Survey Research and Design

How and Why Does Sampling Work?

William Marble

September 14, 2023

So Far...

- High-level overview of survey process
- Data cleaning and management
- But how do surveys actually *work*?

Sampling Process

- Survey sampling: actually recruiting people to take the survey
- (At least) three steps:
  1. define the population
  2. define the sampling frame
  3. draw the sample according to the sampling rule
- We'll talk about different sampling strategies, why some work and some don't, and basic weighting

Population to Sampling Frame to Sample

Population is the universe of units we're interested in

- Defined by the research question
- *Idealized* idea rather than something we can directly observe
- There's a "true value" in the population that we would like to measure

Population to Sampling Frame to Sample

Population is the universe of units we're interested in

- Defined by the research question
- *Idealized* idea rather than something we can directly observe
- There's a "true value" in the population that we would like to measure

Sampling frame is the set of units that could actually be included in the survey

- Try to pick a sampling frame that overlaps perfectly with the population
- There's a "true value" in the sampling frame

Population to Sampling Frame to Sample

Population is the universe of units we're interested in

- Defined by the research question
- *Idealized* idea rather than something we can directly observe
- There's a "true value" in the population that we would like to measure

Sampling frame is the set of units that could actually be included in the survey

- Try to pick a sampling frame that overlaps perfectly with the population
- There's a "true value" in the sampling frame

Sample is the set of units that actually end up in your dataset

- Use some decision rule to selects units from the sampling frame
- In ideal world, selected randomly from the sampling frame
- Use the sample to calculate a *statistic* that is an *estimate* of the true value we target

Idea of Random Sampling

- Sampling means taking a small portion of the whole
- In order for this to work, need the sample to be similar to the population
- Ensured by random sampling
- Analogy: tasting soup

Types of Sampling Strategies

Simple random sampling (SRS)

- Every unit in the sampling frame has an equal probability of selection

Types of Sampling Strategies

Simple random sampling (SRS)

- Every unit in the sampling frame has an equal probability of selection

Cluster sampling

- Put units into groups and sample entire groups

Types of Sampling Strategies

Simple random sampling (SRS)

- Every unit in the sampling frame has an equal probability of selection

Cluster sampling

- Put units into groups and sample entire groups

Stratified sampling

- Put units into groups, then sample a set number from within each group

Types of Sampling Strategies

Simple random sampling (SRS)

- Every unit in the sampling frame has an equal probability of selection

Cluster sampling

- Put units into groups and sample entire groups

Stratified sampling

- Put units into groups, then sample a set number from within each group

The key for each of these to be successful is randomization.

How Does Sampling Work?

The Key Intuition

We want the sample to be very similar to population as a whole in order to produce good estimates

People in our survey should be almost identical to those not in our survey

One way to guarantee this is to sample respondents randomly

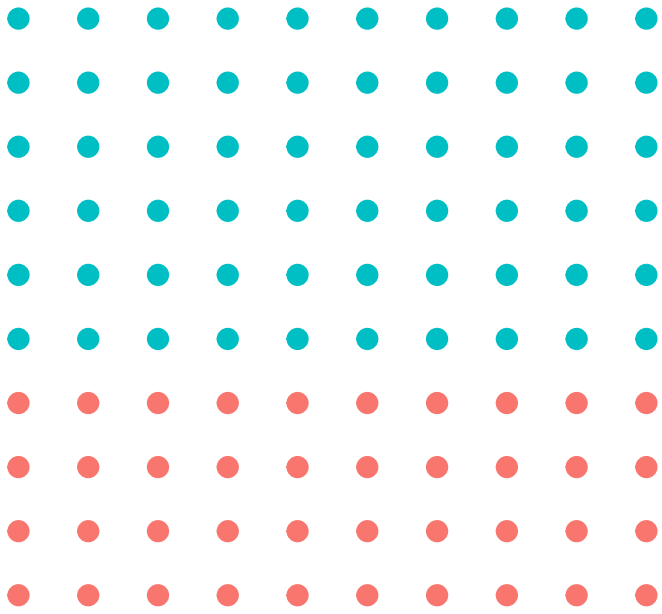
Then respondents should be very similar on every dimension to non-respondents

They won't be *exactly* the same in any given sample — but on average they will be

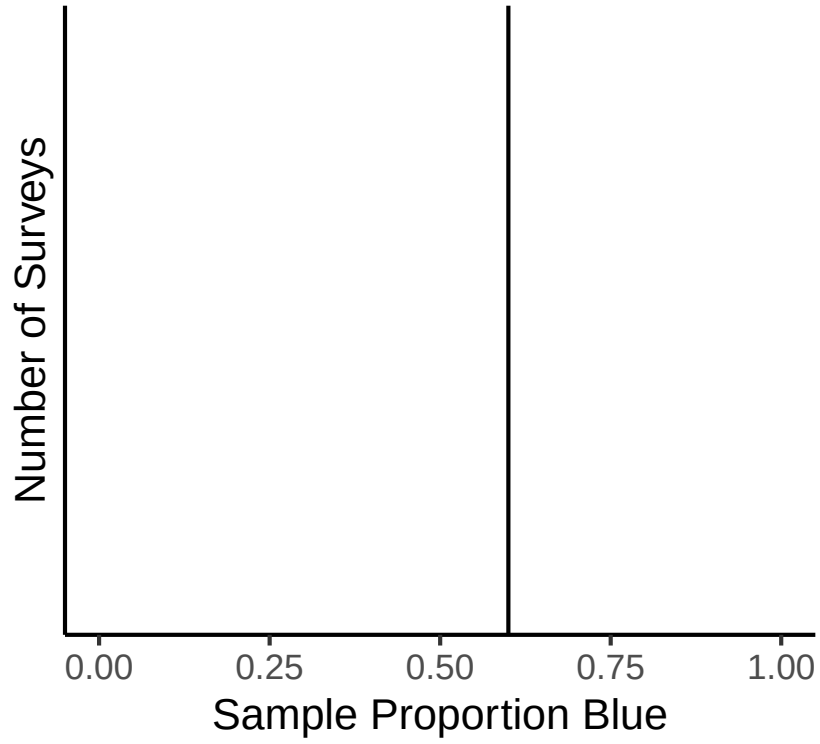
Simulation of Sampling

- Start with a population of 100: there are 60 blue and 40 red
- Simulate surveys:
  1. Randomly select 25 people to survey
  2. Record their color
  3. Calculate the proportion of the sample that is red/blue. This is the *statistic*
- We'll do this many times to simulate *repeated sampling* or many surveys

