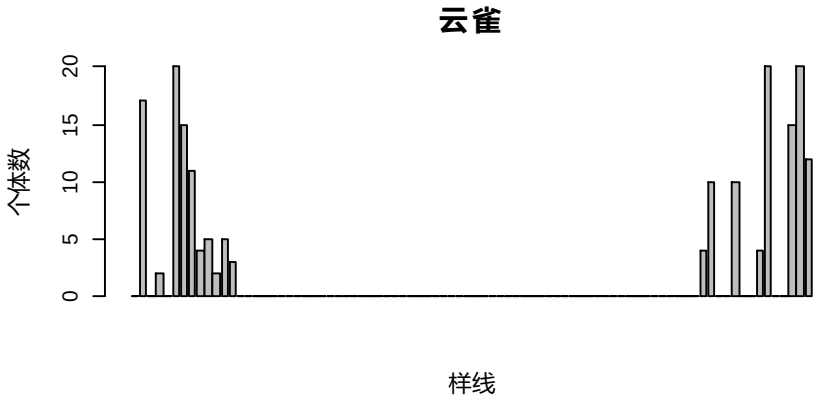

黑龙江林蛙
个体数
0
1
2
3
4
5
6
7
样线

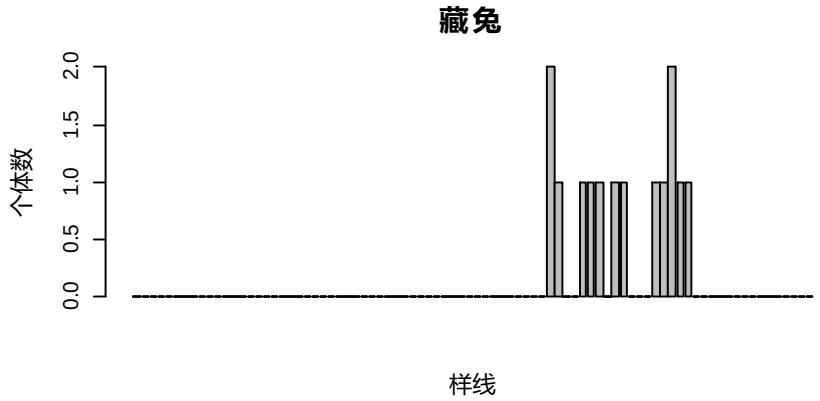

纬度
42
41
40
39
38
93
94
95
96
97
98
99
100
经度

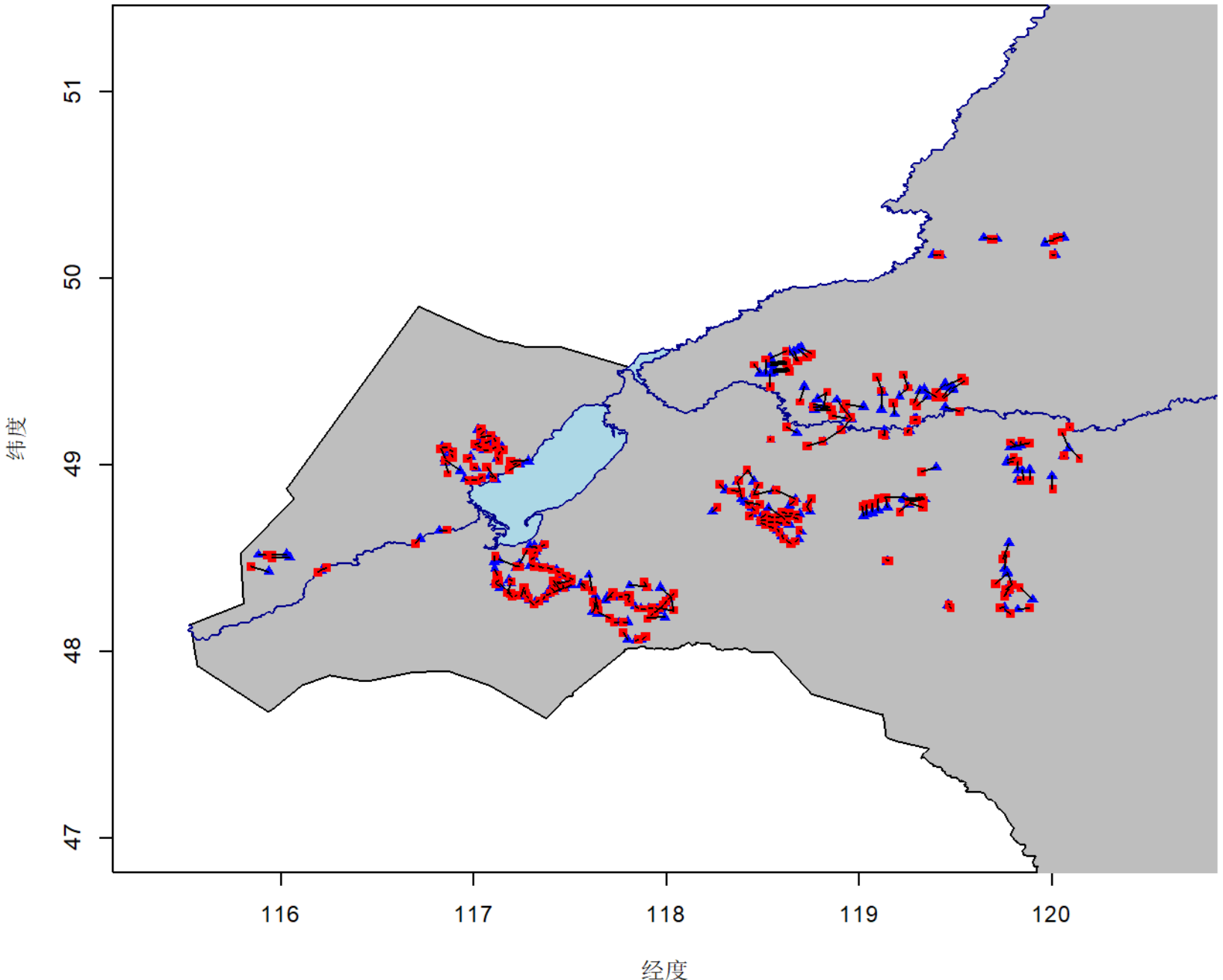

漠鹏
个体数
0
5
10
15
样线

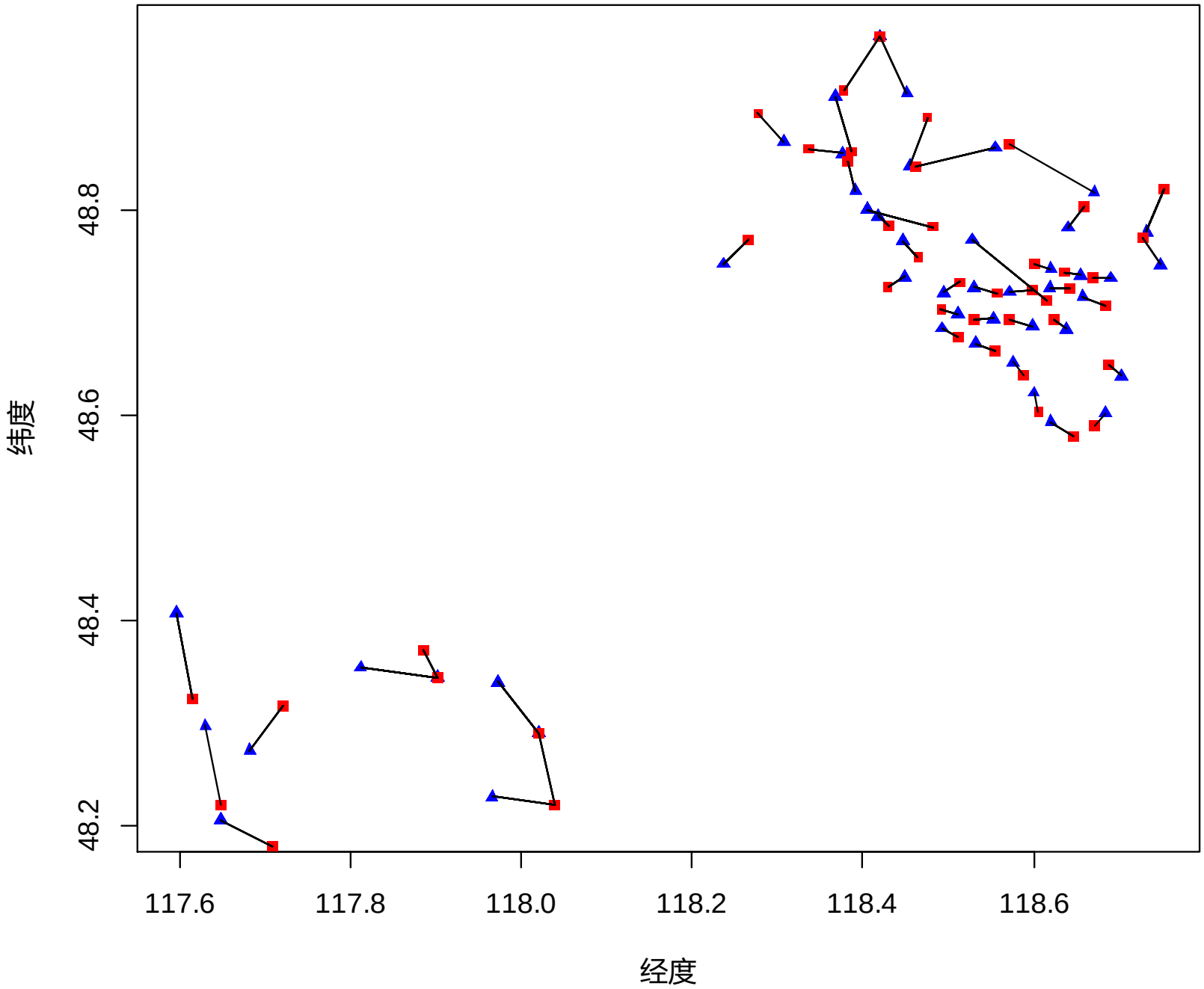

凤头百灵
个体数
0
2
4
6
8
样线


灰斑鸠
个体数
0
5
10
15
20
样线

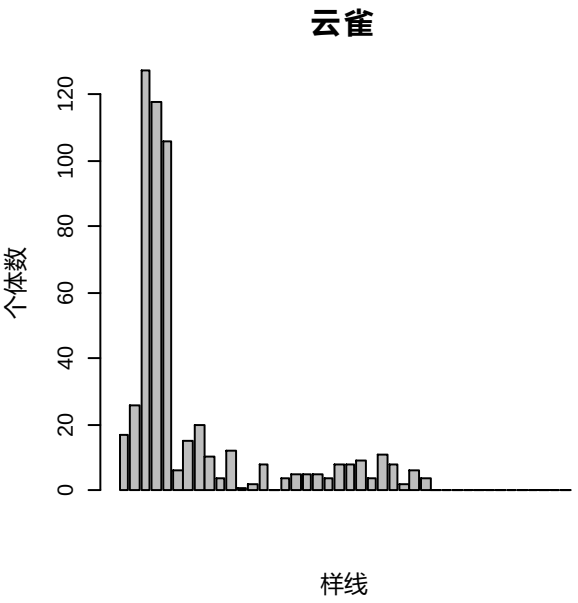

云雀
个体数
0
5
10
15
20
样线

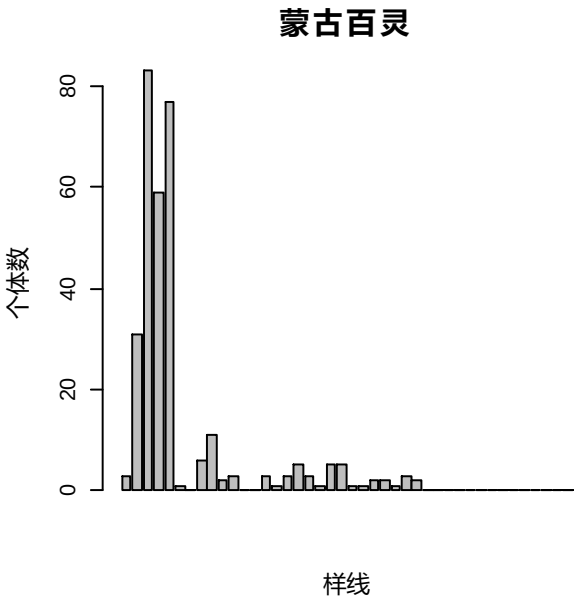

藏兔
个体数
0.0
0.5
1.0
1.5
2.0
样线

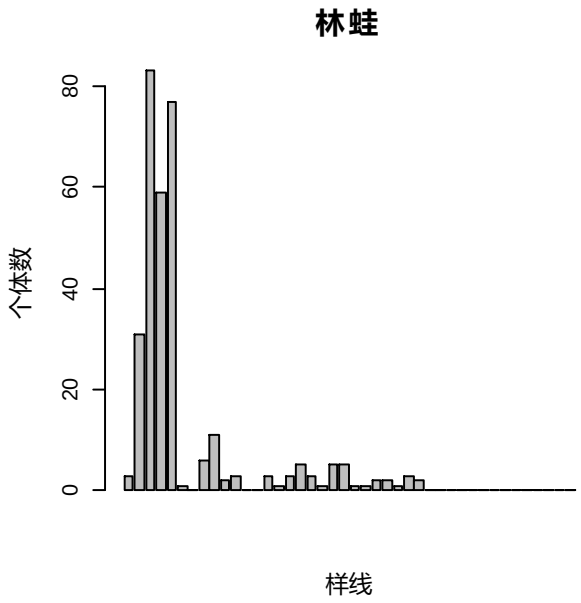

纬度
51
50
49
48
47
116
117
118
119
120
经度

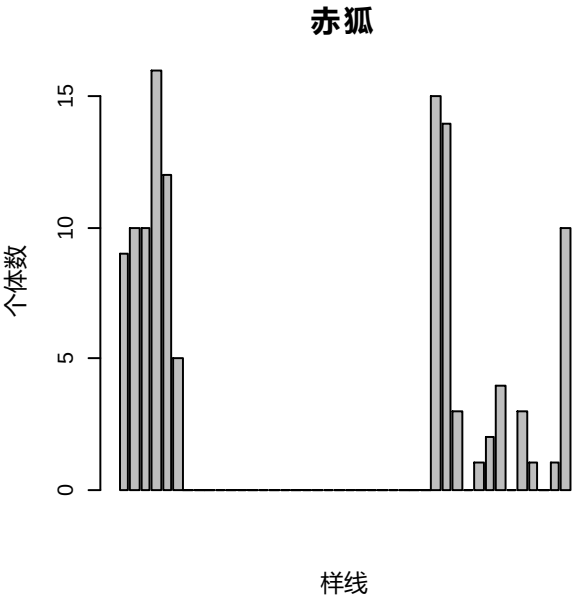

纬度
48.8
48.6
48.4
48.2
117.6
117.8
118.0
118.2
118.4
118.6
经度

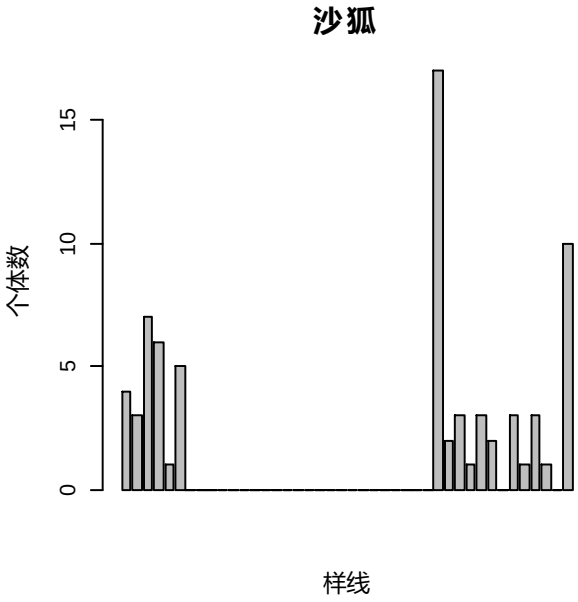
	样线	云雀	蒙古百灵	林蛙	赤狐	沙狐	草兔
1	3-06-079-201-101	17	3	3	9	4	4
2	3-06-079-201-102	26	31	31	10	3	2
3	3-06-079-201-103	127	83	83	10	7	7
4	3-06-079-201-104	118	59	59	16	6	7
5	3-06-079-201-105	106	77	77	12	1	2
