Thursday, Aug 31st, 2017. Lecture 4:

Outline Error
Sampling error
Sampling distribution of the mean
Standard error (definition, calculation and interpretation)
95% confidence interval (interpreation and calculation using the 2SE rule)

Prep for next week: Read Chapter 5

Computer demos: http://phylo.bio.ku.edu/biostats/geneLenDemo.html (note: Keep the sample sizes in the 10 - 200 range for accurate graphs). A better demo (but unrelated to the gene length example given in your text) is: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

1. inference – cloud cartoon
    - **Estimation**
2. Dopamine example.
    - **Point estimate**
    - **Error**
3. Introduce example – book uses length of genes in human genome. Easiest to have a toy in which we know the population. *Cymodocea nodosa* – seagrass from the Mediterranean:
    - It is dioecious,
    - male flowers have 2 anthers
    - we'll assume 50:50 sex ratio
4. Draw population relative frequency distribution
5. Scientific question: "What is the mean # anthers per individual in the population?"
6. tiny study $N = 6$
7. 7 possible outcomes (we only know this because we know the population)
8. black bead = male flower, white bead = female
9. draw 6 beads. calculate $\bar{Y}$
10. *Thought experiment:* Would we get the same answer if we repeated the sampling? Let's repeat the entire experiment 4 more times...
11. histogram of point estimates
12. **Sampling error**
13. show large histograms for $\bar{Y}$
14. show histograms for different N
15. Larger samples $\rightarrow$ smaller sampling error
16. **Sampling distribution**
17. Infinite number of repeated experiments $\rightarrow$ let's look at the true sampling distribution for this study design and population.
18. For large N:
    - The sampling distribution of the mean is bell-shaped – Normal distribution (chapter 10)
    - The mean of the sampling distribution is μ
    - ...
19. What fun fact did we learn about the Normal distribution on Tuesday?

20. Empirical rule: $\approx 95\%$ of values from a normal will be within 2 sd of the mean.
21. If we know the standard deviation of the sampling distribution, we can say something important about our estimation: 95% will be within 2sd of the correct answer.
22. sd of the sampling distribution is so important, we give it a name: **standard error of an estimate**
23. $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$
24. In our example:

n	$\sigma_{\bar{Y}}$
6	0.408
24	0.204
96	0.102
384	0.051

25. draw cutoffs on expected histograms, to show that this works.
26. What is the main problem with using this fun fact about the standard error? We don't know σ
27. $s \approx \sigma$
28. $$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$
29. Dopamine SE
30. Deep Breath:
    - From a sample we can calculate:
      - a sample mean
      - a sample standard deviation
    - $SE_{\overline{Y}} = \frac{s}{\sqrt{n}}$
    - $SE_{\overline{Y}}$ is a good estimate of the standard error.
31. The sampling distribution is Normal$(\mu, SE_{\overline{Y}})$
32. "Ponder this" slide
33. **confidence interval**
34. Dopamine CI
35. CI interpretation
36. Error bars are not standardized... read your figure legends (show Fig 4.4-2)
37. Fargo vs Miami
38. If you estimated a 95% CI for the mean temp of Miami and Fargo, what do you expect?
39. Show the calculations
40. Bad jelly bean counter example, time permitting.

Estimation: the process of inferring a population parameter from sample data.

Point estimate: a single "best guess" of the value of a parameter.

Error: the difference between our estimate and the population parameter value: $\overline{Y} - \mu$

Sampling Error: the error that is caused by the fact that our sample is finite.

Sampling Distribution: the probability distribution of estimates that we might obtain when we sample a population.

The standard error of an estimate: the standard deviation of the estimate's sampling distribution.

Confidence interval: a range of values around our estimate. The interval is likely to contain the value of the population parameter

If we always report a 95% confidence interval, then the population parameter should be within the interval in 95% of our studies.

What is the mean dopamine concentration in the brains of rats? A sample of 7 individuals:

6.8 5.3 6.0 5.9 6.8 7.4 6.2 (nmol/g)

$$\bar{Y} = \frac{\sum_i Y_i}{n} = 6.34 \text{ nmol/g}$$

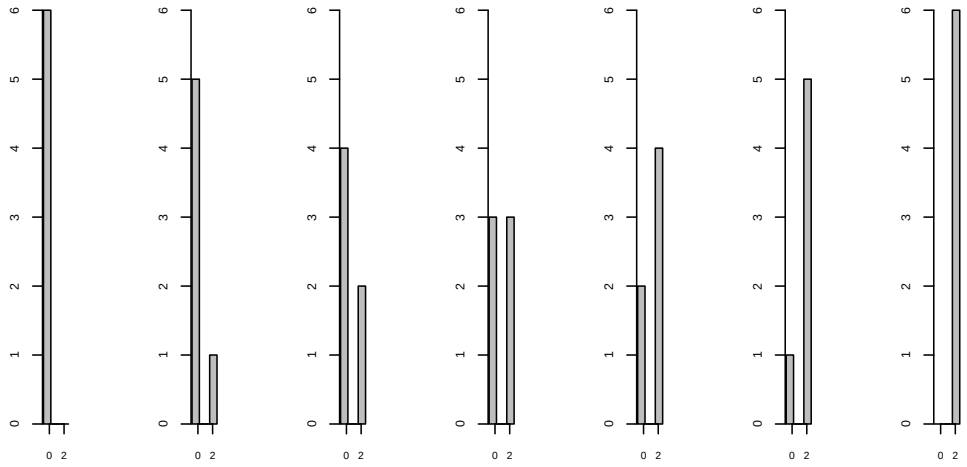
$$s = \sqrt{\frac{\sum_i \left(Y_i - \bar{Y}\right)^2}{n-1}} = .702$$