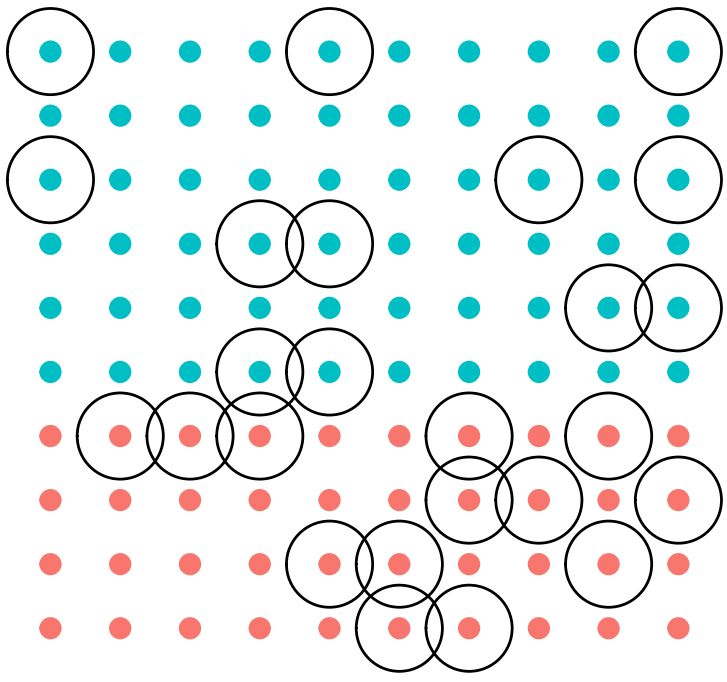
Sample Proportion Blue =



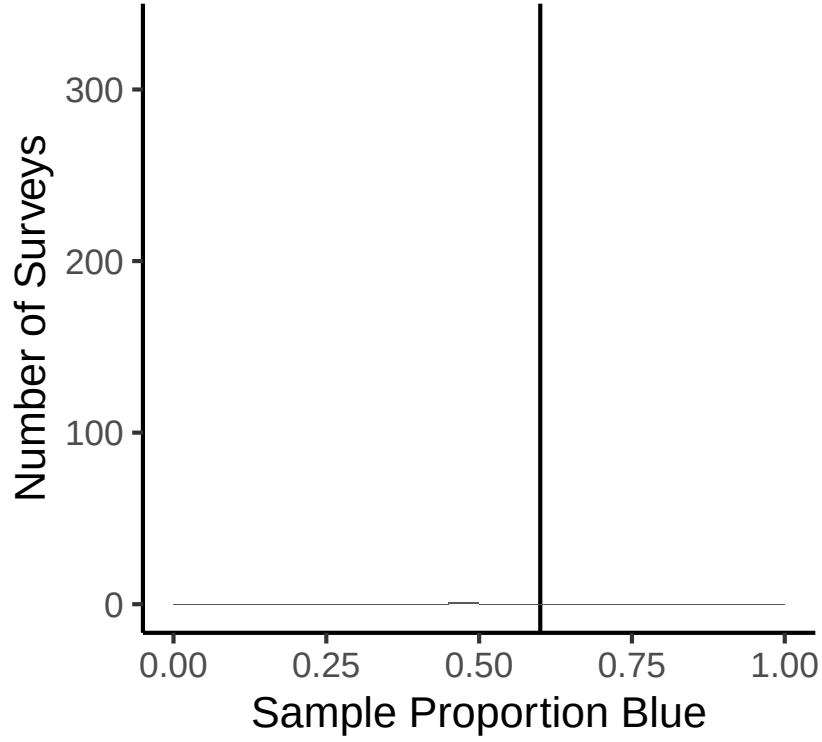
# of Surveys Run = 0



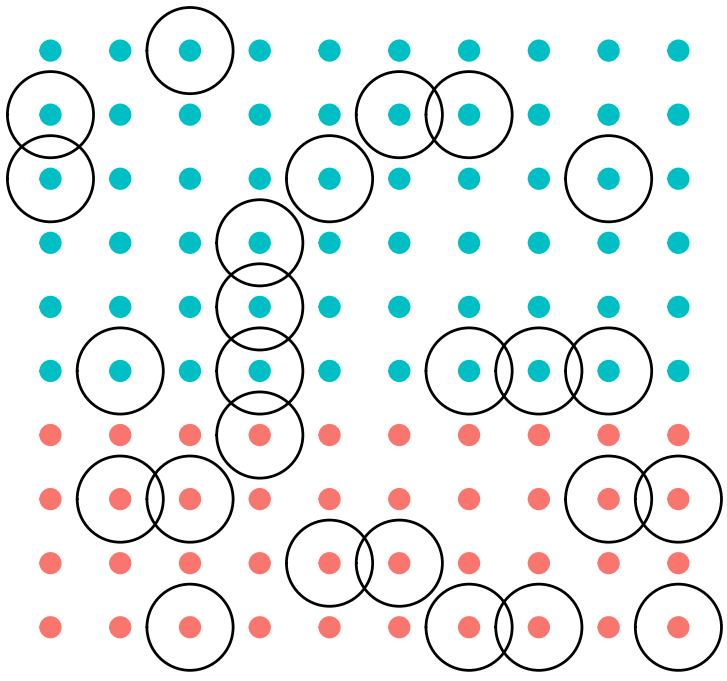
Sample Proportion Blue = 0.48



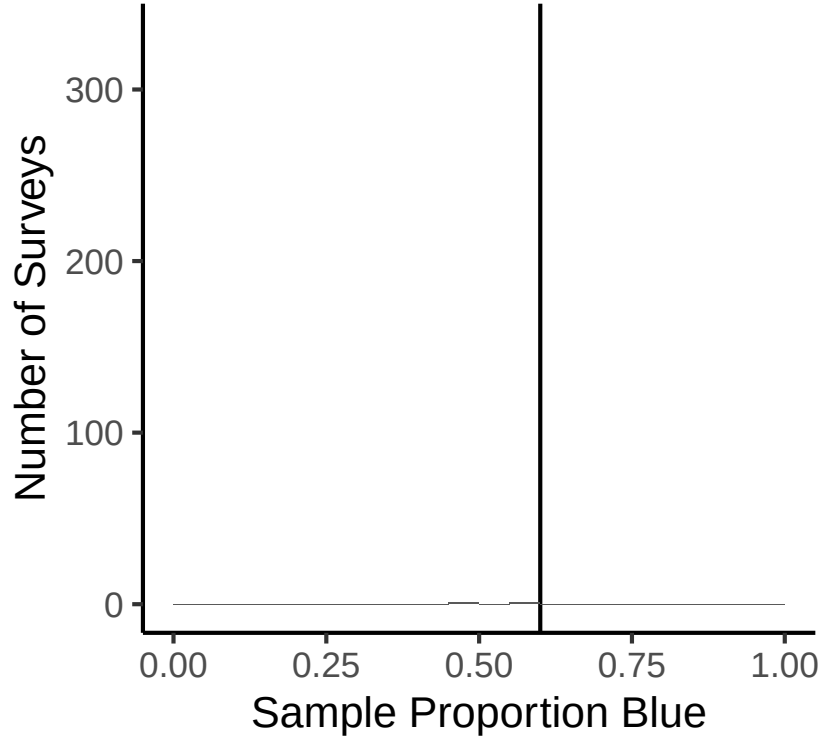
# of Surveys Run = 1



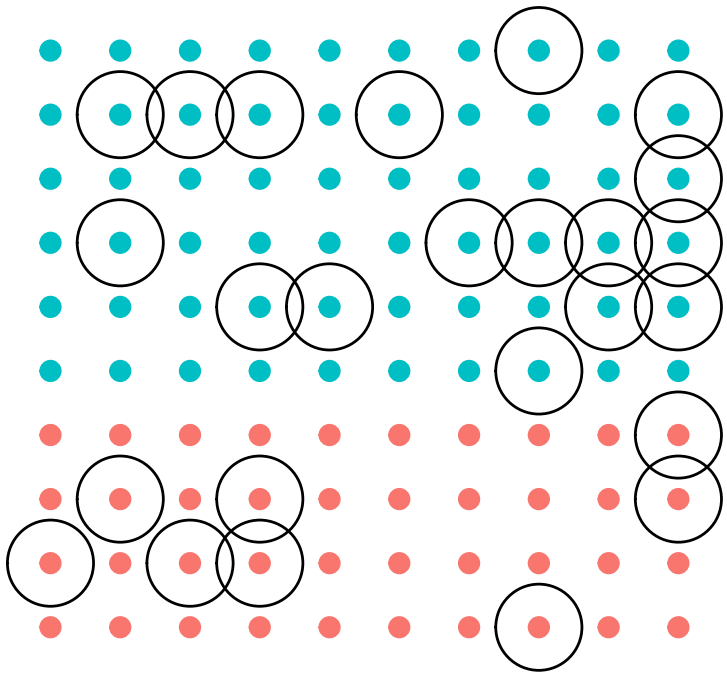
Sample Proportion Blue = 0.56



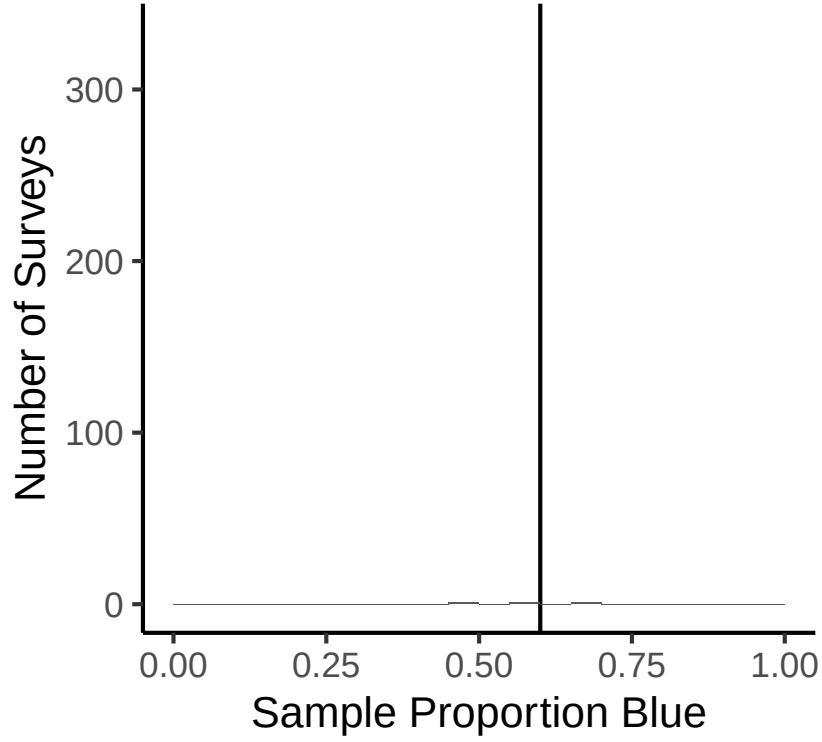
# of Surveys Run = 2



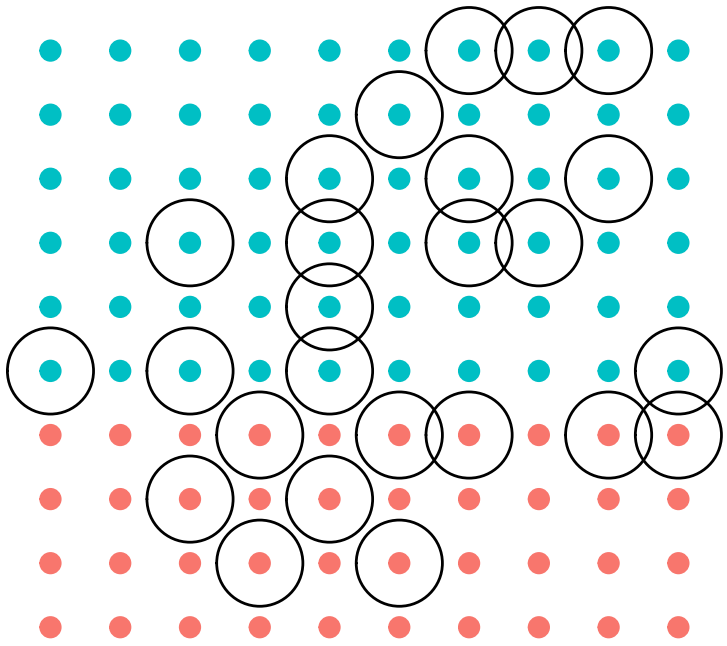
Sample Proportion Blue = 0.68



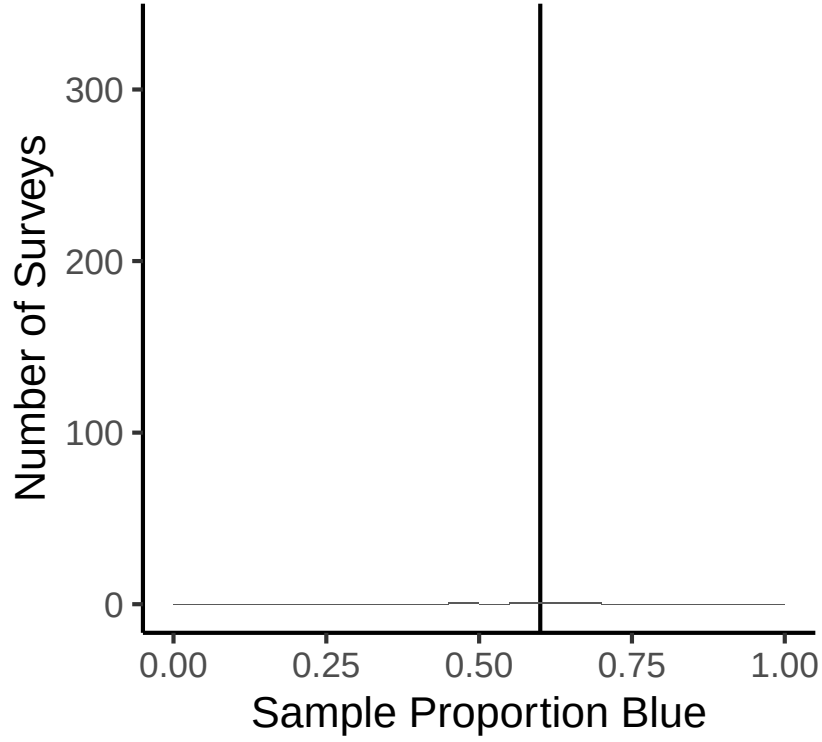
# of Surveys Run = 3



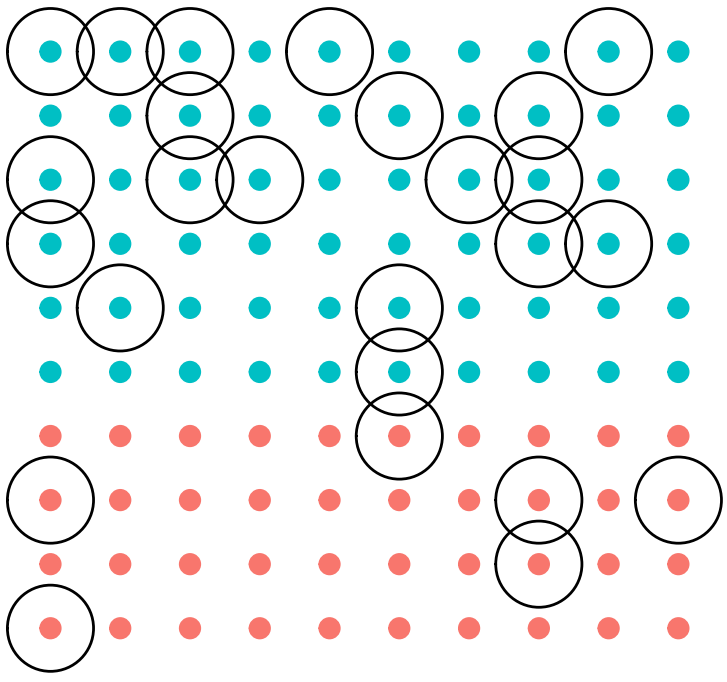
Sample Proportion Blue = 0.64



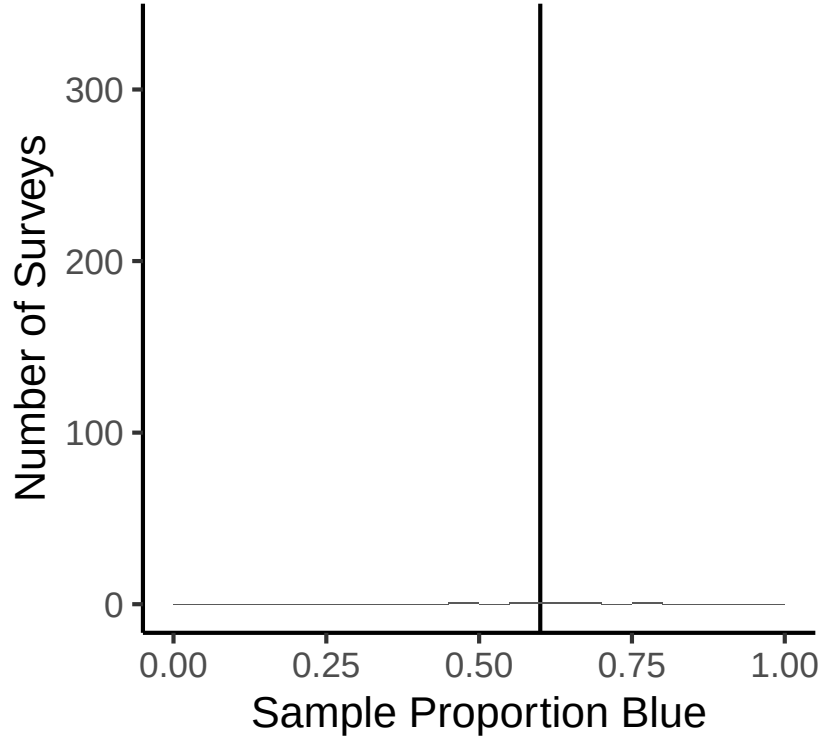
# of Surveys Run = 4



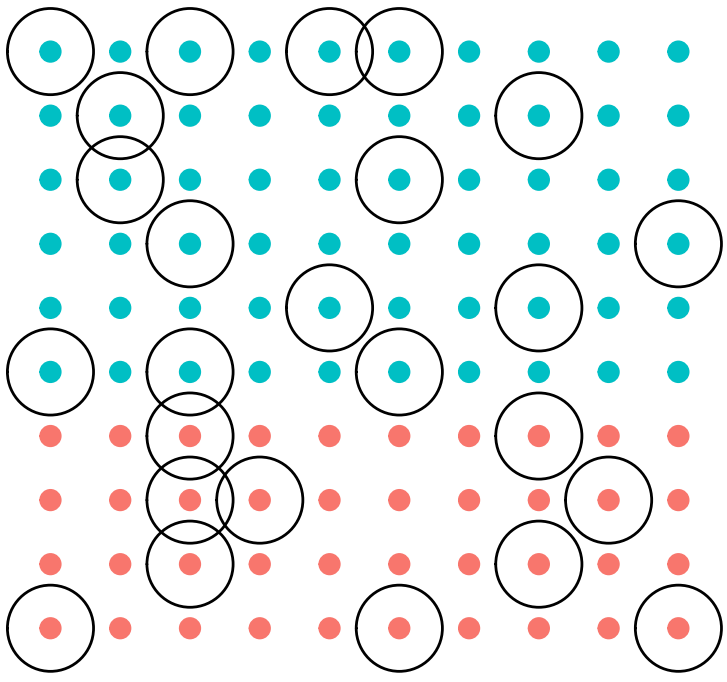