6	3-06-079-201-106	6	1	1	5	5	2
7	3-06-079-201-107	15	0	0	0	0	0
8	3-06-079-201-108	20	6	6	0	0	0
9	3-06-079-201-109	10	11	11	0	0	0
10	3-06-079-201-110	4	2	2	0	0	0
11	3-06-079-201-111	12	3	3	0	0	0
12	3-06-079-201-112	1	0	0	0	0	0
13	3-06-079-201-113	2	0	0	0	0	0
14	3-06-079-201-114	8	3	3	0	0	0
15	3-06-079-201-115	0	1	1	0	0	0
16	3-06-079-201-116	4	3	3	0	0	0
17	3-06-079-201-117	5	5	5	0	0	0
18	3-06-079-201-118	5	3	3	0	0	0
19	3-06-079-201-119	5	1	1	0	0	0
20	3-06-079-201-120	4	5	5	0	0	0
21	3-06-079-201-121	8	5	5	0	0	0
22	3-06-079-201-122	8	1	1	0	0	0
23	3-06-079-201-123	9	1	1	0	0	0
24	3-06-079-201-124	4	2	2	0	0	0
25	3-06-079-201-125	11	2	2	0	0	0
26	3-06-079-201-126	8	1	1	0	0	0
27	3-06-079-201-128	2	3	3	0	0	0
28	3-06-079-201-129	6	2	2	0	0	0
29	3-06-079-201-130	4	0	0	0	0	0
30	3-06-079-201-201	0	0	0	15	17	9
31	3-06-079-202-101	0	0	0	14	2	6
32	3-06-079-202-201	0	0	0	3	3	0
33	3-06-079-202-202	0	0	0	0	1	0