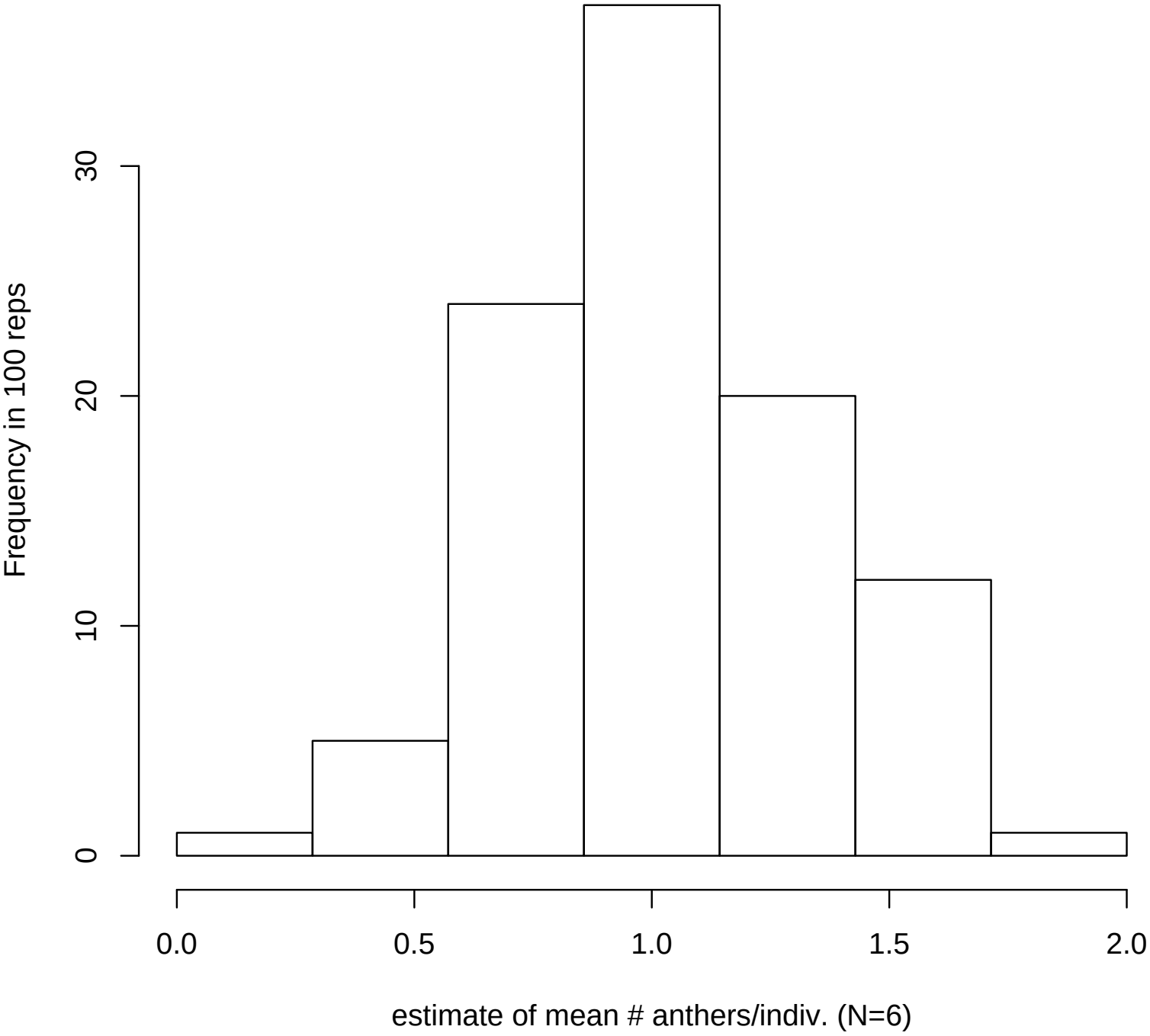
$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{0.702}{\sqrt{7}} = \frac{0.702}{2.6457} = 0.265$$