Sample Proportion Blue = 0.76



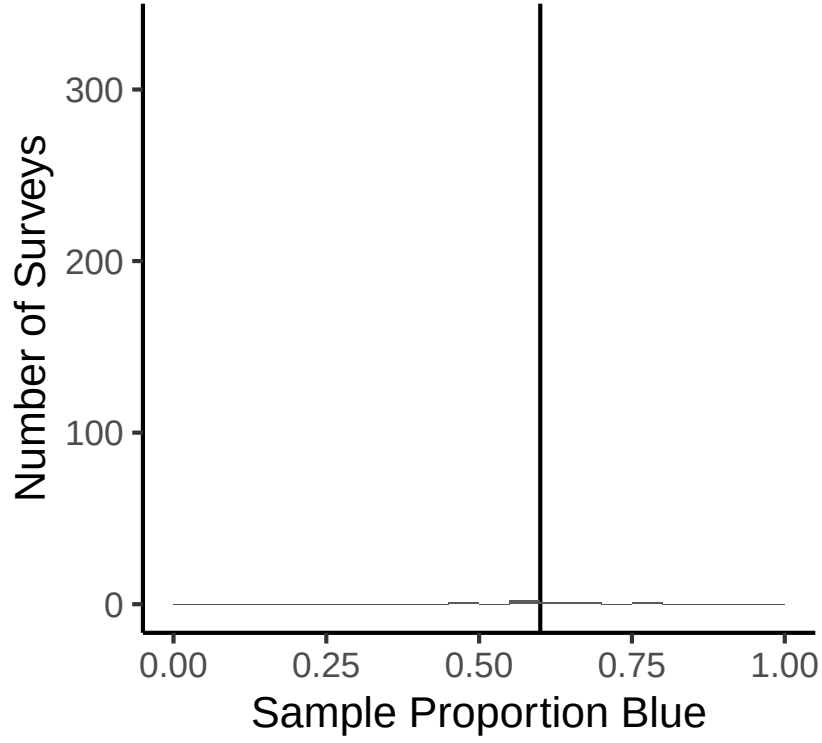
# of Surveys Run = 5



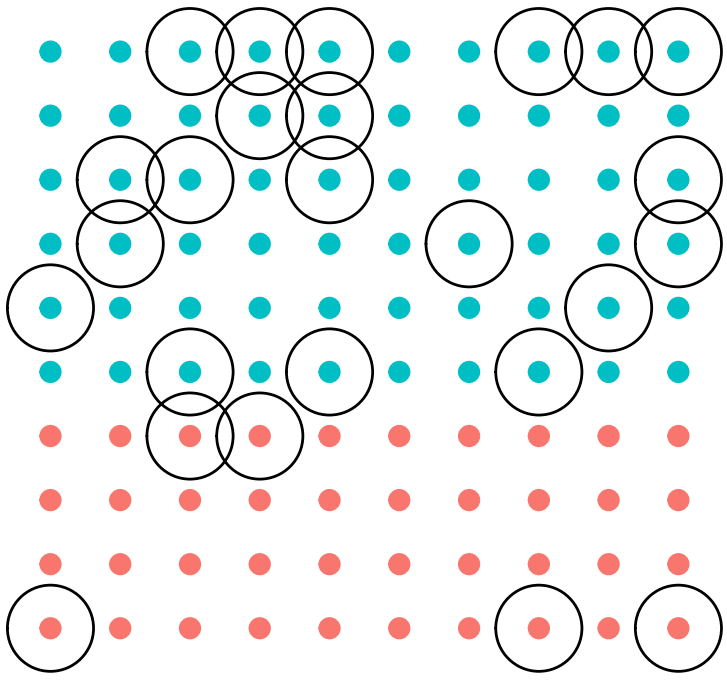
Sample Proportion Blue = 0.6



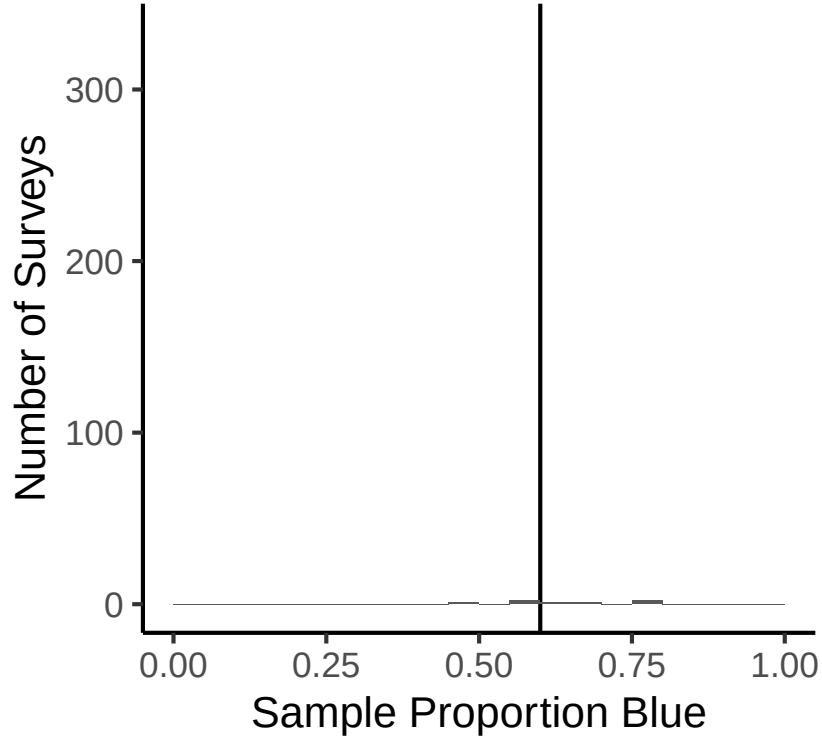
# of Surveys Run = 6



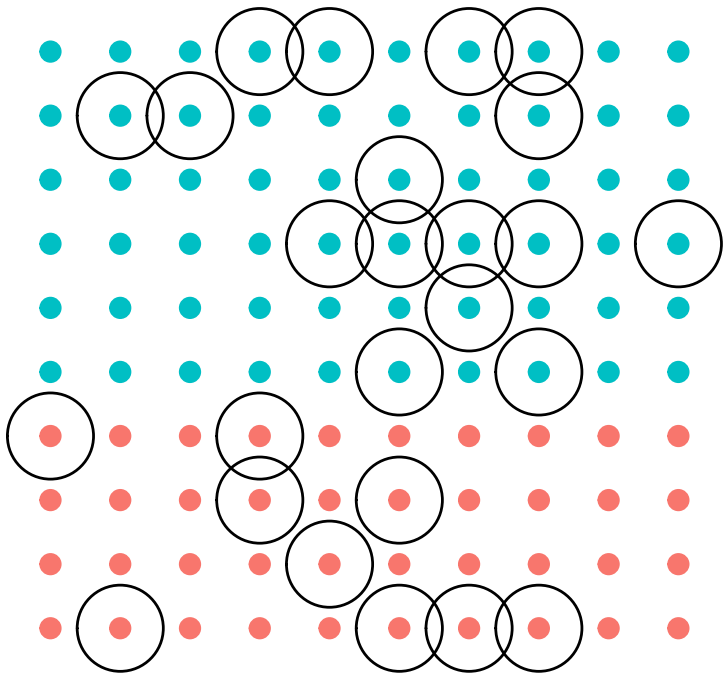
Sample Proportion Blue = 0.8



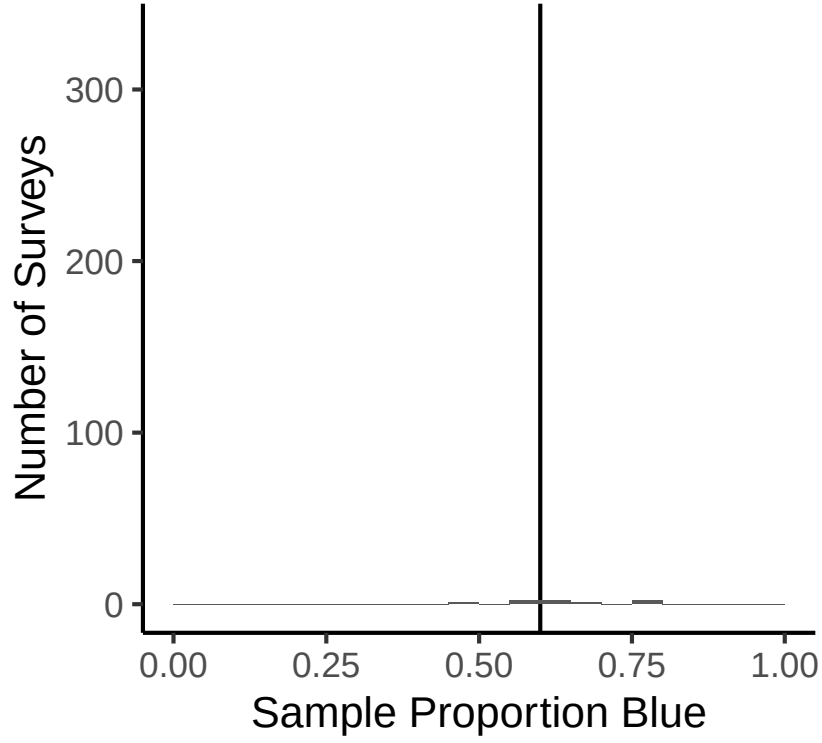
# of Surveys Run = 7



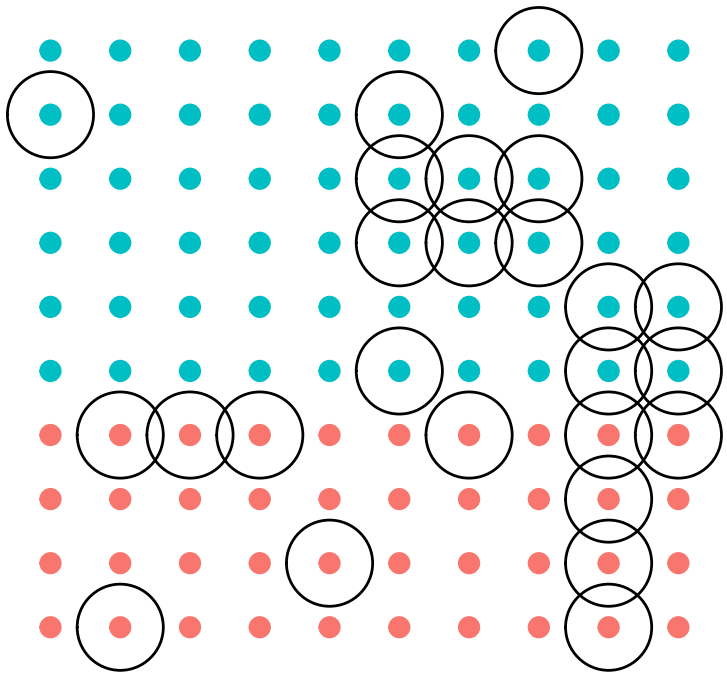
Sample Proportion Blue = 0.64



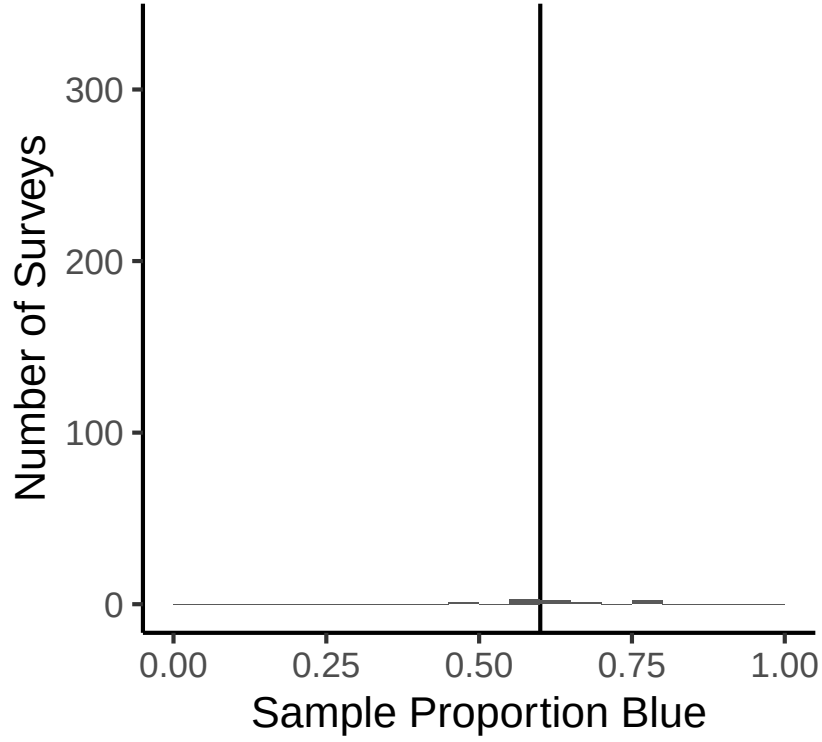
# of Surveys Run = 8



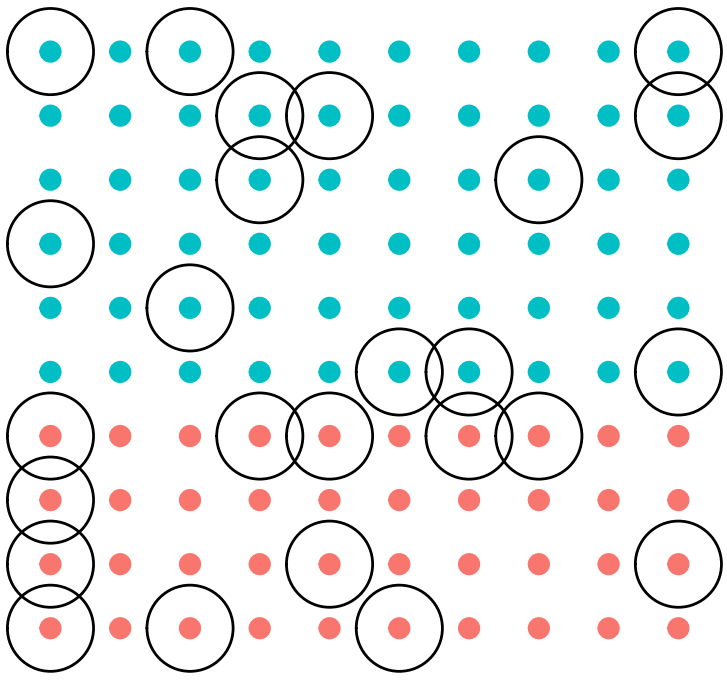
Sample Proportion Blue = 0.56



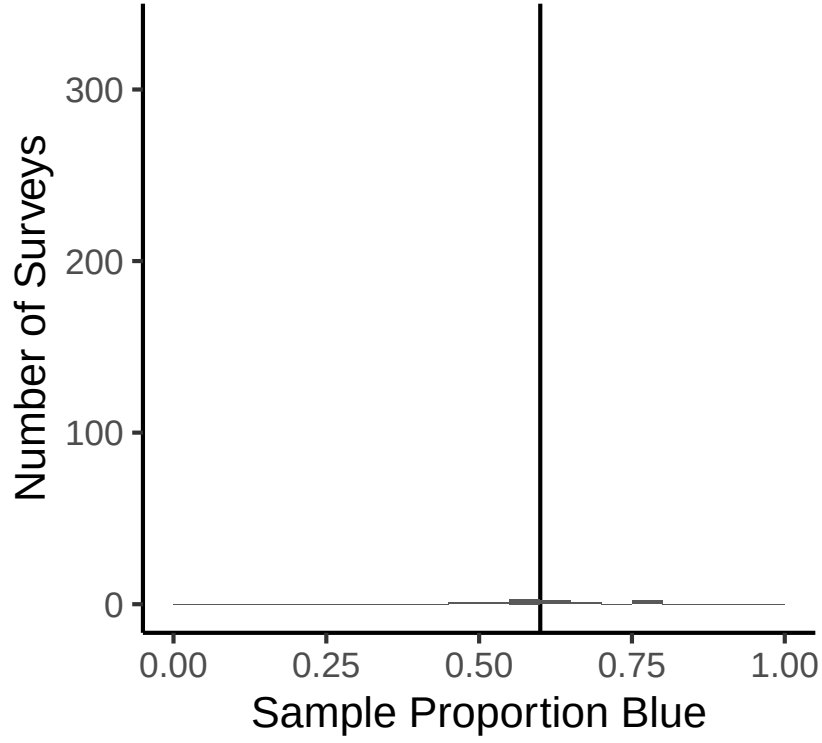
# of Surveys Run = 9



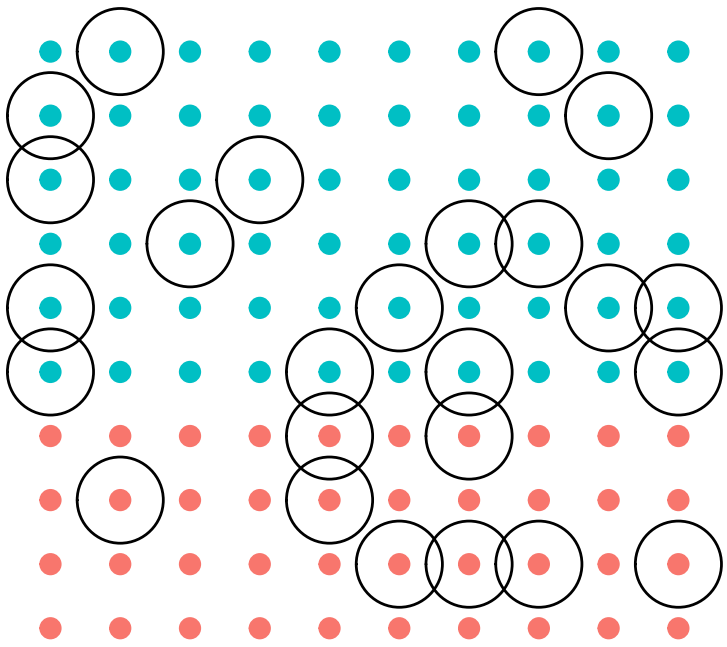