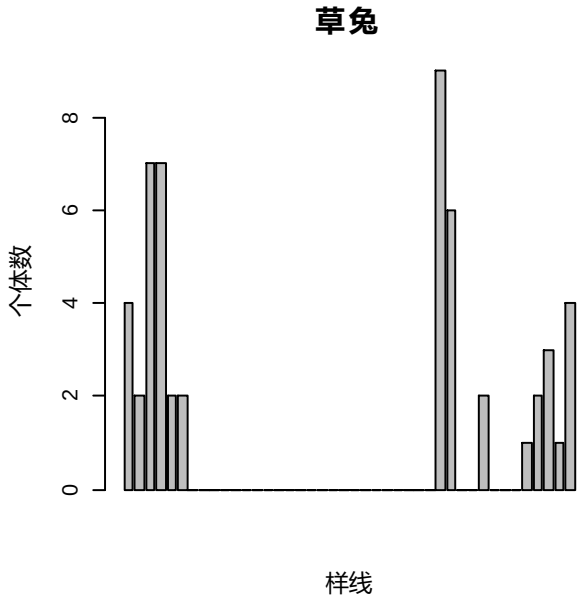

云雀
个体数
0
20
40
60
80
100
120
样线


蒙古百灵
个体数
0
20
40
60
80
样线

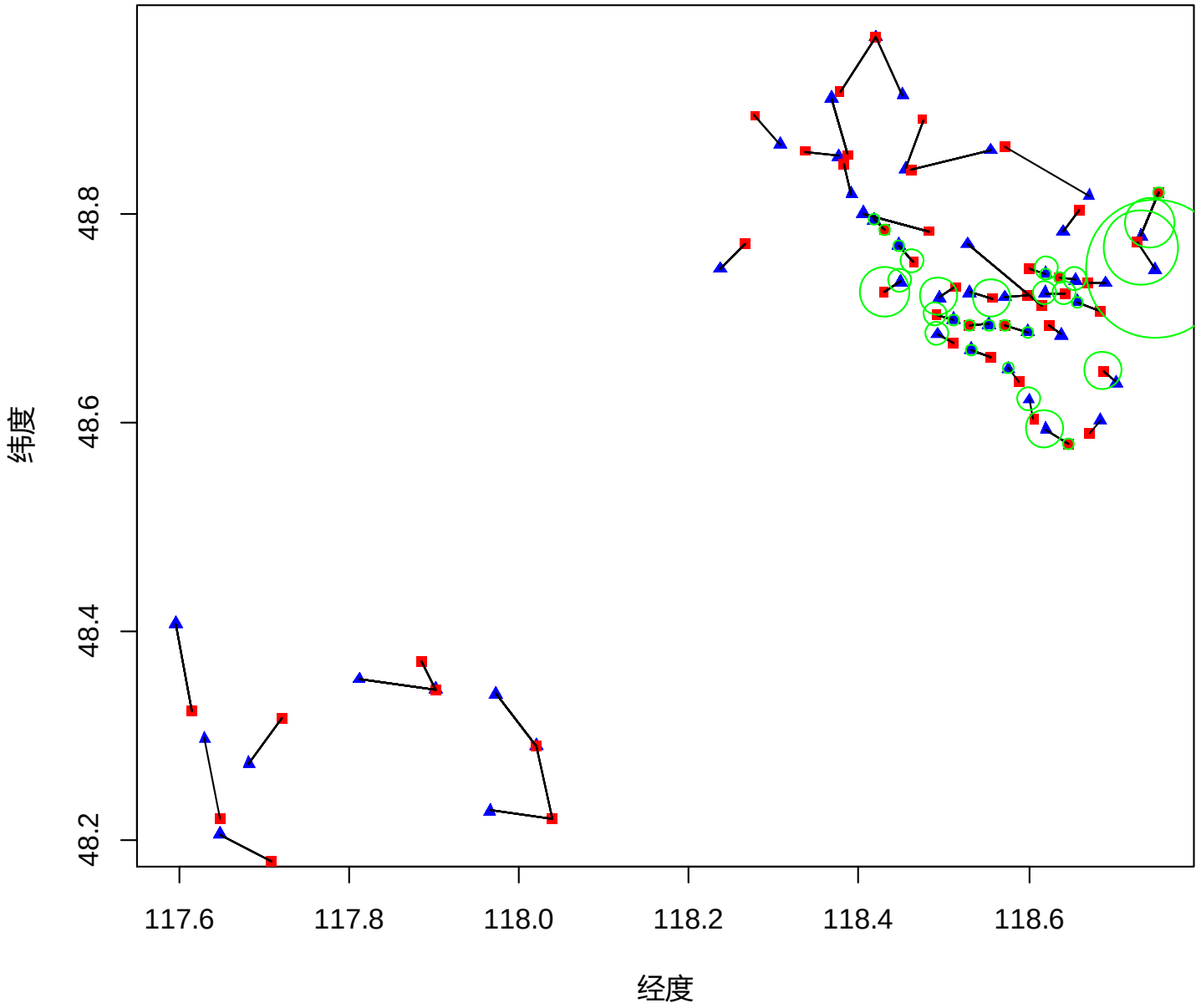

林蛙
个体数
0
20
40
60
80
样线

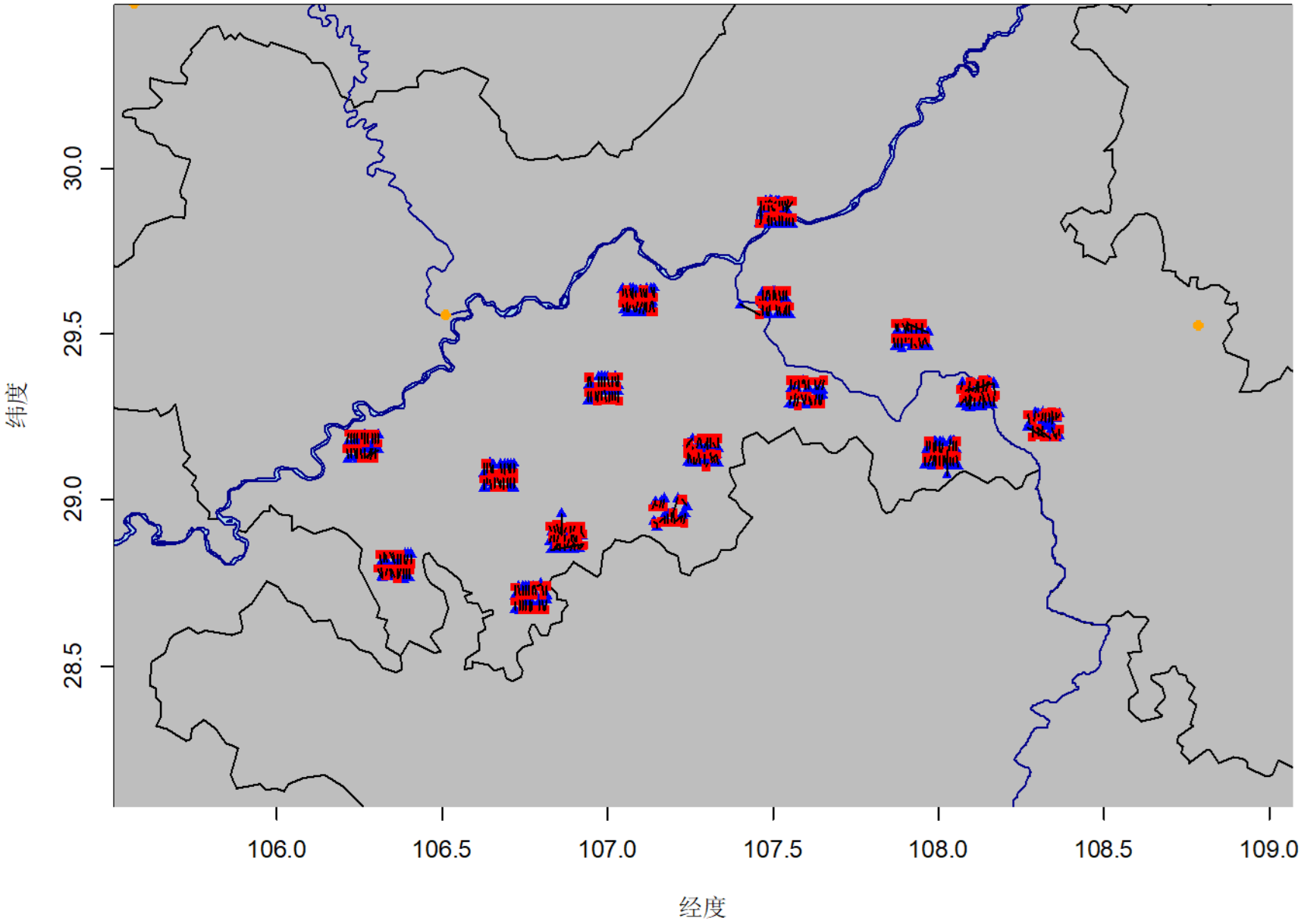

赤狐
个体数
0
5
10
15
样线


沙狐
个体数
0
5
10
15