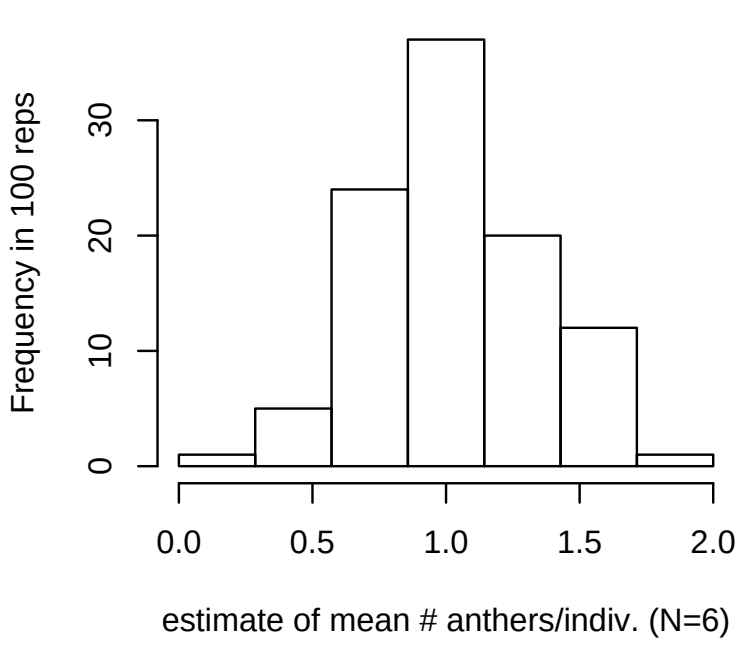
We'd report: $\bar{Y} = 6.34 \pm 0.265(SE)$ nmol/g

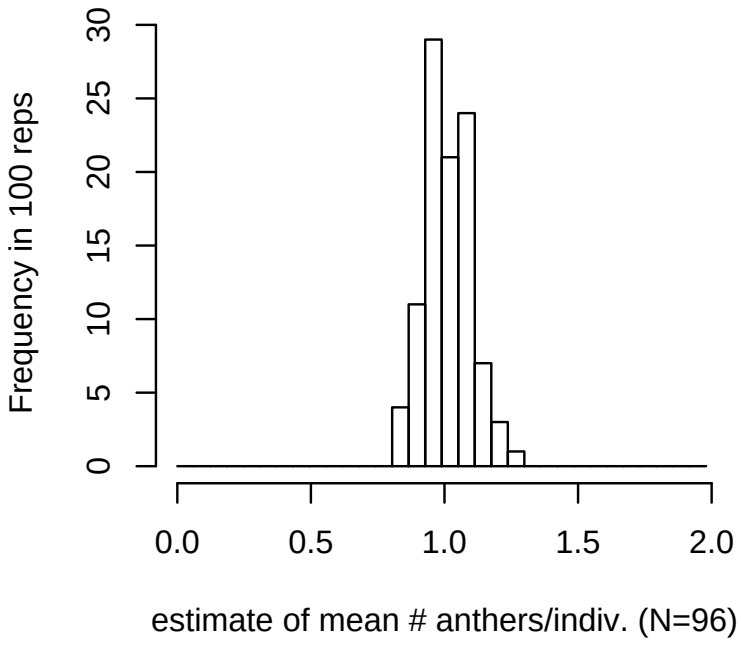
We are 95% confident that ...

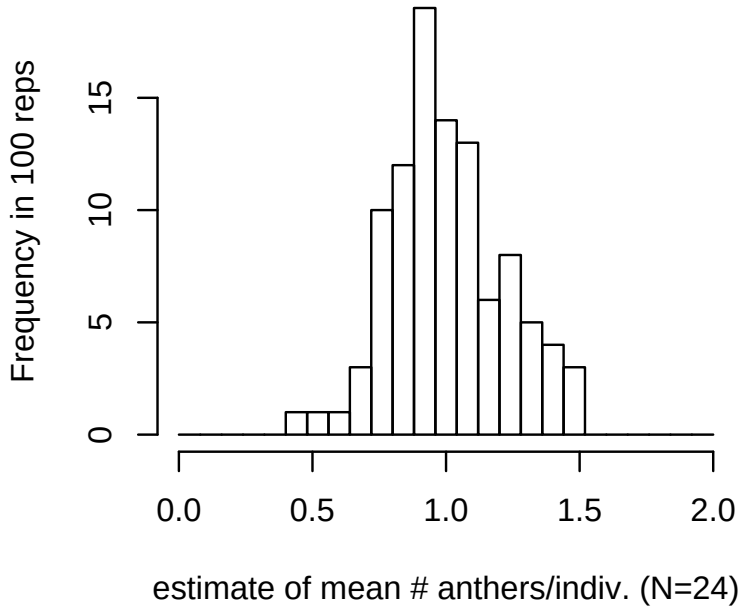
$$\begin{aligned} \overline{Y} - 2SE_{\overline{Y}} &< \mu < \overline{Y} + 2SE_{\overline{Y}} \\ 6.34 - 0.53 &< \mu < 6.34 + 0.53 \\ 5.81 \text{ nmol/g} &< \mu < 6.87 \text{ nmol/g} \end{aligned}$$