Sample Proportion Blue = 0.52



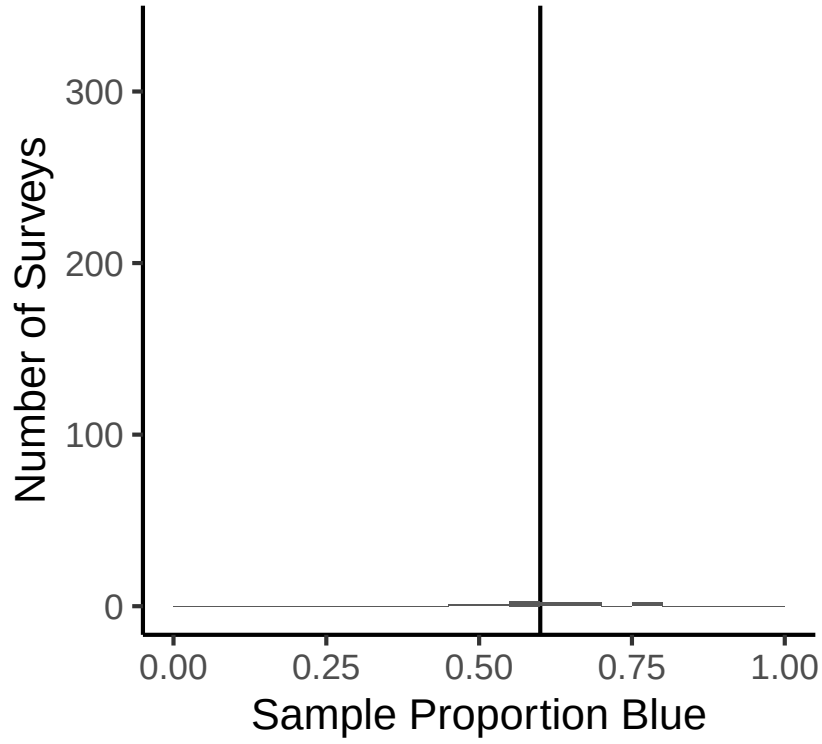
# of Surveys Run = 10



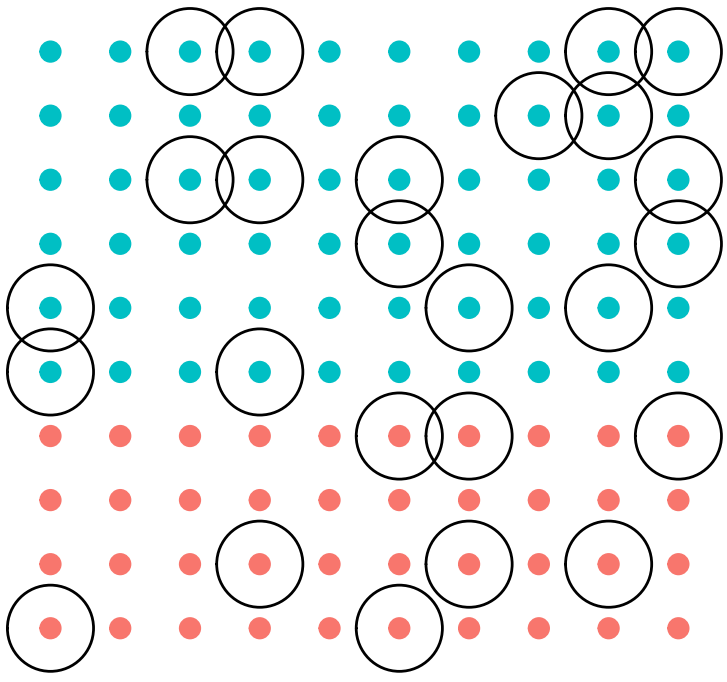
Sample Proportion Blue = 0.68



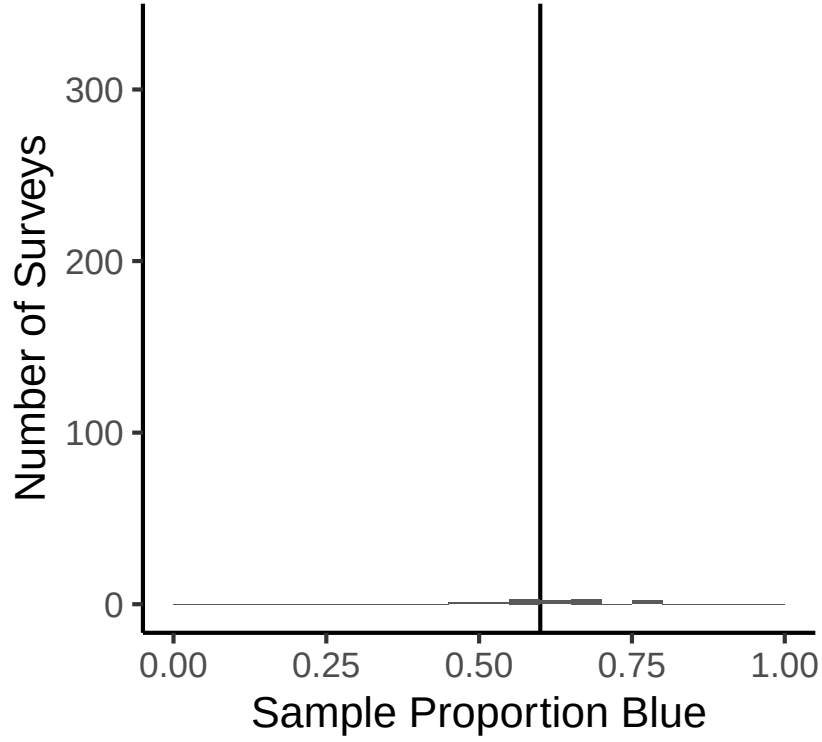
# of Surveys Run = 11



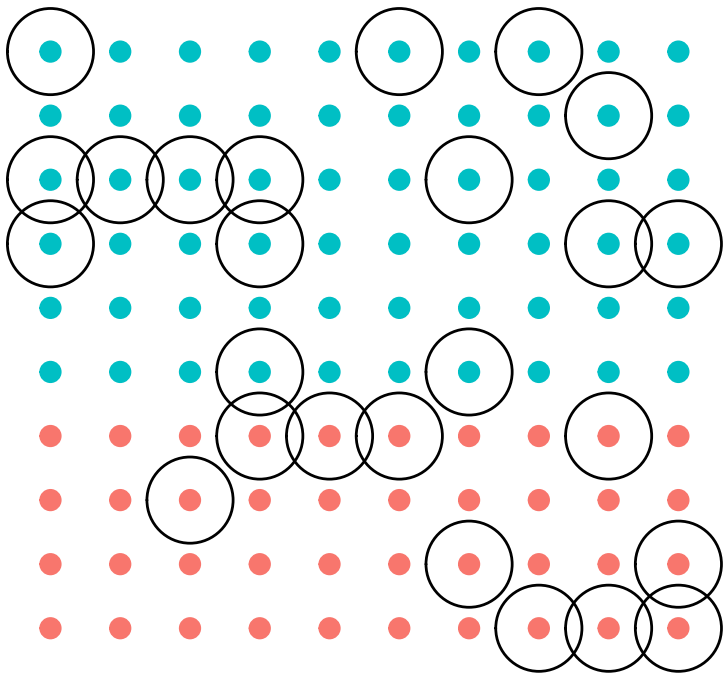
Sample Proportion Blue = 0.68



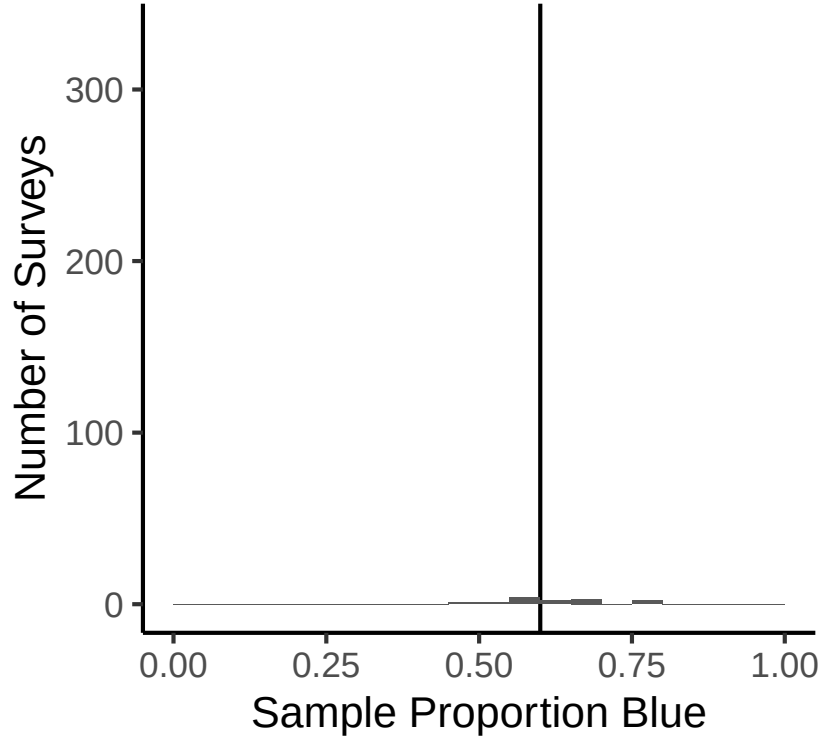
# of Surveys Run = 12



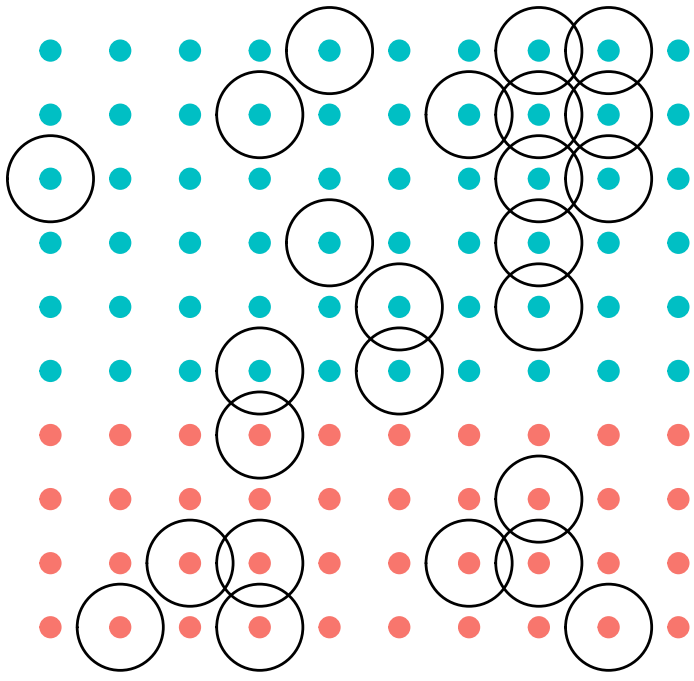
Sample Proportion Blue = 0.6



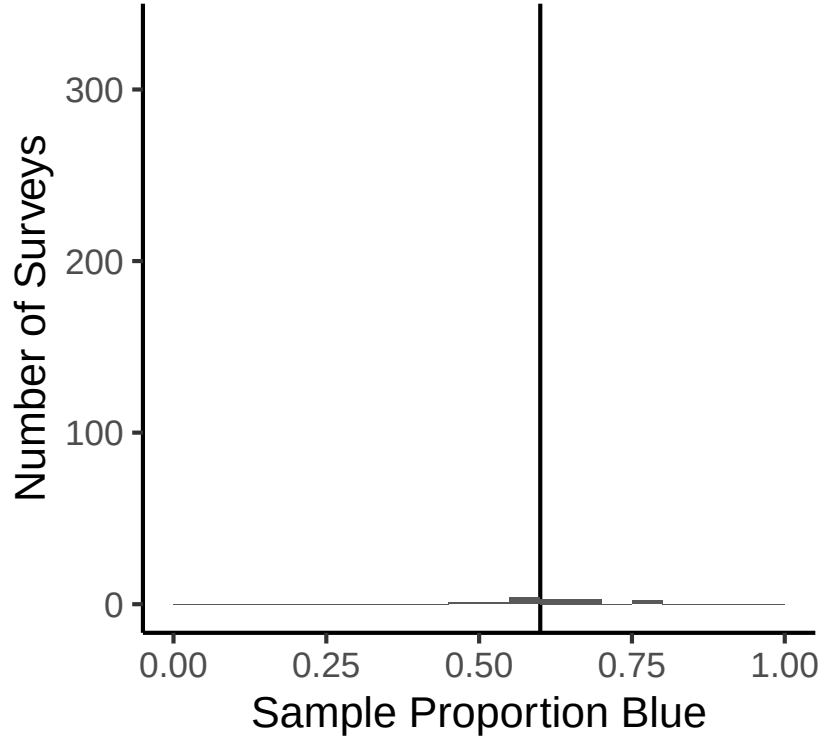
# of Surveys Run = 13



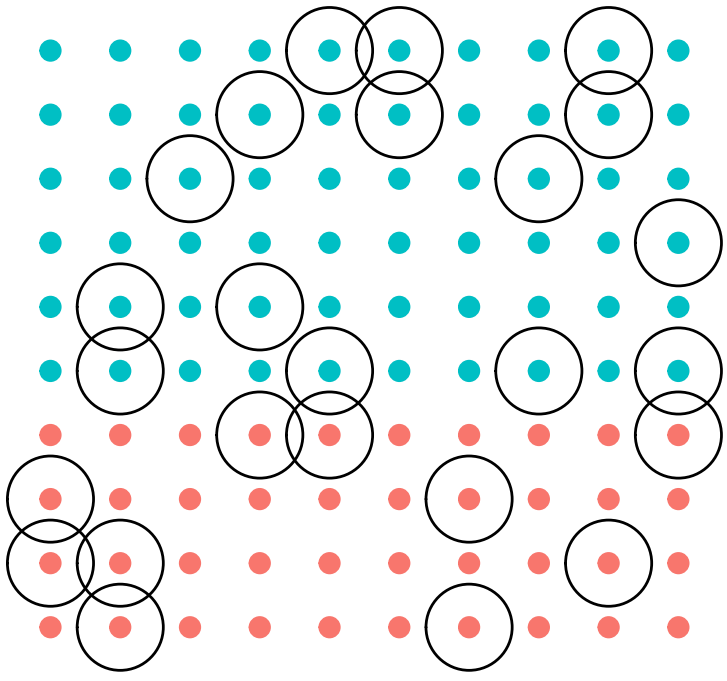
Sample Proportion Blue = 0.64



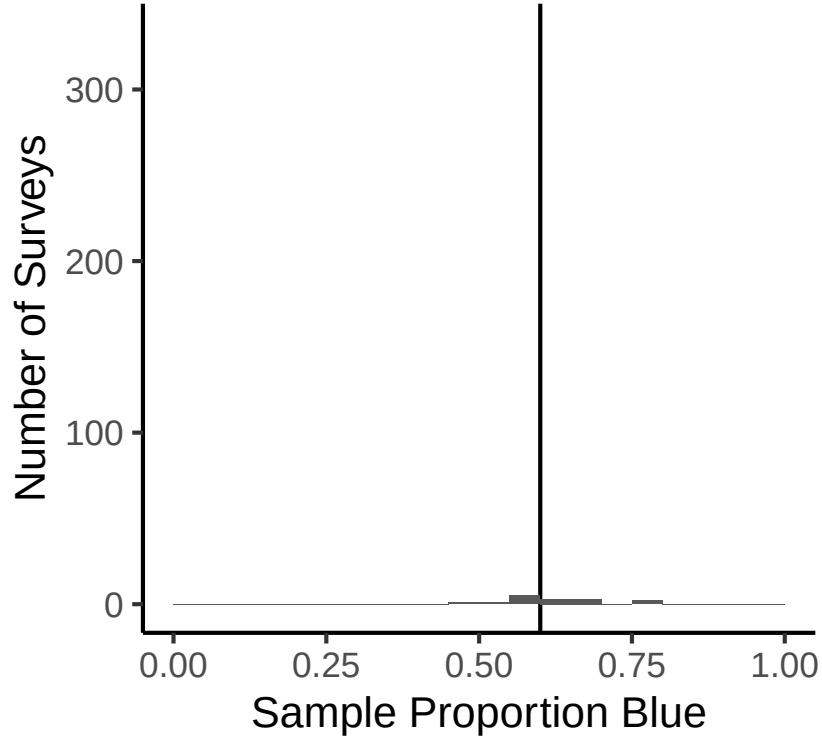
# of Surveys Run = 14



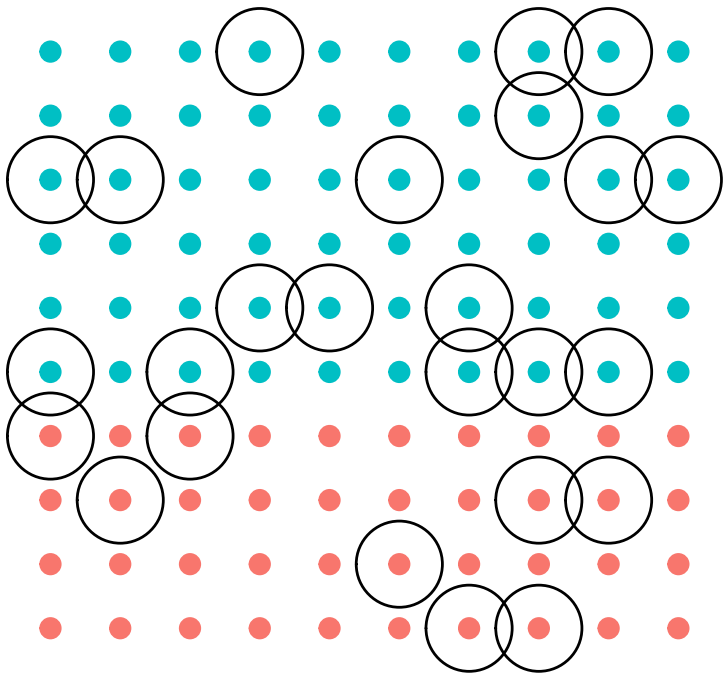