样线

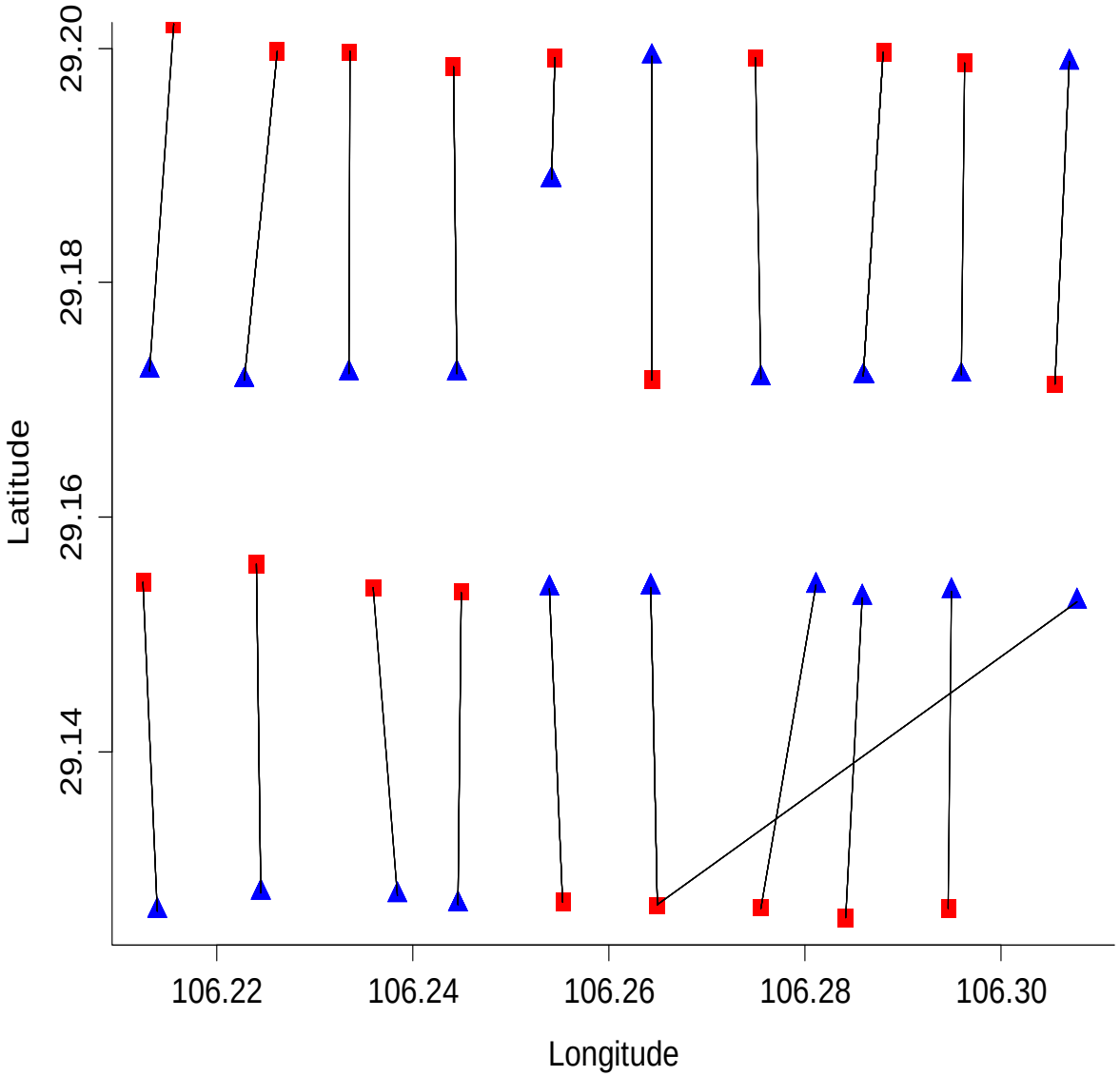

草兔
个体数
0
2
4
6
8
样线

云雀

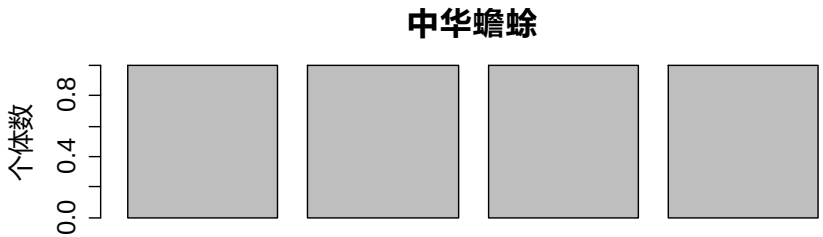


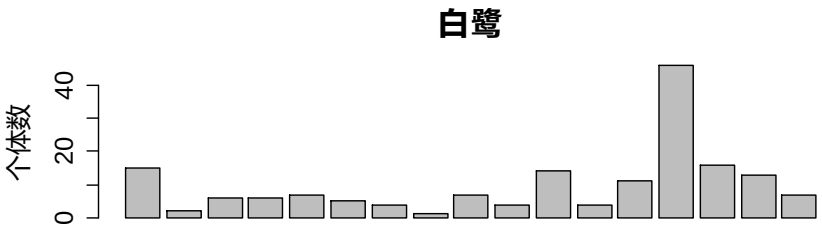

30.0
29.5
29.0
28.5
纬度
106.0
106.5
107.0
107.5
108.0
108.5
109.0
经度

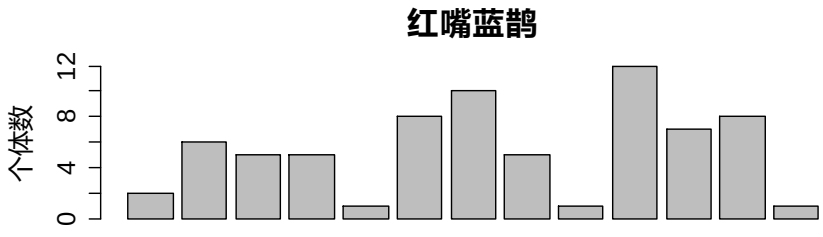