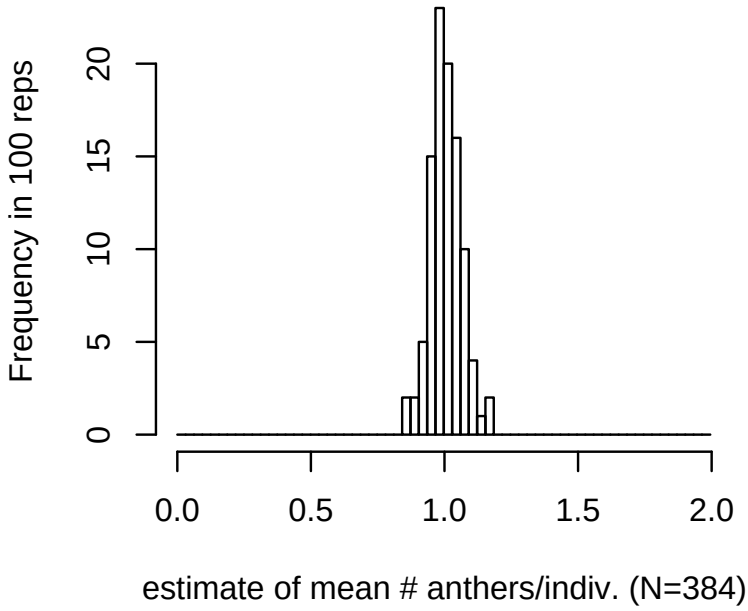

0
1
2
3
4
5
6
0
2

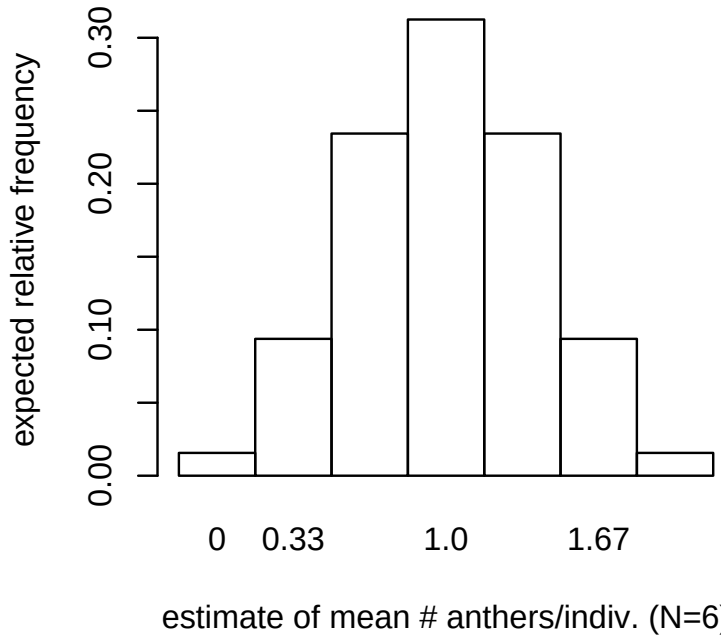

Frequency in 100 reps
0
10
20
30
0.0
0.5
1.0
1.5
2.0
estimate of mean # anthers/indiv. (N=6)

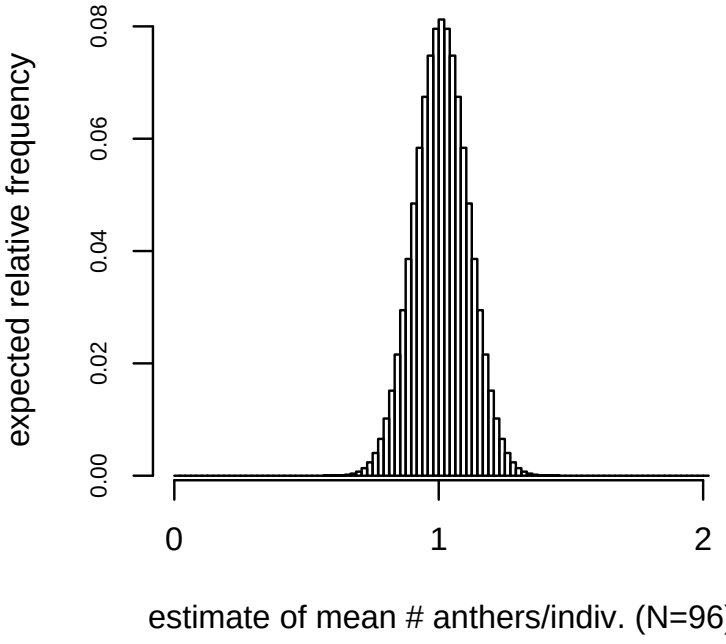

Frequency in 100 reps
0
10
20
30
0.0
0.5
1.0
1.5
2.0
estimate of mean # anthers/indiv. (N=6)

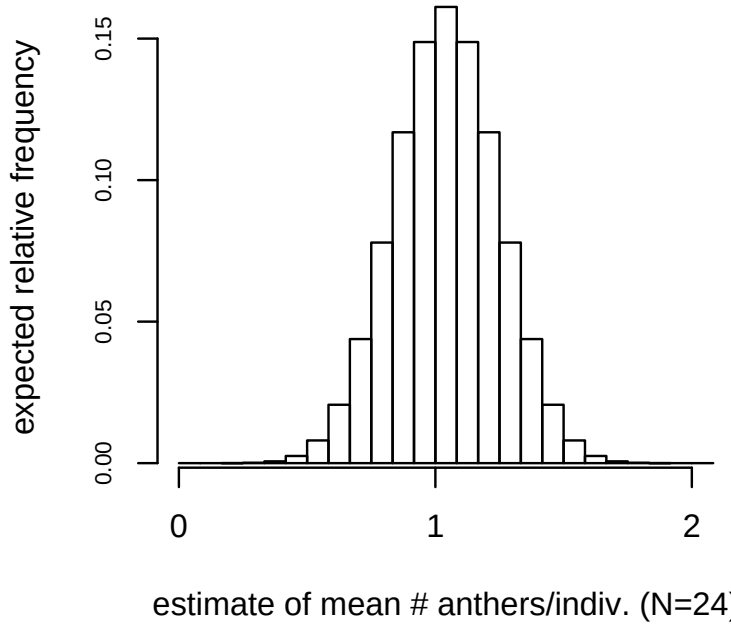

Frequency in 100 reps
0
5
10
15
20
25
30
0.0
0.5
1.0
1.5
2.0
estimate of mean # anthers/indiv. (N=96)