Sample Proportion Blue = 0.6



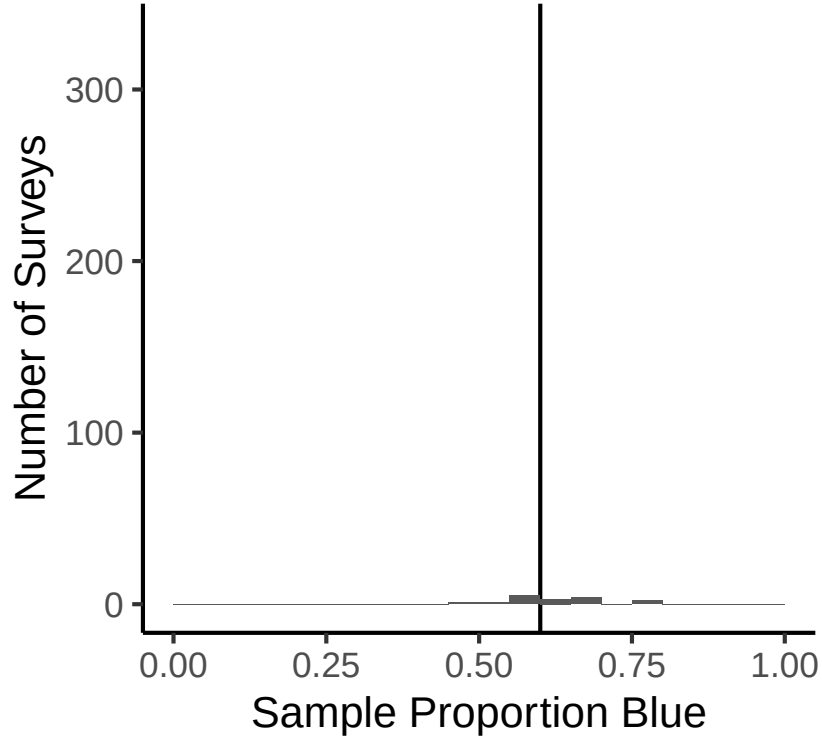
# of Surveys Run = 15



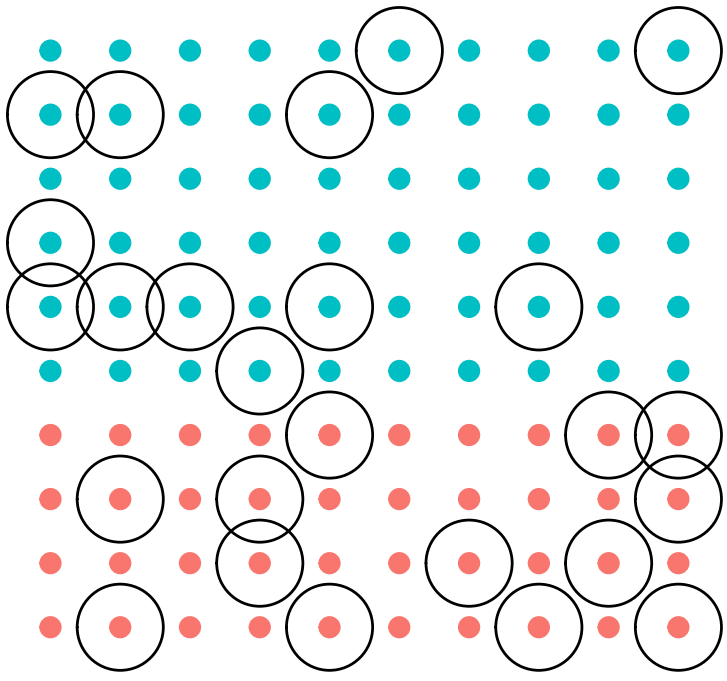
Sample Proportion Blue = 0.68



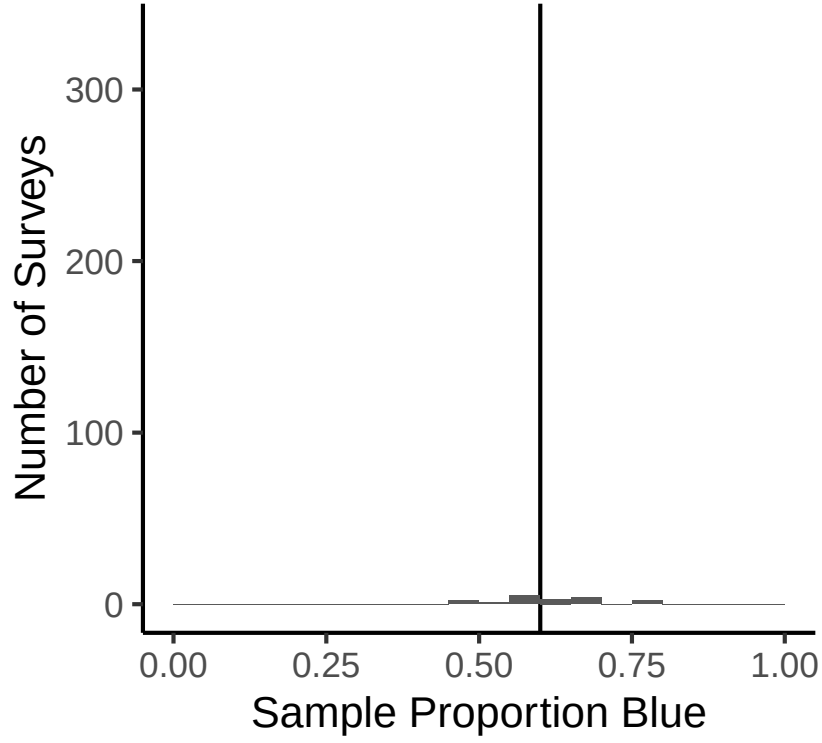
# of Surveys Run = 16



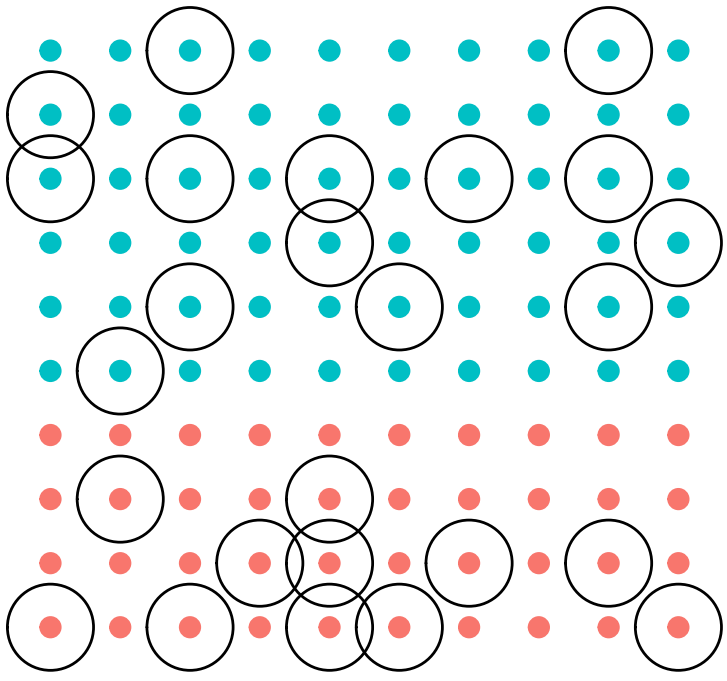
Sample Proportion Blue = 0.48



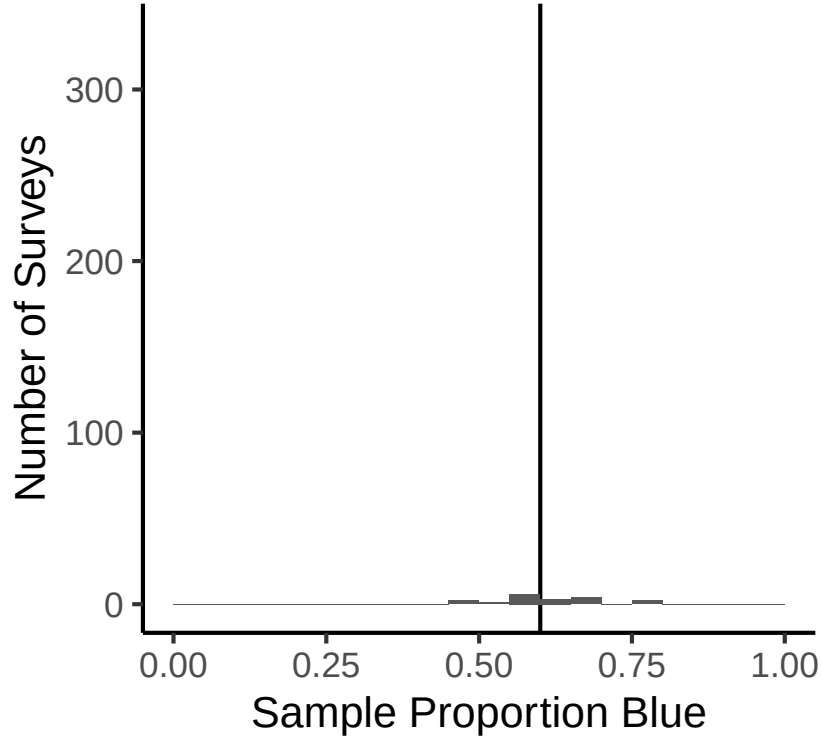
# of Surveys Run = 17



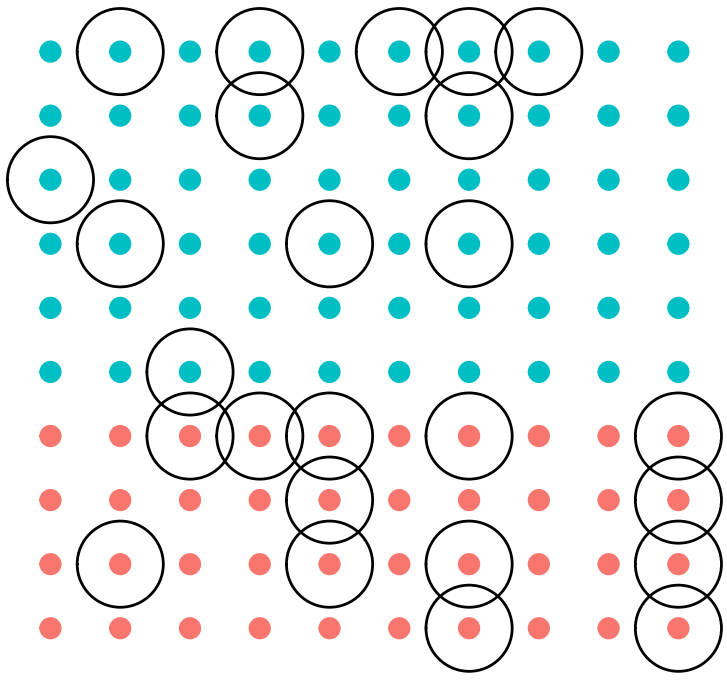
Sample Proportion Blue = 0.56



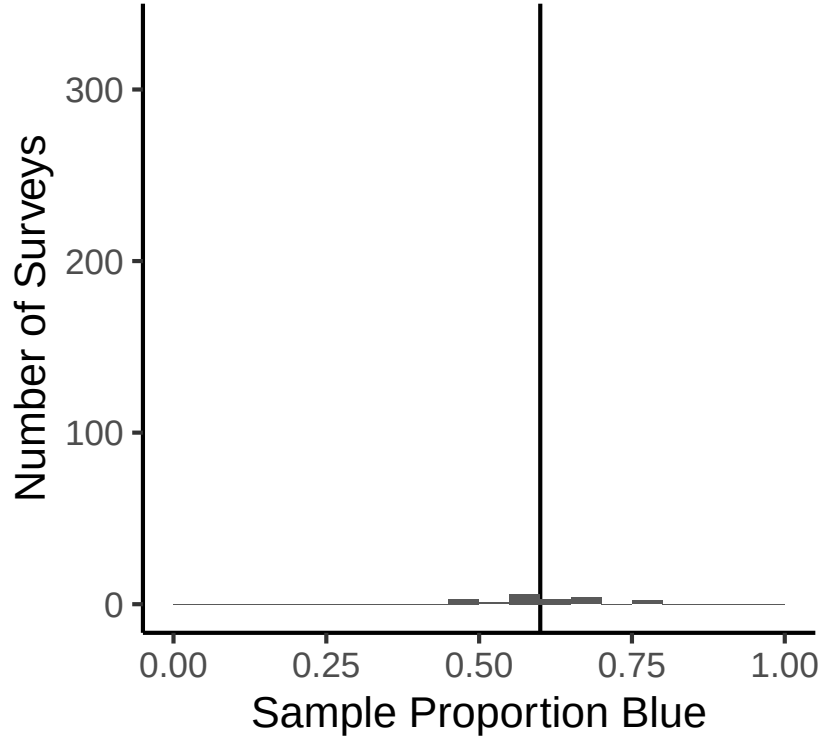
# of Surveys Run = 18



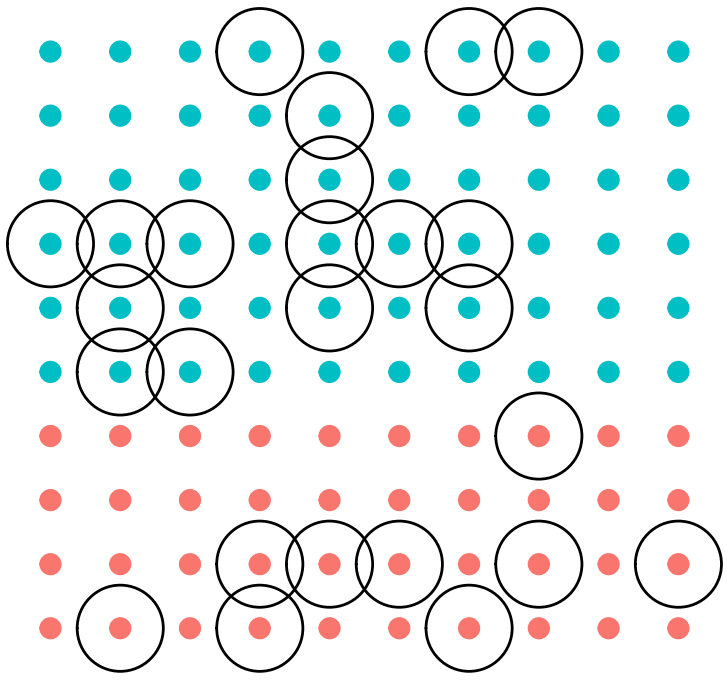
Sample Proportion Blue = 0.48



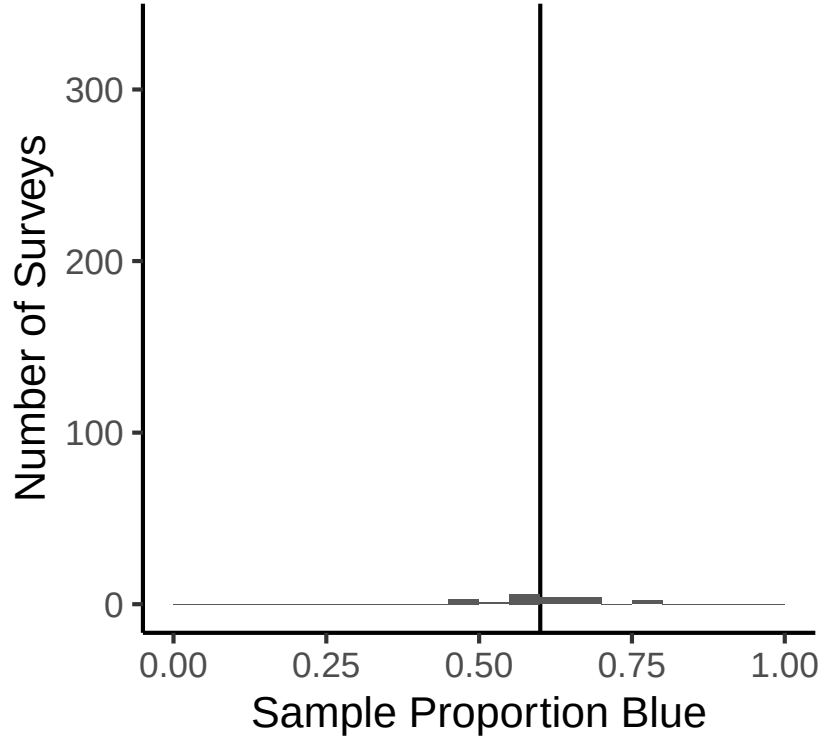
# of Surveys Run = 19



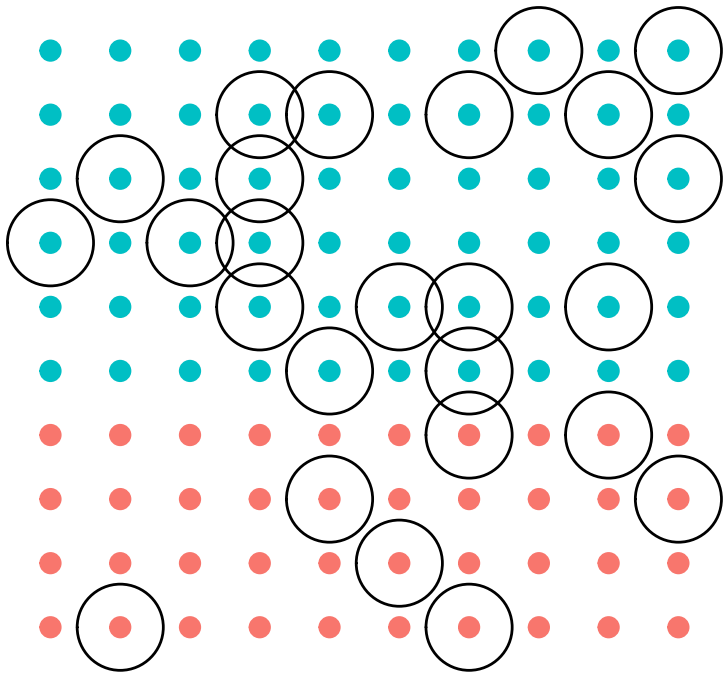
Sample Proportion Blue = 0.64