样区10的样线

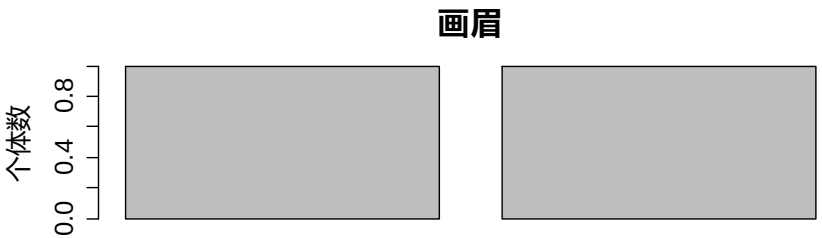

Latitude
29.20
29.18
29.16
29.14
106.22
106.24
106.26
106.28
106.30
Longitude

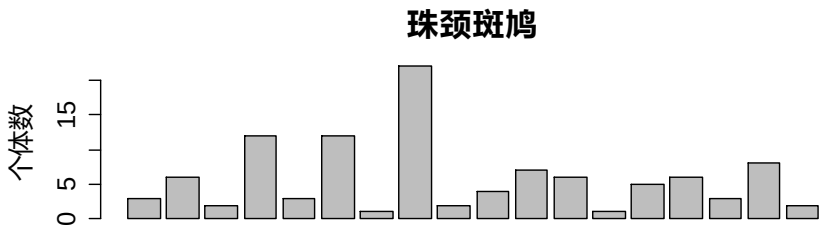
样区10各样线的物种数

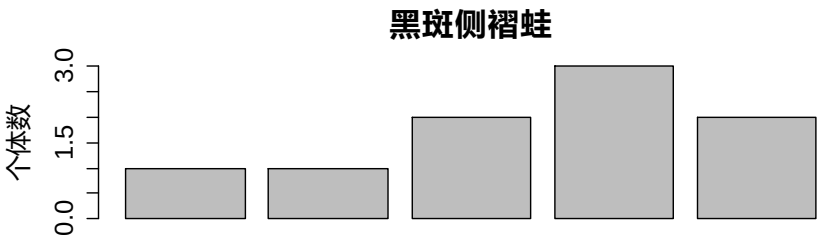