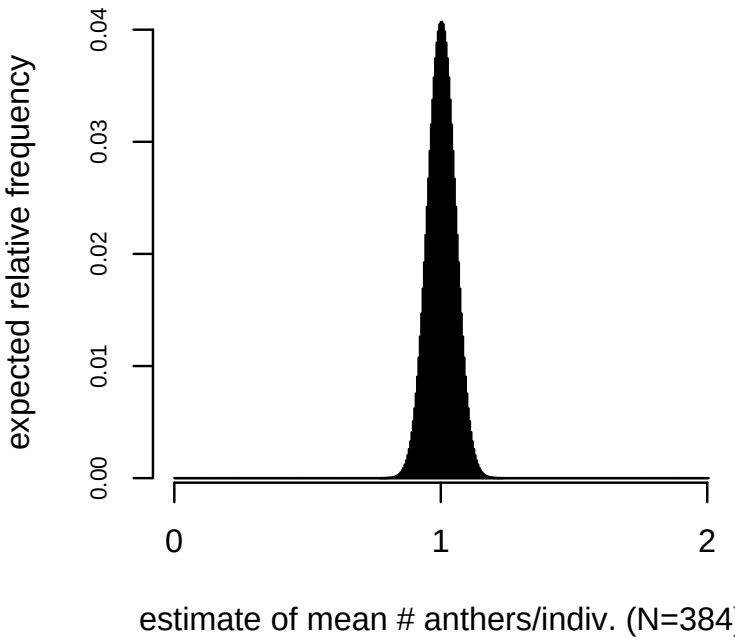

Frequency in 100 reps
0
5
10
15
0.0
0.5
1.0
1.5
2.0
estimate of mean # anthers/indiv. (N=24)


Frequency in 100 reps
0
5
10
15
20
0.0
0.5
1.0
1.5
2.0
estimate of mean # anthers/indiv. (N=384)


expected relative frequency
0.00
0.10
0.20
0.30
0
0.33
1.0
1.67
estimate of mean # anthers/indiv. (N=6)


expected relative frequency
0.00
0.02
0.04
0.06
0.08
0
1
2
estimate of mean # anthers/indiv. (N=96)


expected relative frequency
0.00
0.05
0.10
0.15
0
1
2
estimate of mean # anthers/indiv. (N=24)


expected relative frequency
0.00
0.01
0.02
0.03
0.04
0
1
2
estimate of mean # anthers/indiv. (N=384)

Fun facts about the sampling distribution of $\bar{Y}$

(1) bell-shaped. A Normal distribution.