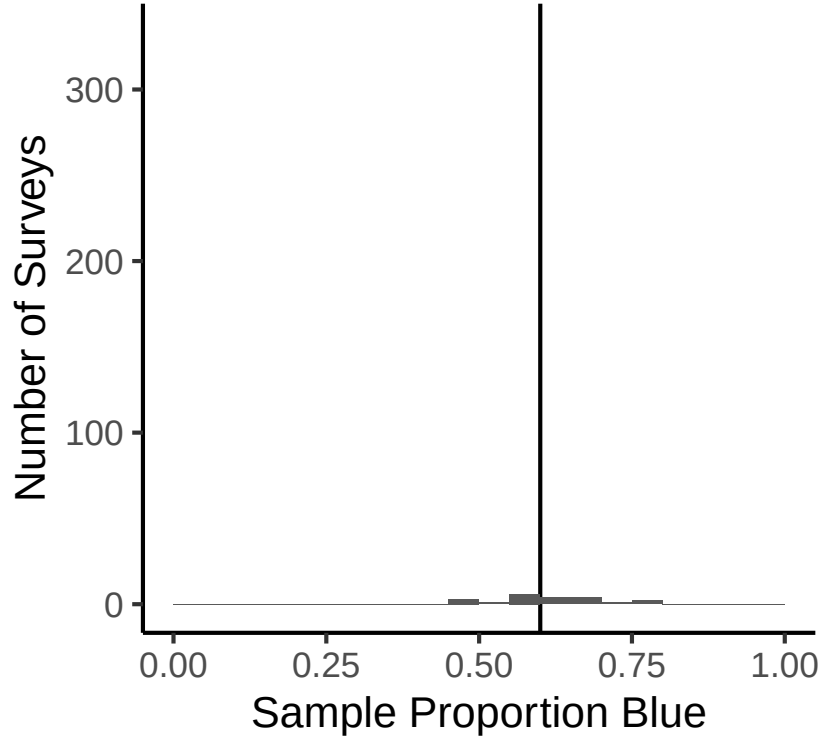
# of Surveys Run = 20



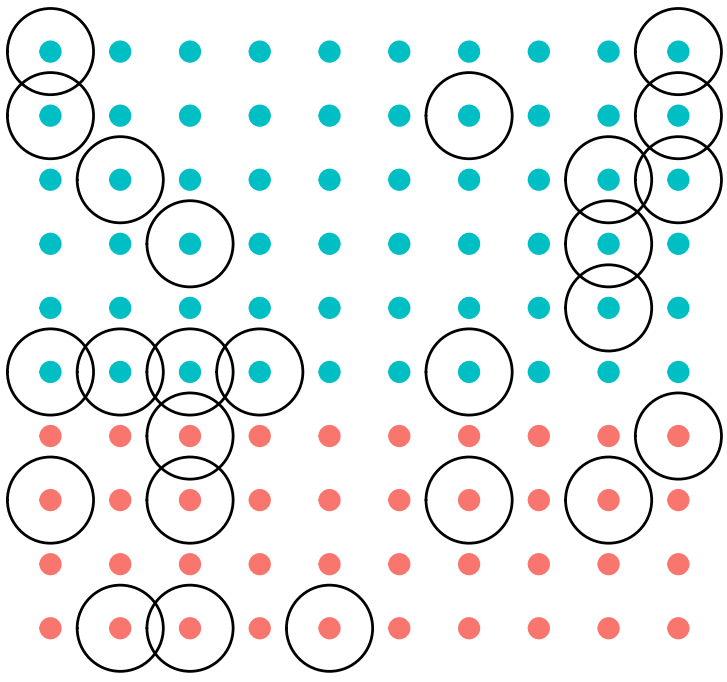
Sample Proportion Blue = 0.72



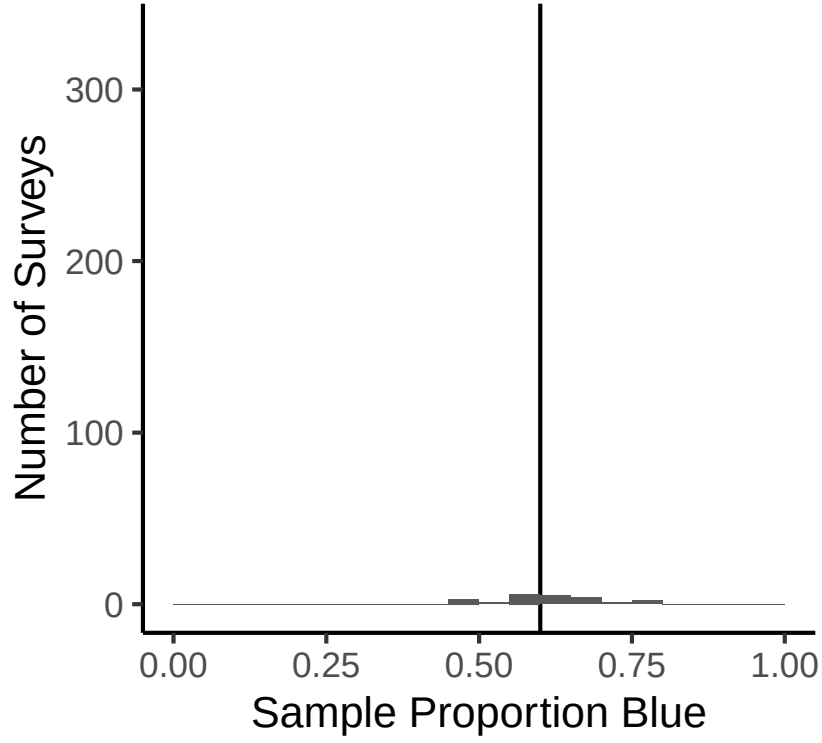
# of Surveys Run = 21



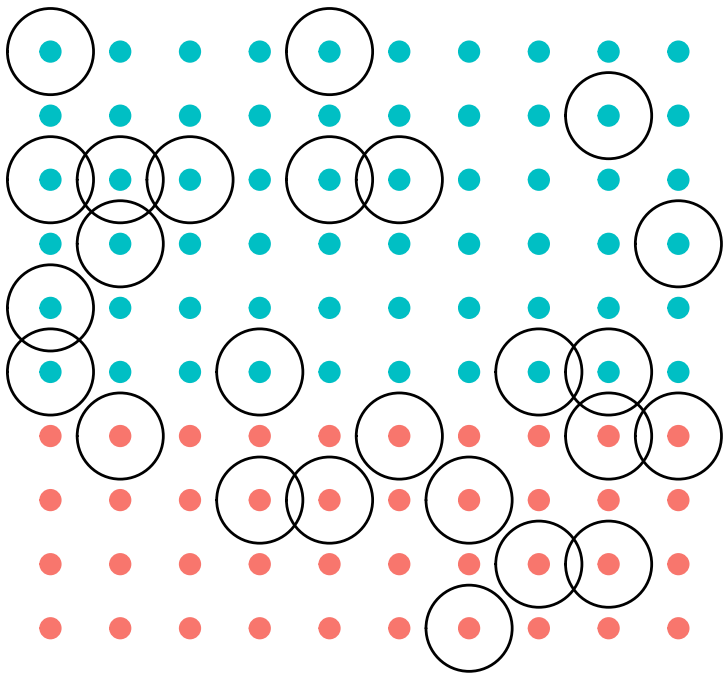
Sample Proportion Blue = 0.64



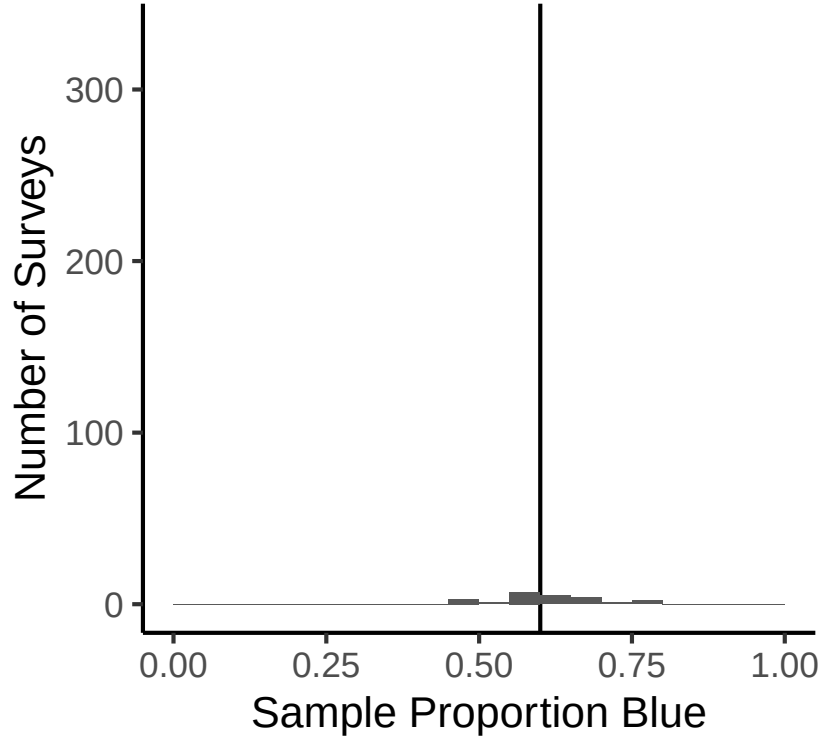
# of Surveys Run = 22



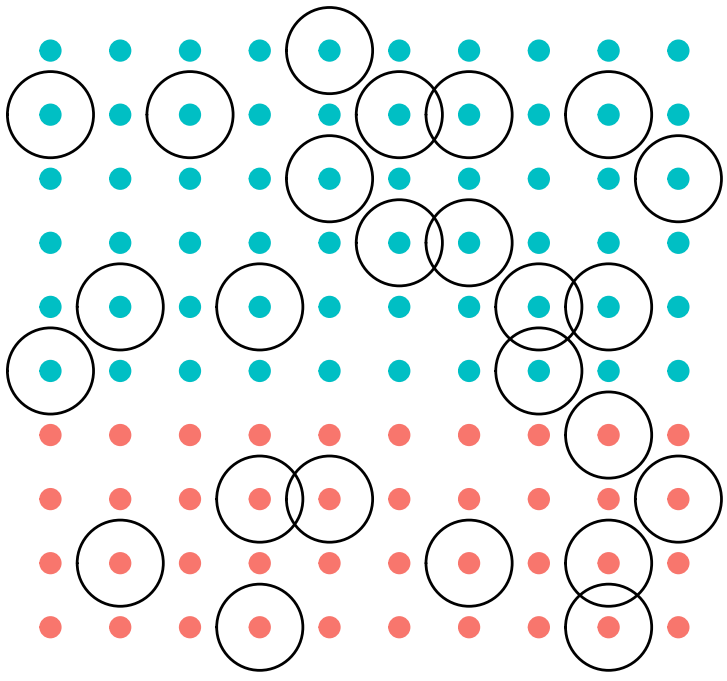
Sample Proportion Blue = 0.6



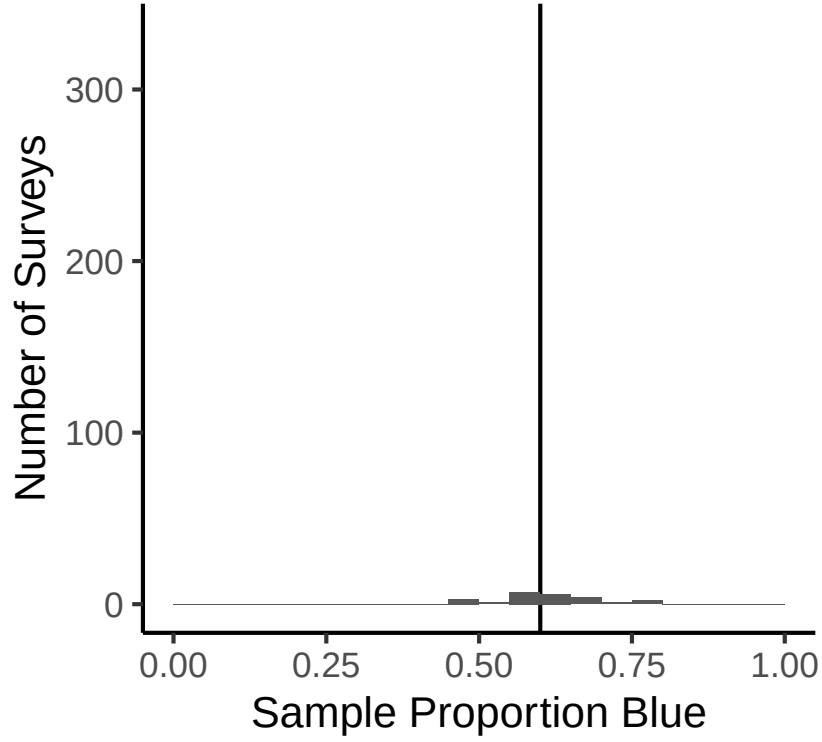
# of Surveys Run = 23



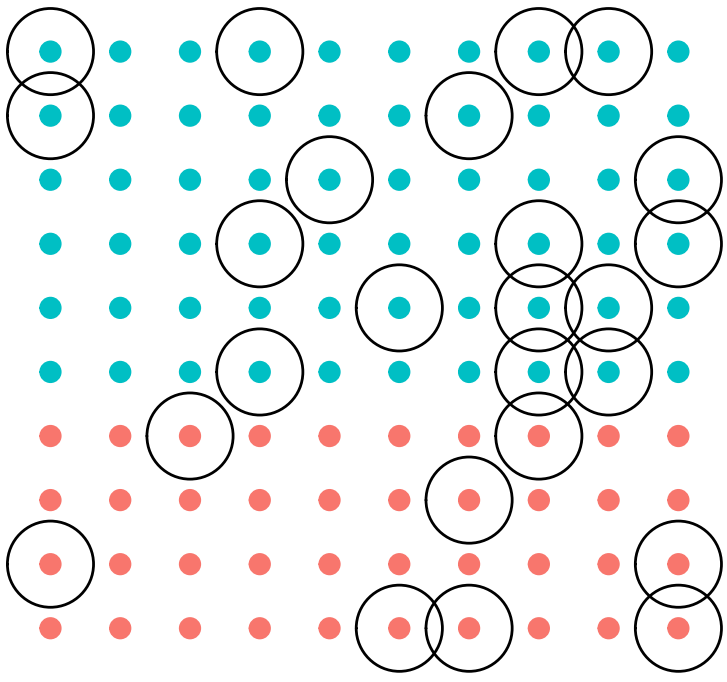
Sample Proportion Blue = 0.64



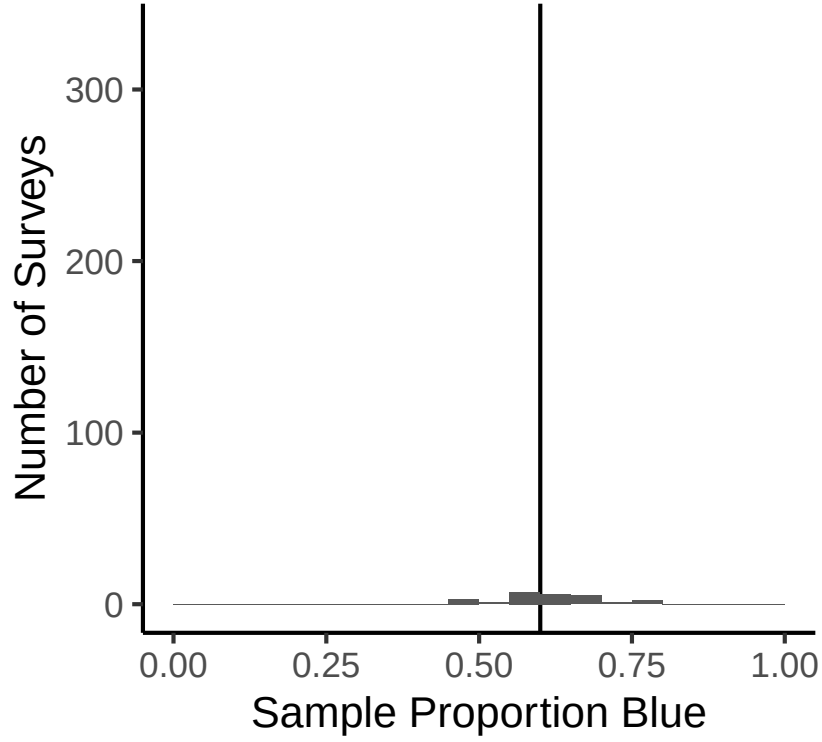
# of Surveys Run = 24



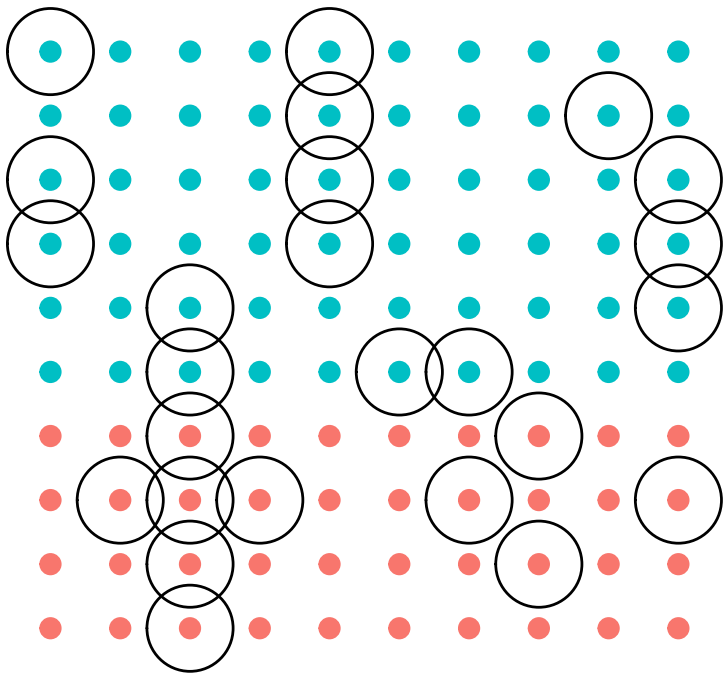
Sample Proportion Blue = 0.68