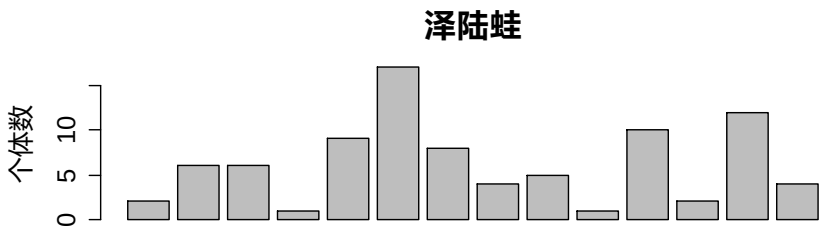




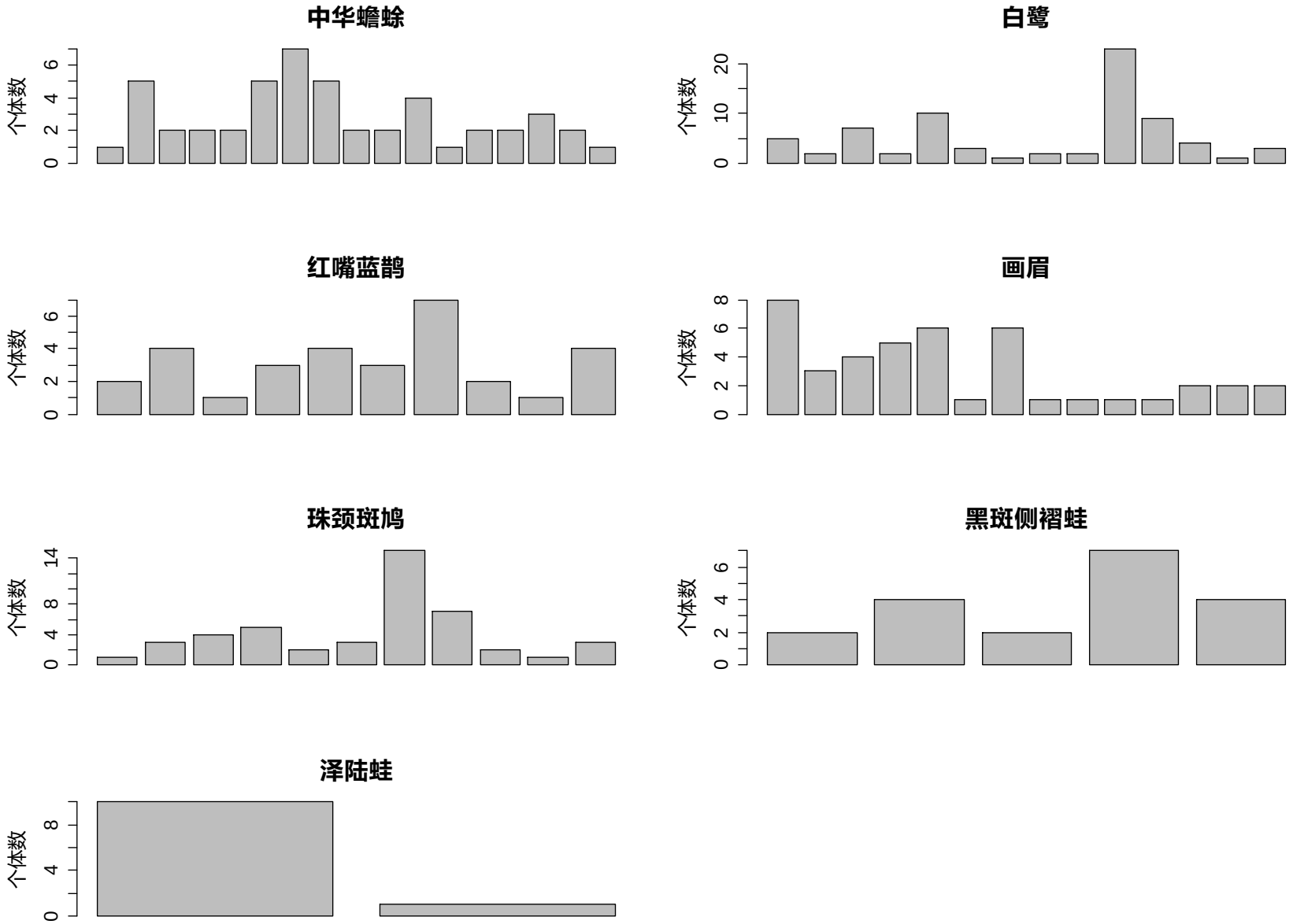




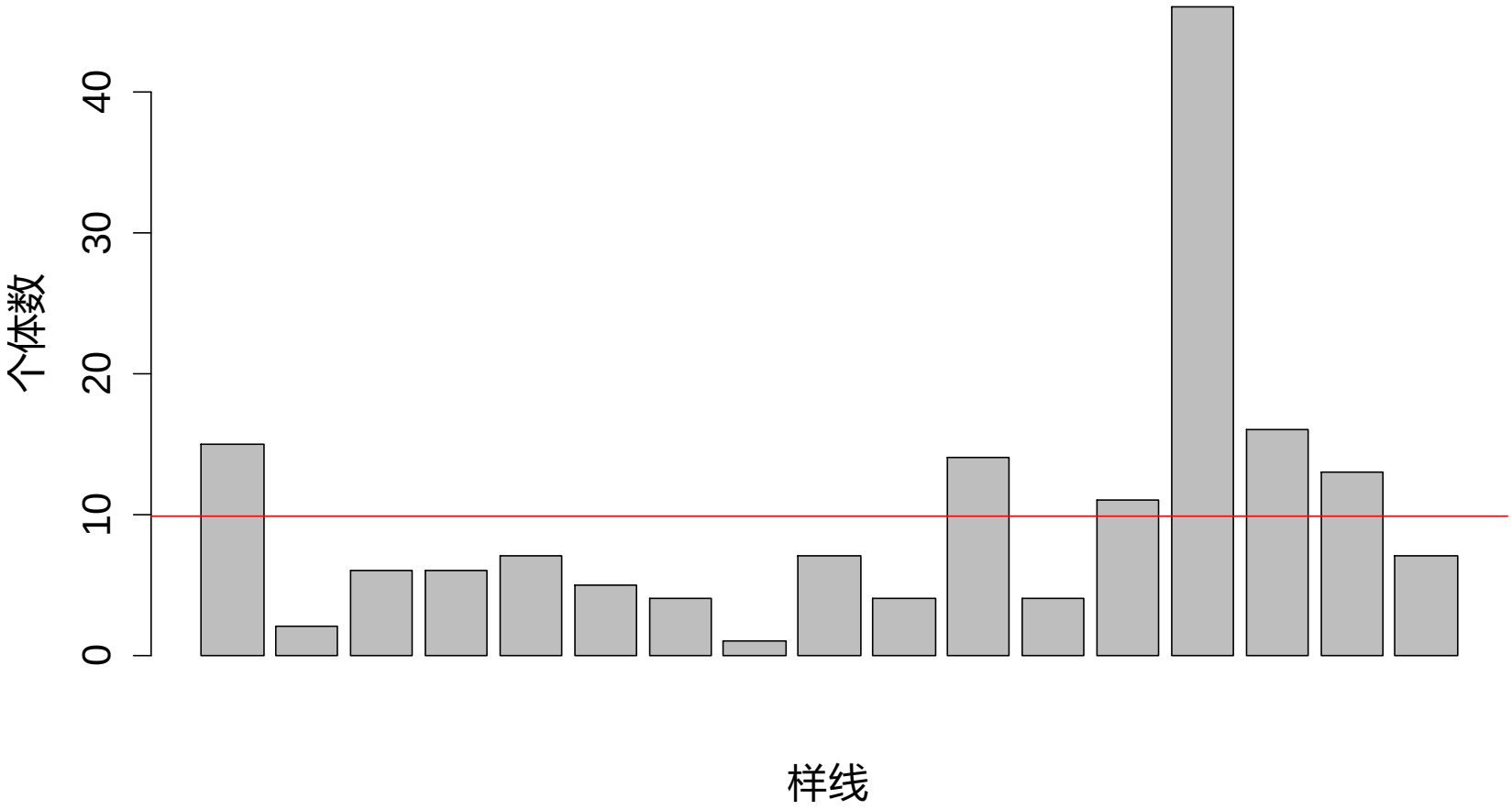




样区15各样线的物种数

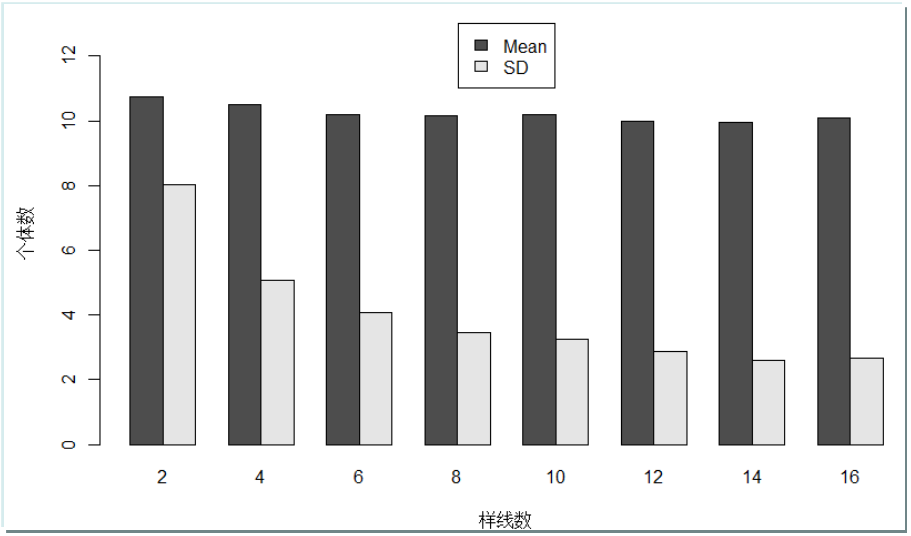


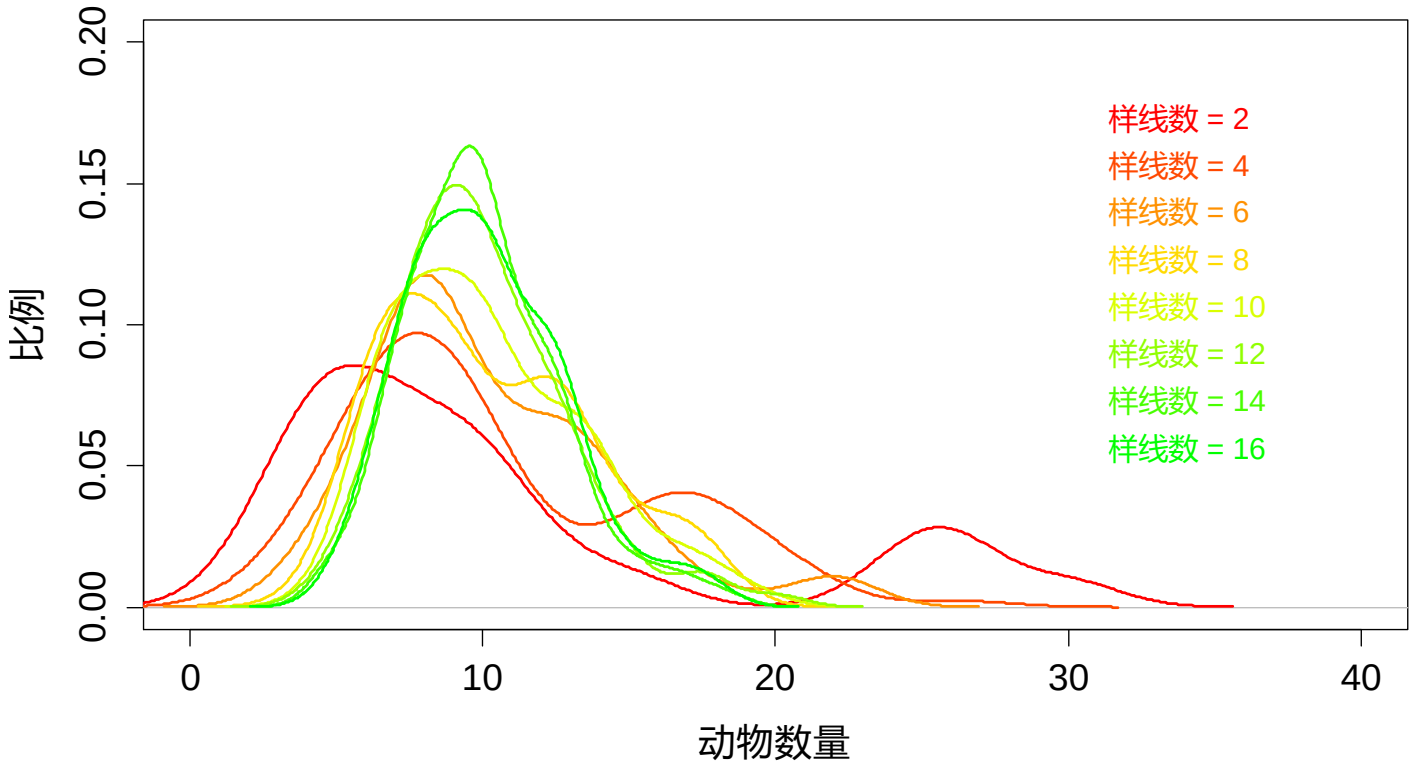
样区10 各样线白鹭的数量