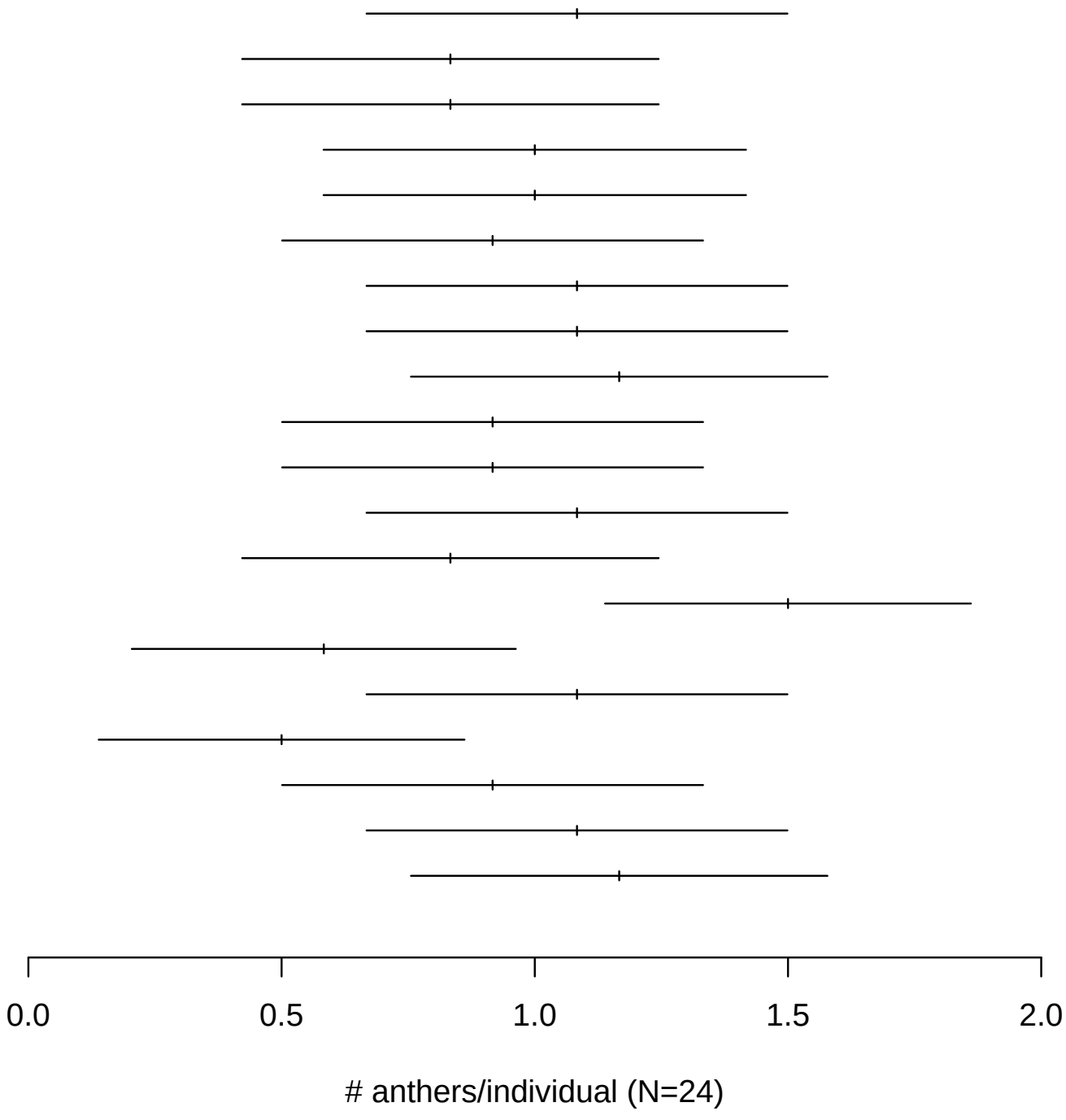
(2) mean is μ

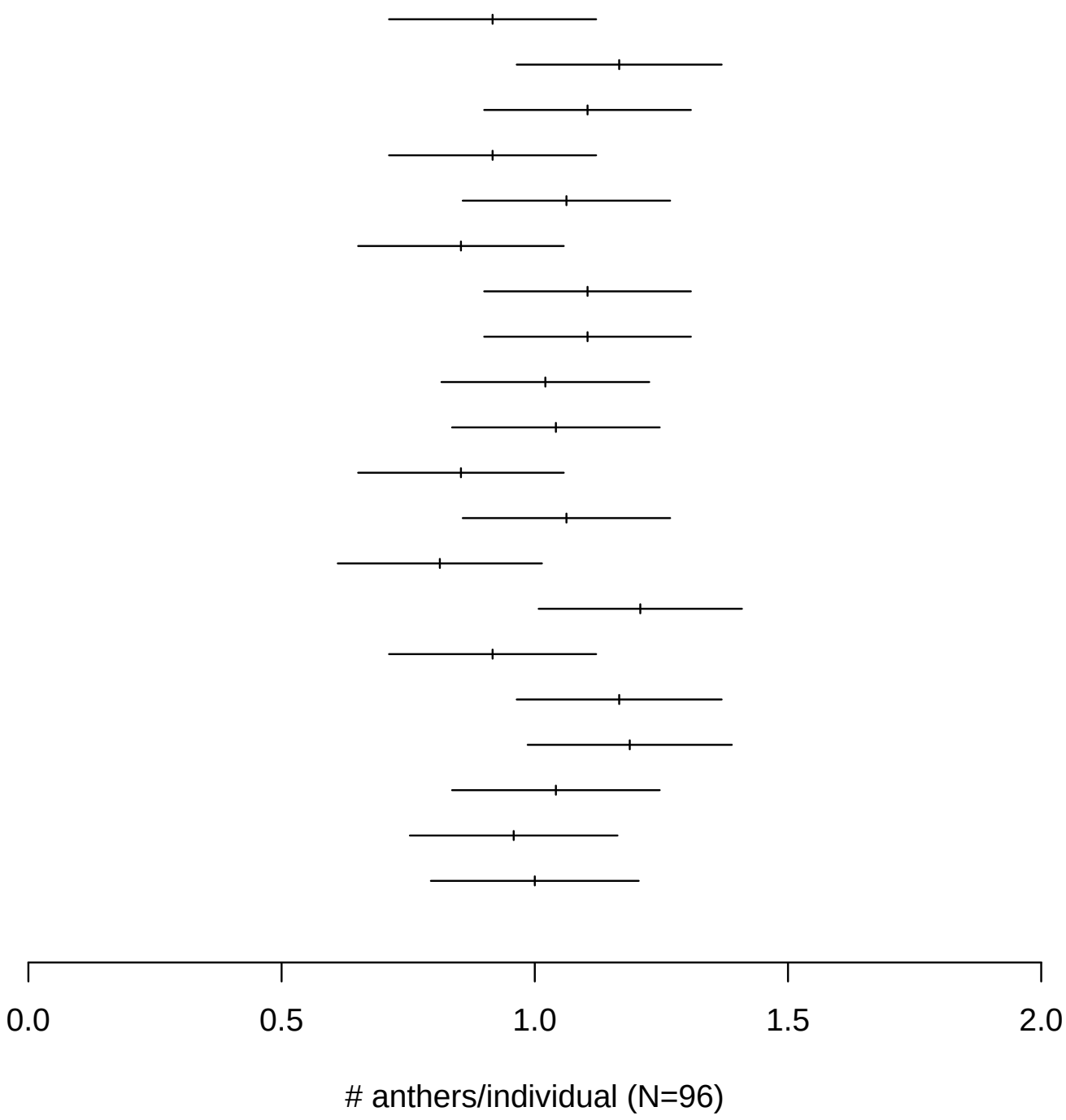
Empirical rule: Typically, 95% of our observations are within 2SD of the mean