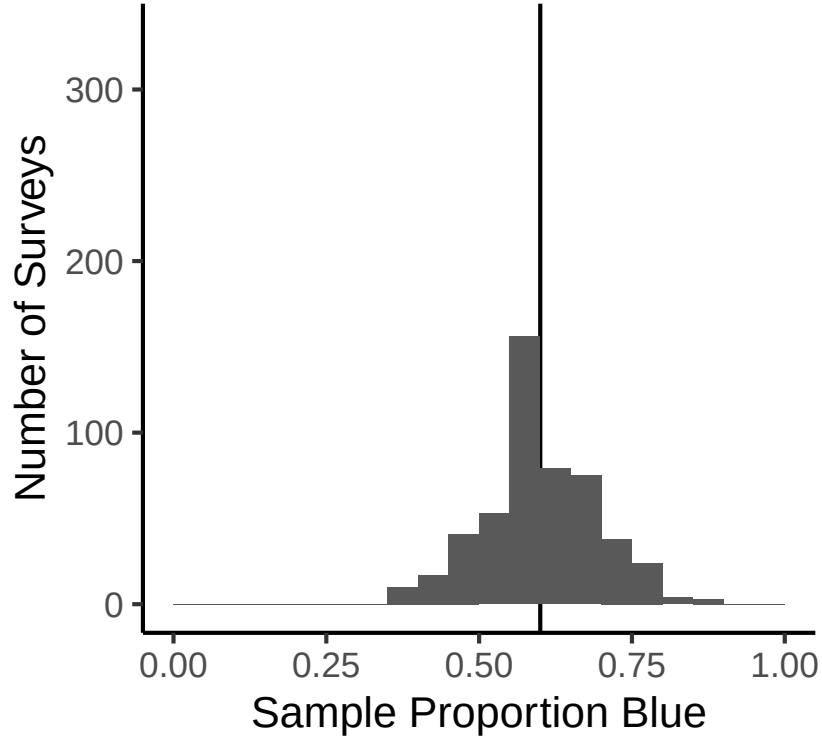
# of Surveys Run = 25



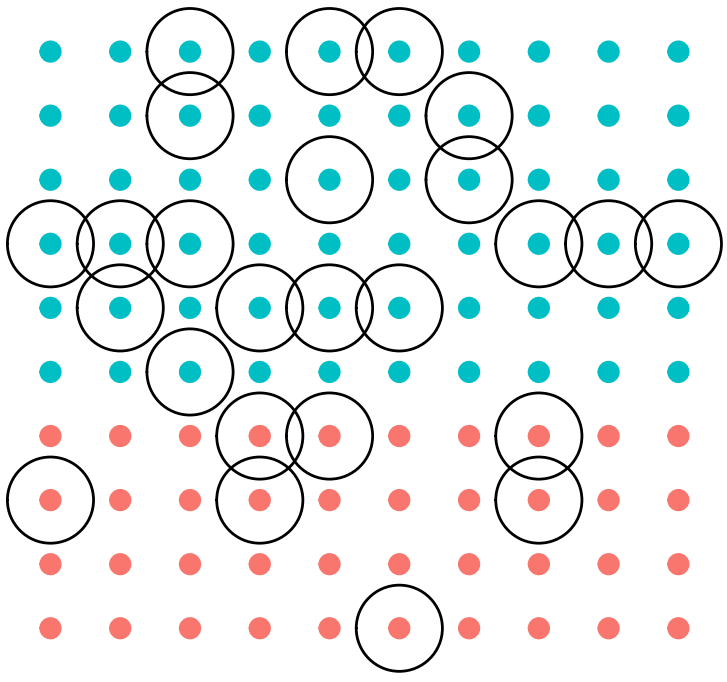
Sample Proportion Blue = 0.6



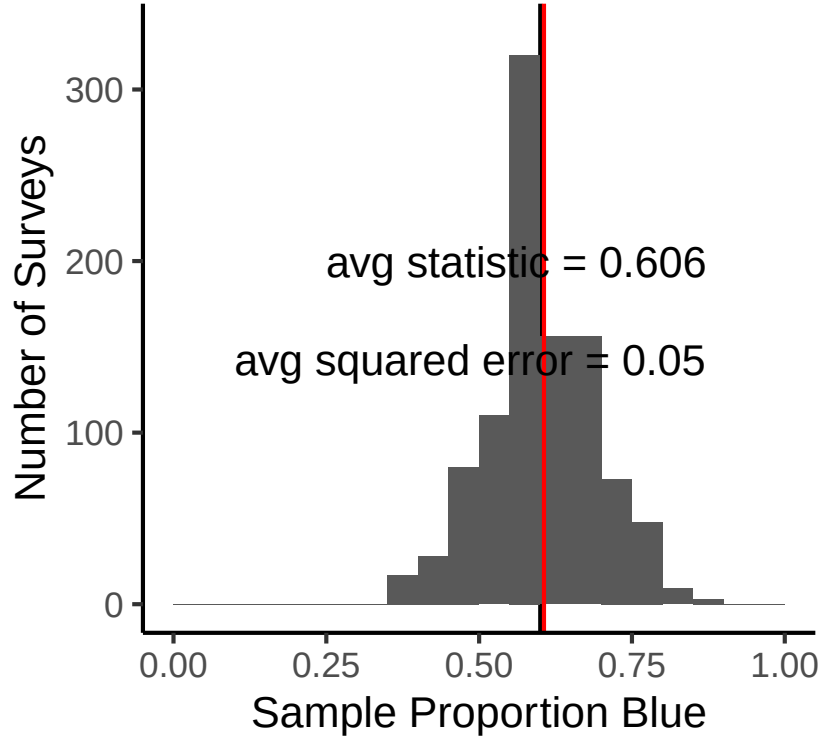
# of Surveys Run = 500



Sample Proportion Blue = 0.72



# of Surveys Run = 1000



Sampling Illustration

- Each time we take a sample, we get a slightly different result
- When we take many samples, we generate a *sampling distribution* of the statistic
- On average, we got the right answer
- But often we had small errors and less commonly we had large errors

What About Sample Size?

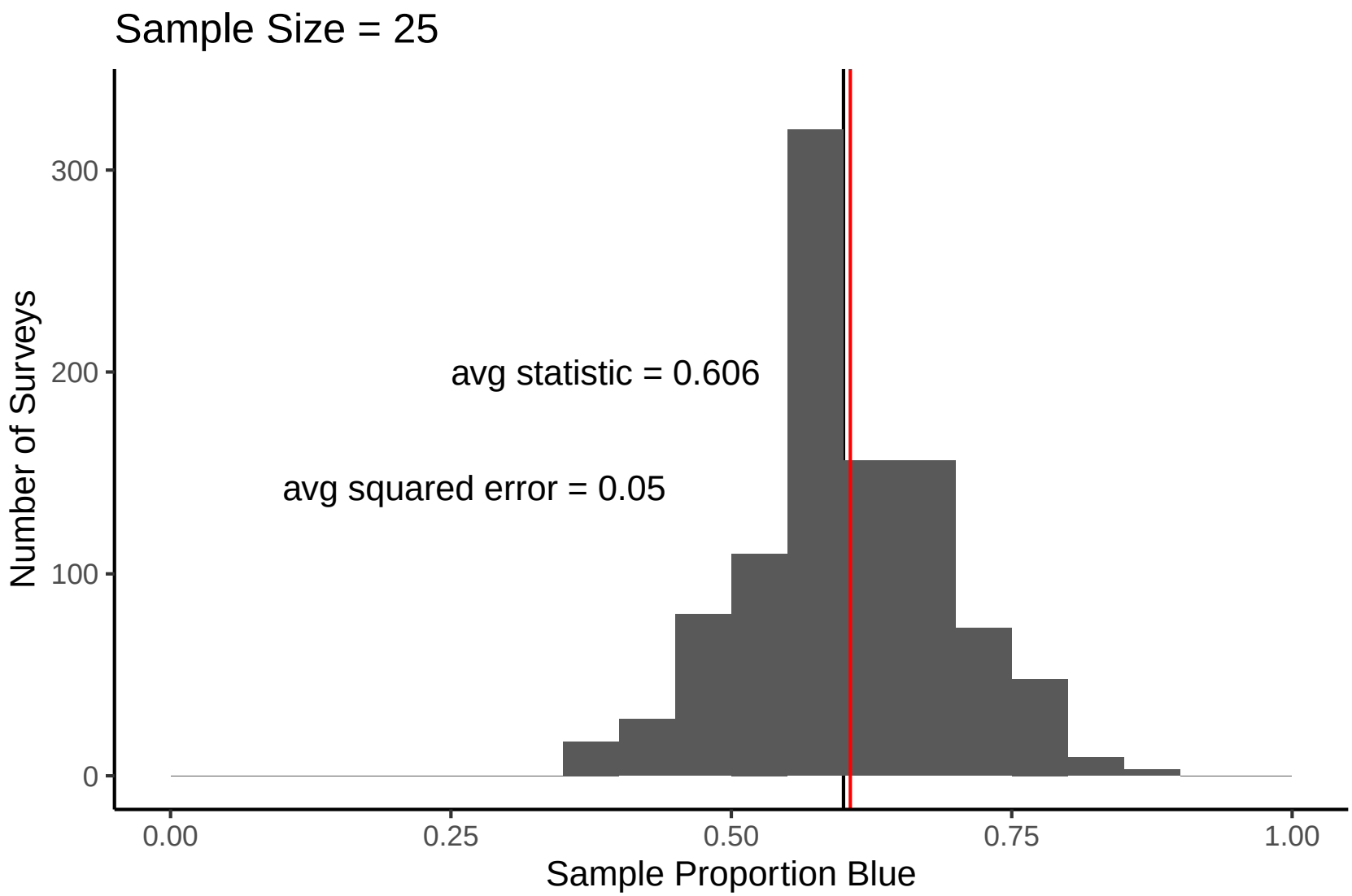
- We had a sample size of 25

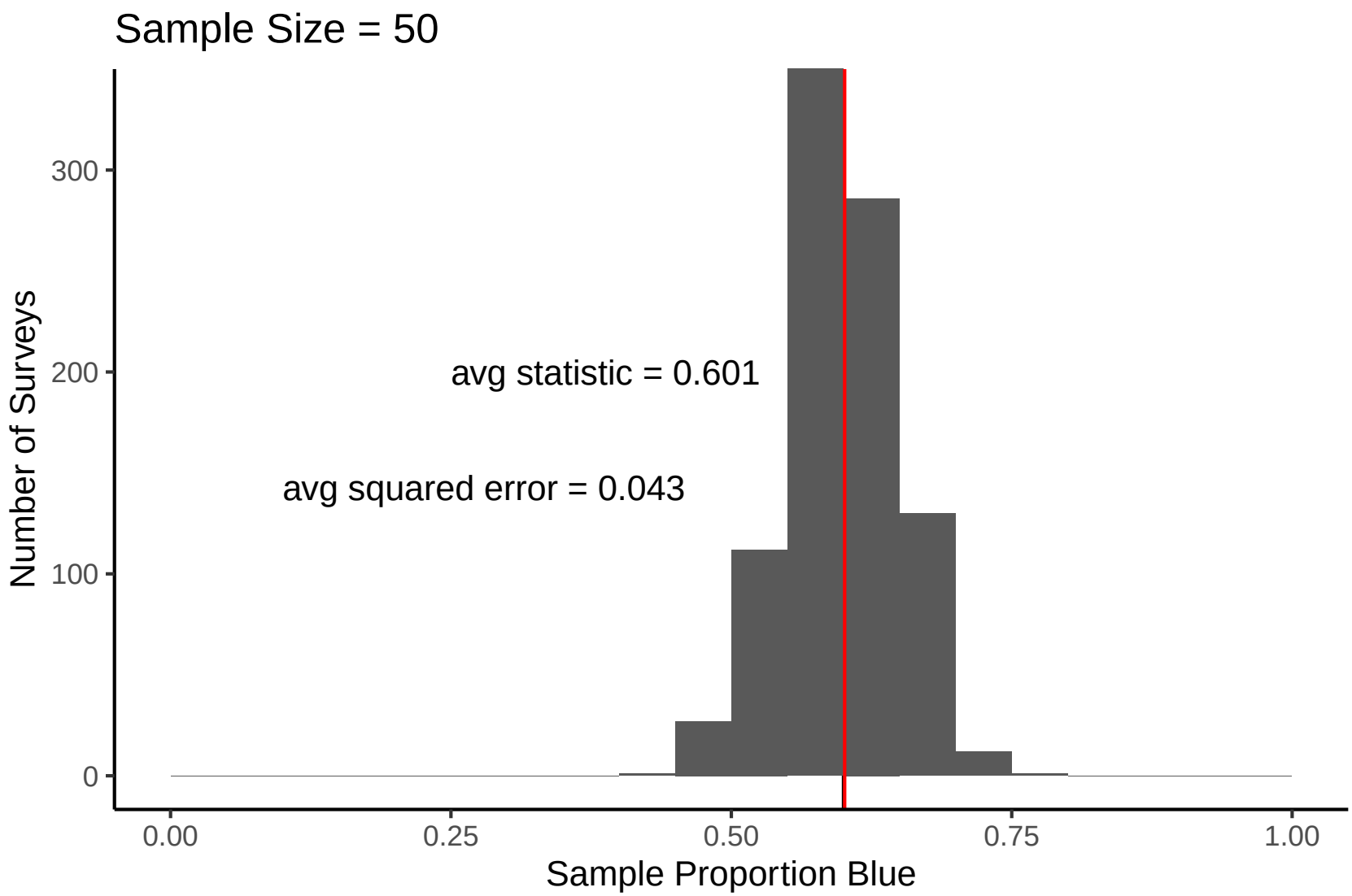
What About Sample Size?