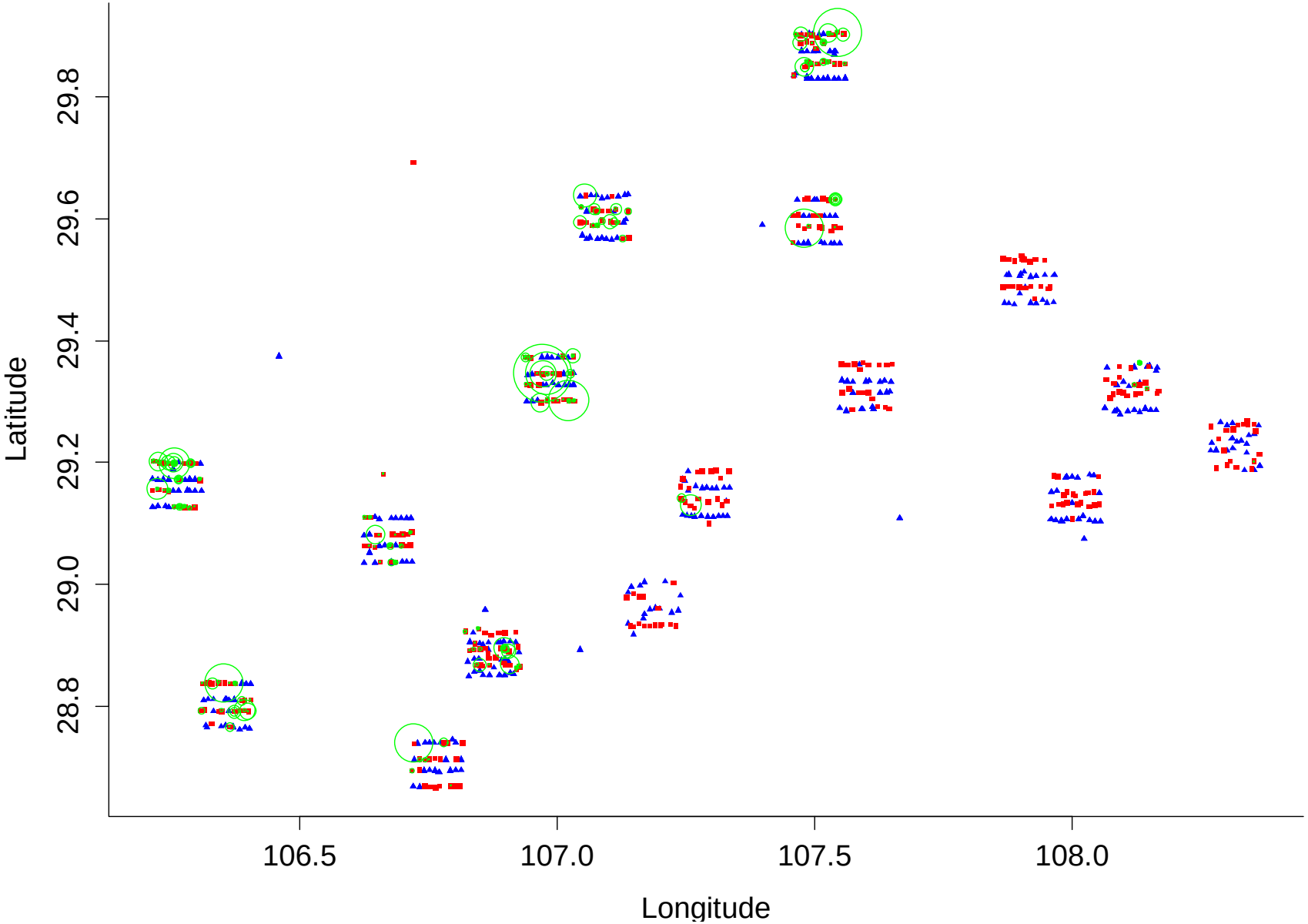
样线数量与估计的动物密度

样线数	2	4	6	8	10	12	14	16
均值	10.75	10.52	10.21	10.18	10.21	9.99	9.96	10.08
标准差	8.04	5.09	4.07	3.48	3.26	2.89	2.59	2.68
95%低	-5.34	0.35	2.07	3.22	3.69	4.22	4.77	4.72
95%高	26.83	20.69	18.34	17.13	16.72	15.77	15.15	15.45