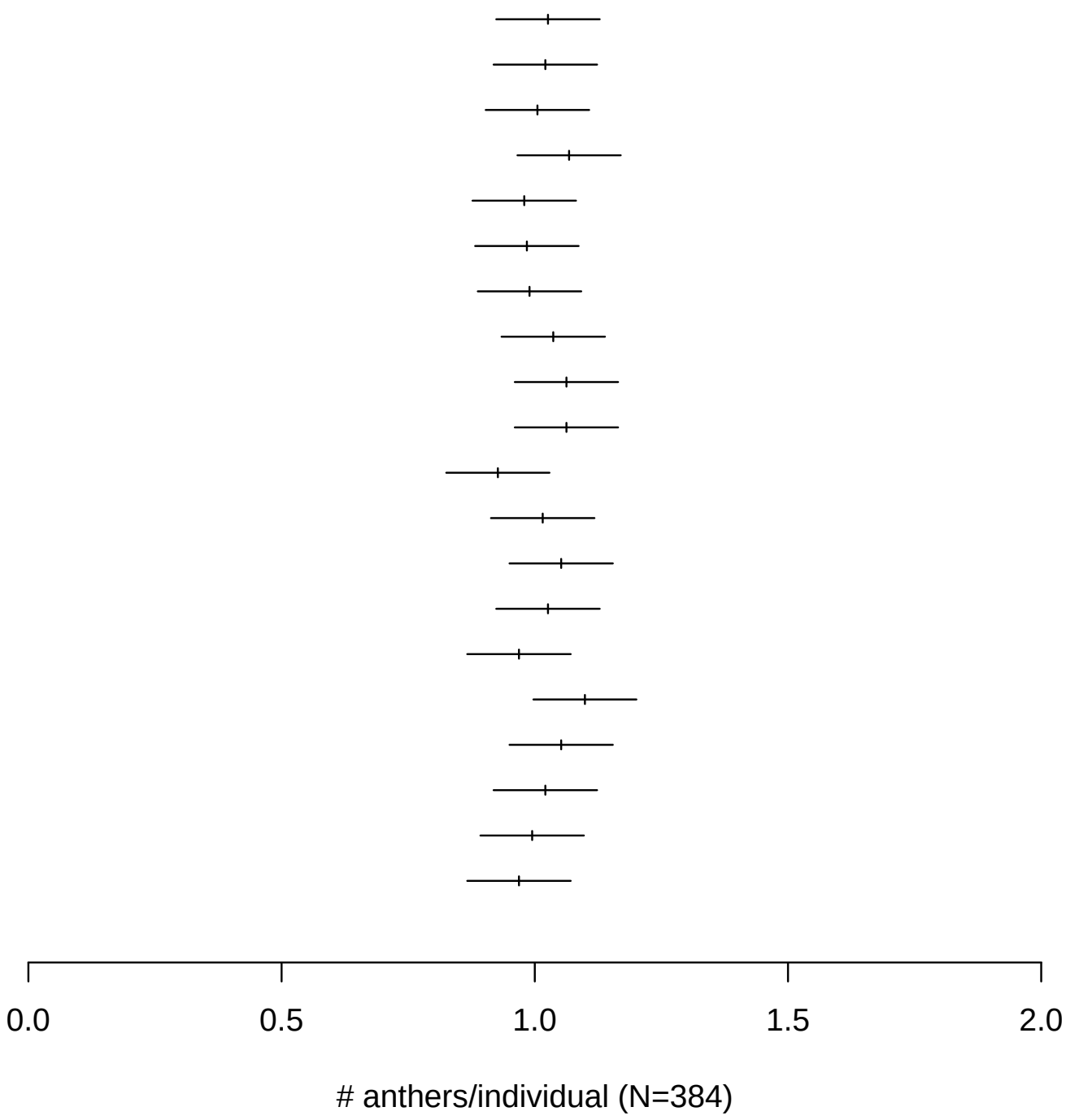
Ponder this:

Approximately 95% of the $\bar{Y}$ estimates will be within $2\text{SE}_{\bar{Y}}$ of μ.

So, if we create an interval to include points within $2\text{SE}_{\bar{Y}}$ of $\bar{Y}$, then approximately 95% of our intervals should contain μ.