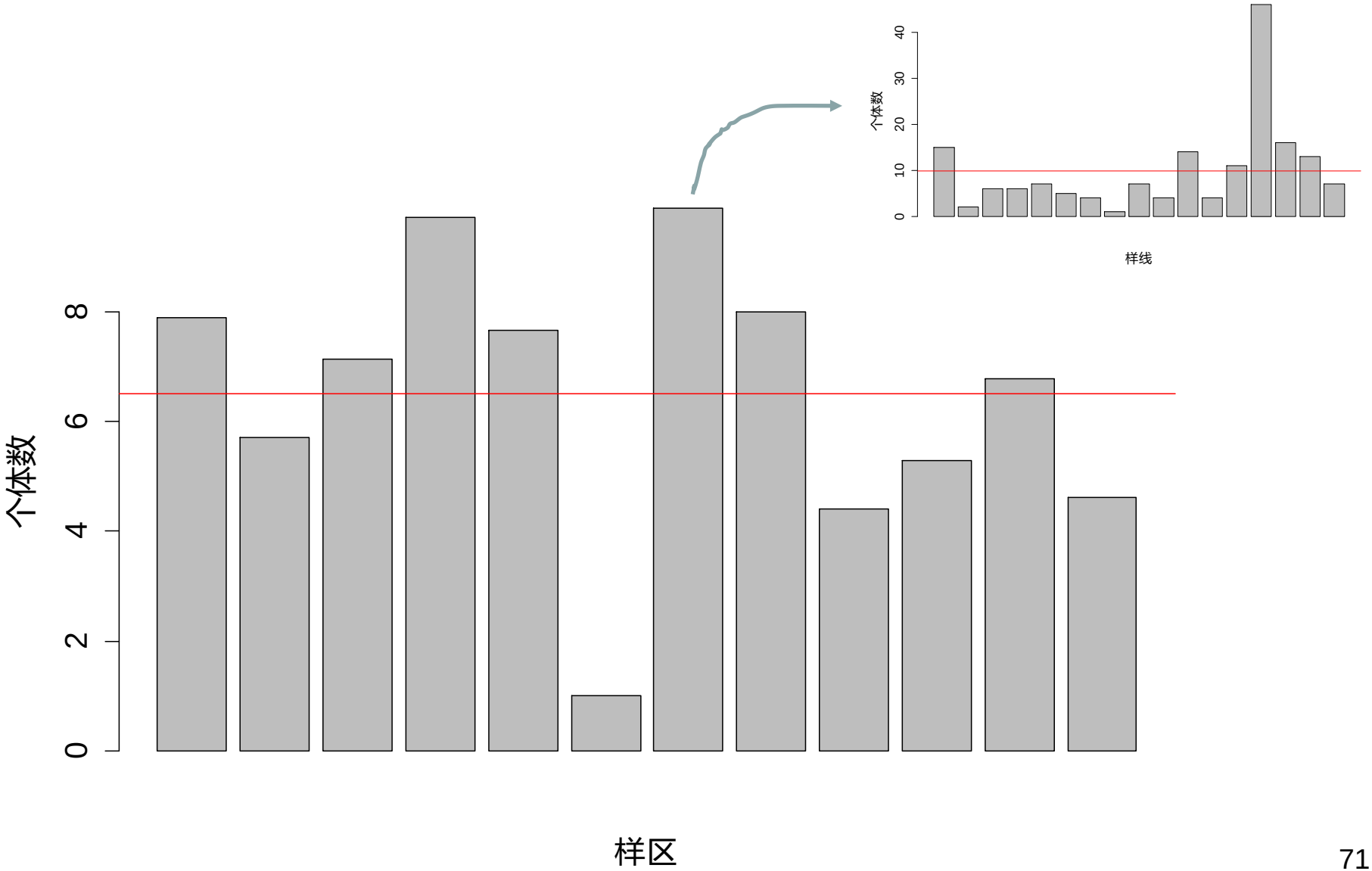




白鹭在各样区每条样线的数量



白鹭在各样区每条样线的平均数量



Conclusion regarding to wildlife survey

- All animals follow cluster distribution in space.
- Line transaction method can not be used for accurately estimate population sizes.

All models are wrong; some models are useful.

George E. P. Box, William Hunter and Stuart Hunter, *Statistics for Experimenters*, second edition, 2005, page 440.

- George Edward Pelham Box (October 18, 1919 – March 28, 2013).
- A British mathematician and professor of statistics at the University of Wisconsin.
- A pioneer in the areas of quality control, time series analysis, design of experiments and Bayesian inference.
- He was the son-in-law of Sir Ronald Fisher.

http://cen.acs.org/content/cen/articles/91/i13/Design-Experiments-Makes-Comeback/_jcr_content/articlebody/subpar/boxed_text_0/par/media_0.img.jpg/1466564026093.jpg

野生动物调查方法 Wildlife survey methods

Krebs, Charles J. 2014. Ecological Methodology. Third edition. Page 22.

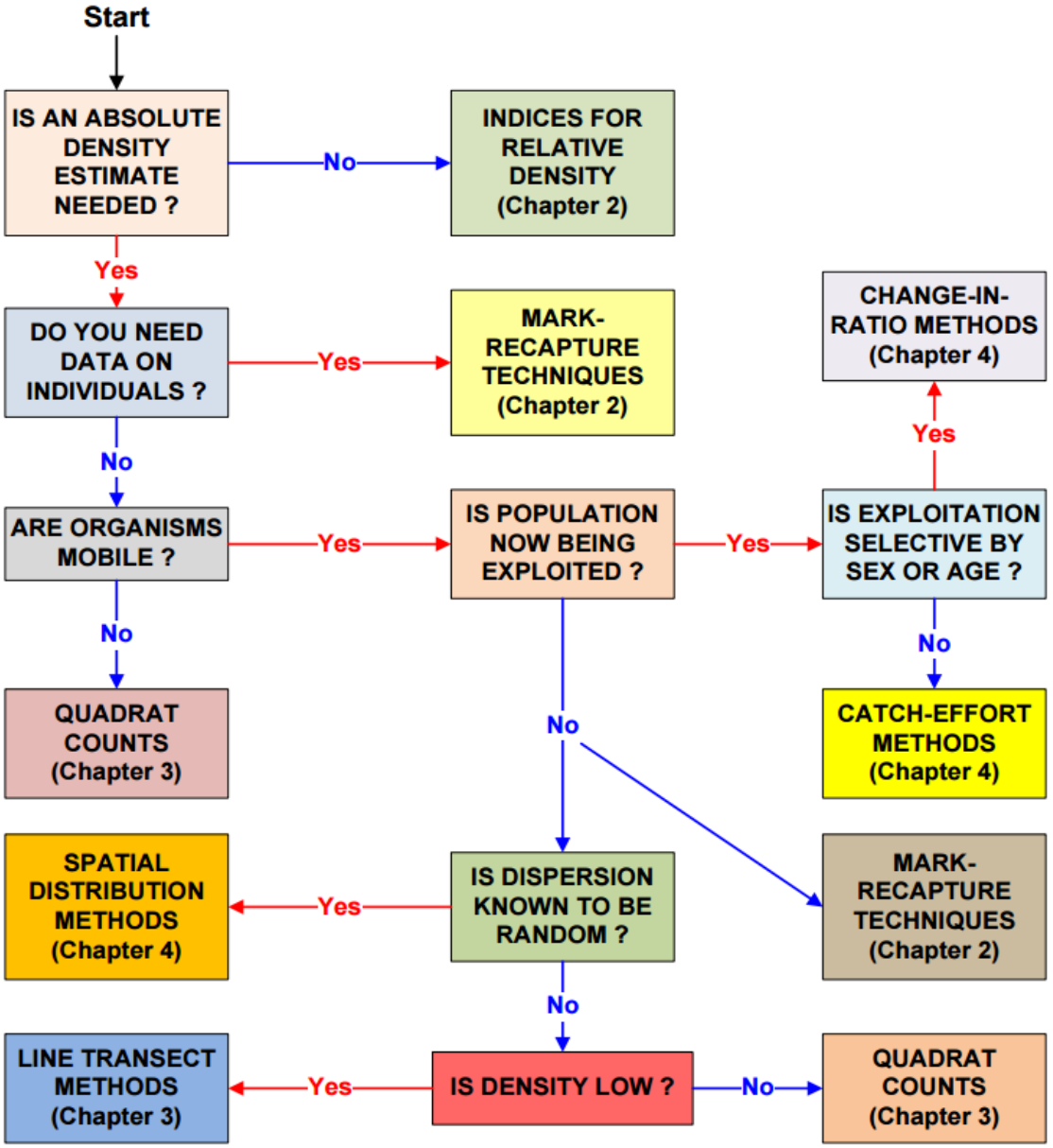


Figure A. Sequence of decisions by which a technique for estimating abundance can be chosen. (Modified from Caughley 1977.)

Assignment

- For a given population (e.g. N=1000), try simple random sampling (n=100), systematic sampling (n=100), stratified sampling (you need to make groups for this) (n=100), and cluster sampling (n=100).