0.0
0.5
1.0
1.5
2.0
anthers/individual (N=24)


0.0
0.5
1.0
1.5
2.0
anthers/individual (N=96)


0.0
0.5
1.0
1.5
2.0
anthers/individual (N=384)

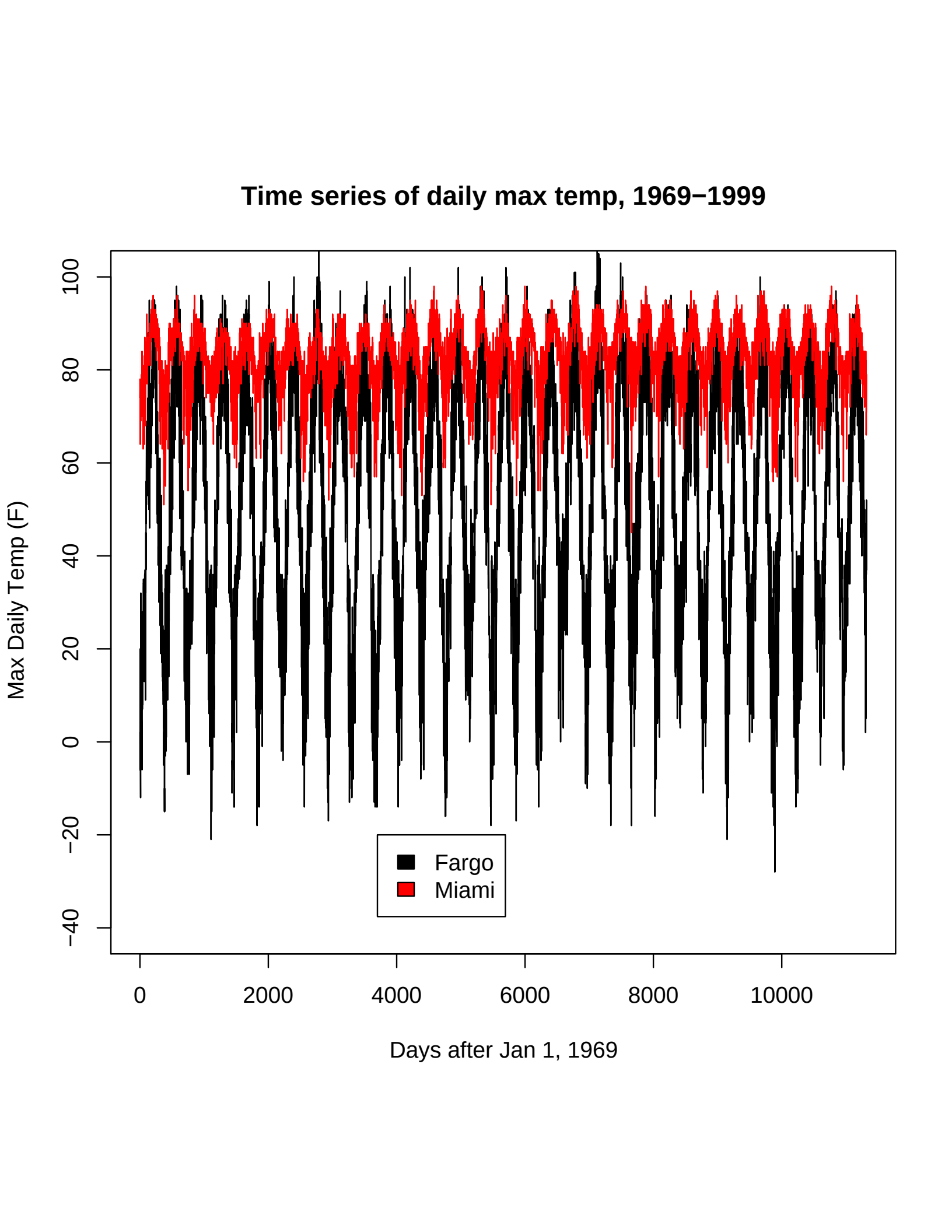
Two groups of researchers want to estimate the mean length of a gene in the human genome.

Group 1 uses $n = 10$

Group 2 uses $n = 100$

They both will report 95% Confidence Intervals. Which statement is true:

1. Group 1's CI will be more likely to contain μ
2. Group 2's CI will be more likely to contain μ
3. The CI's for both groups are equally likely to contain μ


Time series of daily max temp, 1969−1999
Max Daily Temp (F)
100
80
60
40
20
0
−20
−40
0
2000
4000
6000
8000
10000
Days after Jan 1, 1969
Fargo
Miami

Y_i	$(Y_i - \overline{Y})^2$		Y_i	$(Y_i - \overline{Y})^2$
35	410.06		83	0.88
81	663.06		87	9.38
69	189.06		73	119.63
6	2425.56		87	9.38
84	826.56		80	15.50
47	68.06		89	25.63
89	1139.06		84	0.00
82	715.56		76	63.00
70	217.56		81	8.63
72	280.56		90	36.75
13	1785.06		92	65.00
36	370.56		73	119.63
16	1540.56		88	16.50
59	14.06		83	0.88
47	68.06		84	0.00
78	517.56		93	82.13
$\overline{Y}$	55.25		$\overline{Y}$	83.94
s^2	748.73		s^2	38.20
s	27.36		s	6.18
$SE_{\overline{Y}}$	6.84		$SE_{\overline{Y}}$	1.55
$\overline{Y} - 2SE_{\overline{Y}}$	41.57		$\overline{Y} - 2SE_{\overline{Y}}$	80.85
$\overline{Y} + 2SE_{\overline{Y}}$	68.93		$\overline{Y} + 2SE_{\overline{Y}}$	87.03

Temperature data from 16 random samples of maximum daily temperatures in Fargo (first table and left column in the second table) and Miami (right columns in the second table).

The example shows the steps that are required to calculate an approximate 95%-confidence interval using the $2\text{SE}_{\bar{Y}}$ rule.

Note that the smaller variation in temperatures in Miami leads to noticeably narrower confidence intervals.