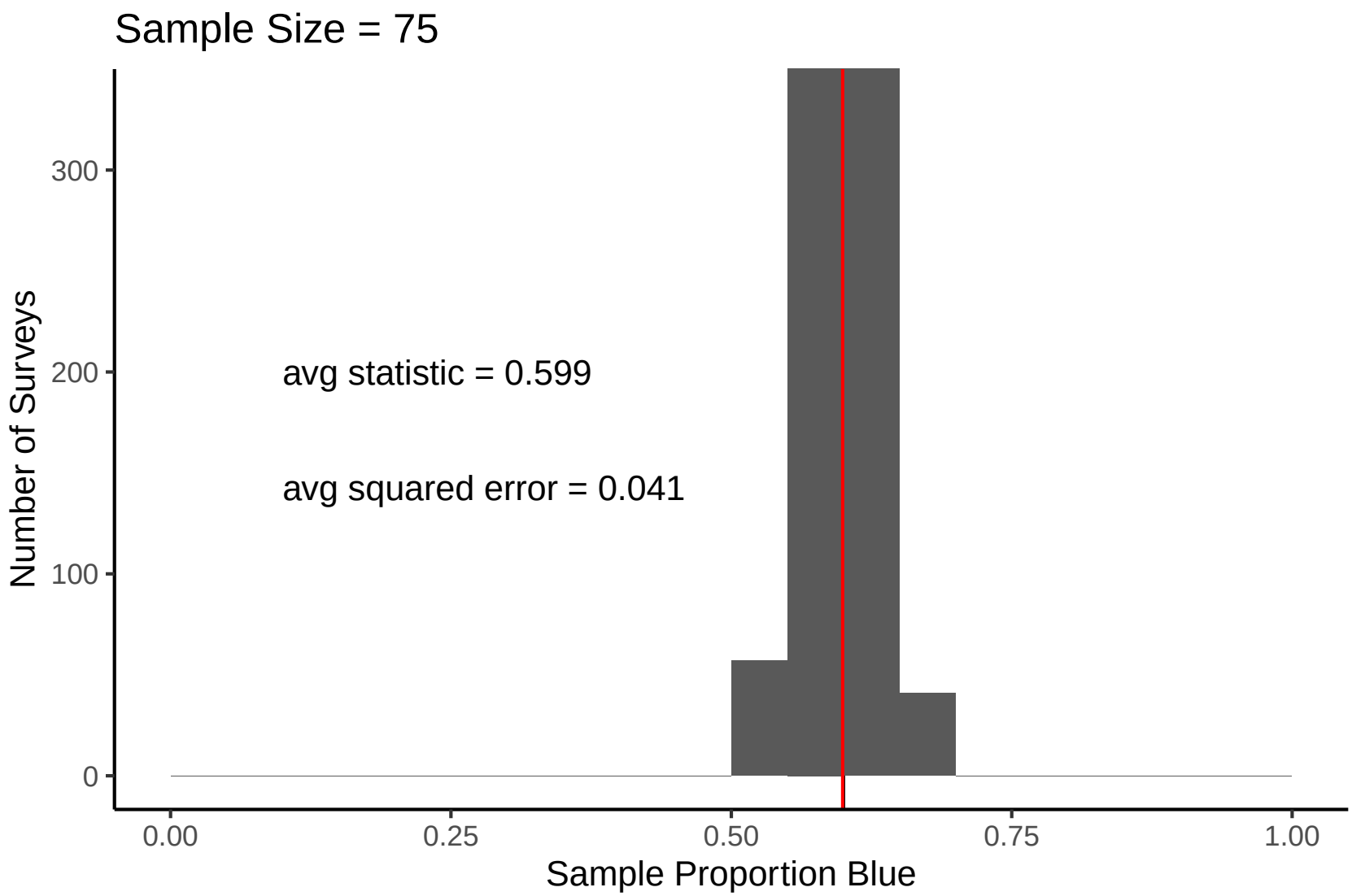
- We had a sample size of 25
- What do you think happens if we change the sample size to 50? 75? 10?

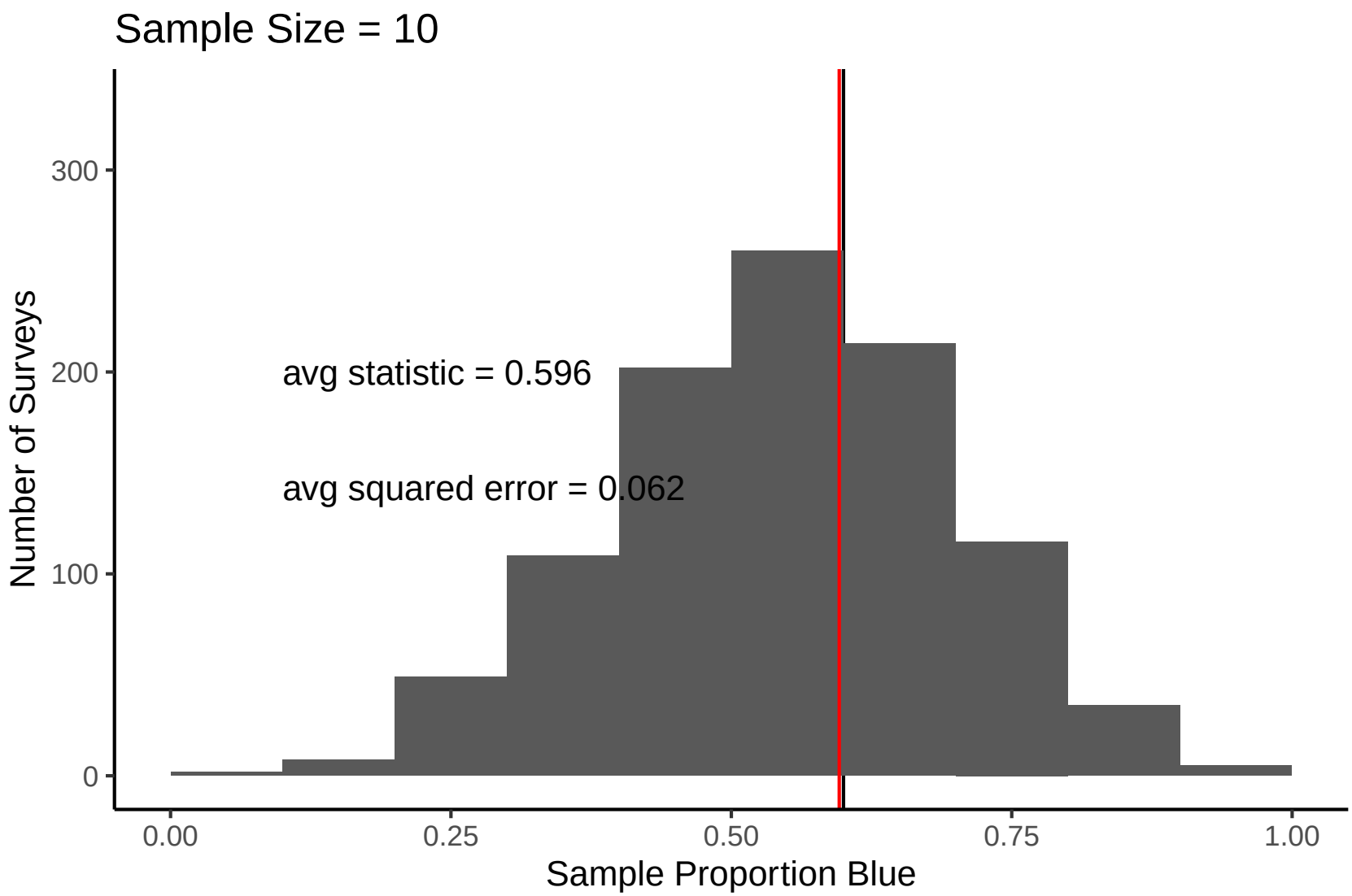
What About Sample Size?

- We had a sample size of 25
- What do you think happens if we change the sample size to 50? 75? 10?


Sample Size = 25
Number of Surveys
300
200
100
0
avg statistic = 0.606
avg squared error = 0.05
0.00
0.25
0.50
0.75
1.00
Sample Proportion Blue


Sample Size = 50
Number of Surveys
300
200
100
0
avg statistic = 0.601
avg squared error = 0.043
0.00
0.25
0.50
0.75
1.00
Sample Proportion Blue


Sample Size = 75
Number of Surveys
300
200
100
0
avg statistic = 0.599
avg squared error = 0.041
0.00
0.25
0.50
0.75
1.00
Sample Proportion Blue


Sample Size = 10
Number of Surveys
300
200
100
0
avg statistic = 0.596
avg squared error = 0.062
0.00
0.25
0.50
0.75
1.00
Sample Proportion Blue

Sampling Distributions

- Every survey gives a different statistic because different units are sampled
- The distribution of this statistic under *repeated sampling* is the *sampling distribution*

Sampling Distributions

- Every survey gives a different statistic because different units are sampled
- The distribution of this statistic under *repeated sampling* is the *sampling distribution*
- The width of the distribution (made precise shortly) is the *sampling variance*
- The center of the distribution is the *sample mean*

Sampling Distributions

- Every survey gives a different statistic because different units are sampled
- The distribution of this statistic under *repeated sampling* is the *sampling distribution*
- The width of the distribution (made precise shortly) is the *sampling variance*
- The center of the distribution is the *sample mean*
- Goal: have the sampling distribution centered at the true population value and narrow in width

Takeaways from the Simulations

1. On average, under random sampling, we get the right answer
2. Any given survey has *sampling error*
3. Error seems to decrease as sample size increases

Takeaways from the Simulations

1. On average, under random sampling, we get the right answer
2. Any given survey has *sampling error*
3. Error seems to decrease as sample size increases

In reality, we only observe one survey. We use statistical theory to quantify properties of the design under repeated sampling.

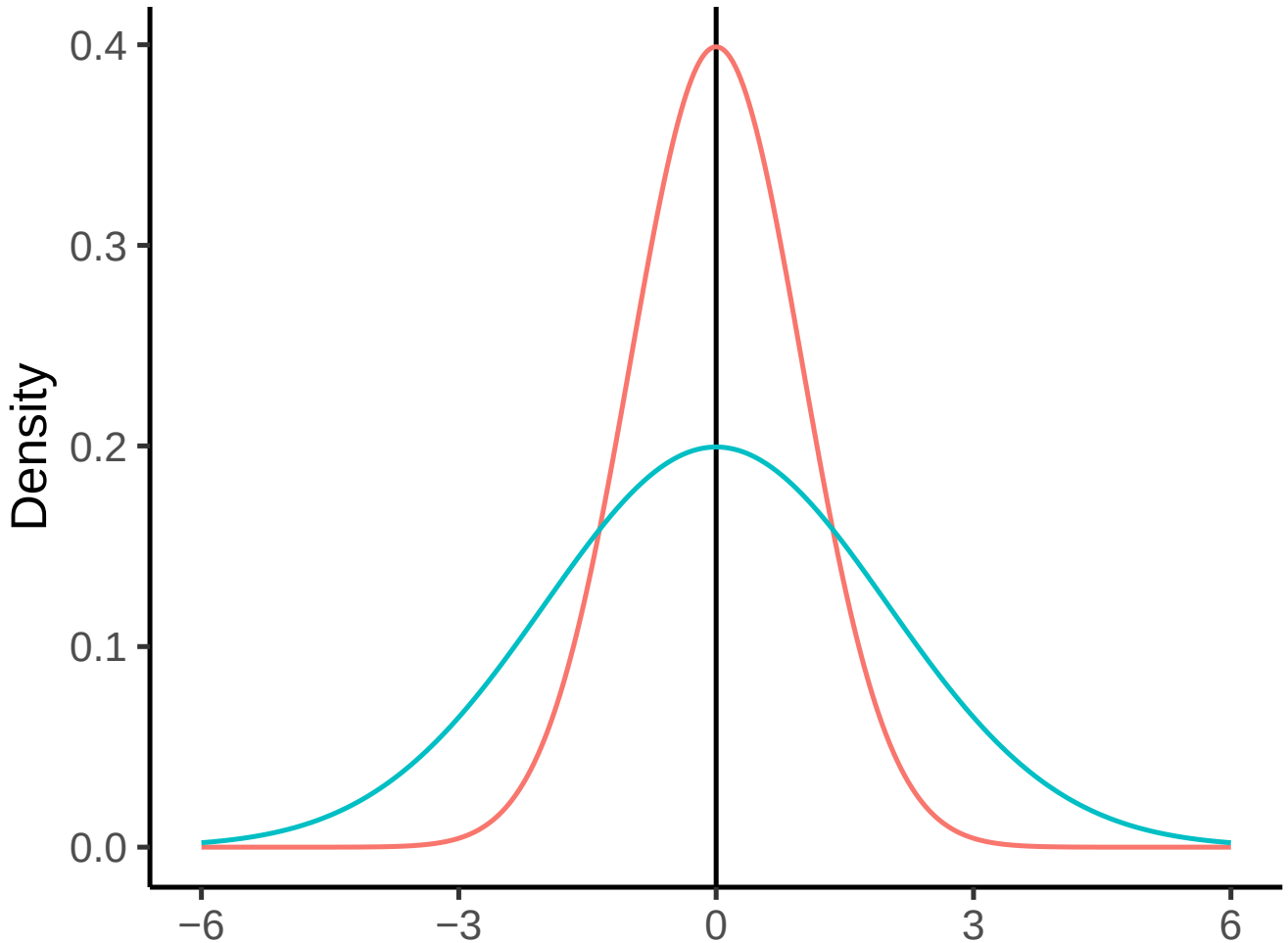
Statistics

Overview

- Theory can tell us about the sampling distribution
- We'll use a *little* math to derive properties of sampling strategies
- We'll cover mean, variance, and standard error

Mean

Often we're interest in "measures of central tendency" — averages.



Mean

The mean (aka the "arithmetic mean") is the most common measure.

Mean

The mean (aka the "arithmetic mean") is the most common measure.

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Mean

The mean (aka the "arithmetic mean") is the most common measure.

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\left(y_1 + y_2 + \cdots + y_n\right)\end{aligned}$$

Mean

The mean (aka the "arithmetic mean") is the most common measure.

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\left(y_1 + y_2 + \cdots + y_n\right)\end{aligned}$$

Mean

The mean (aka the "arithmetic mean") is the most common measure.

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\left(y_1 + y_2 + \cdots + y_n\right)\end{aligned}$$

The mean is our "best guess" for the value of a randomly chosen unit.

Mean

The mean (aka the "arithmetic mean") is the most common measure.

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}\left(y_1 + y_2 + \cdots + y_n\right)\end{aligned}$$

The mean is our "best guess" for the value of a randomly chosen unit.

In `R`: `mean()`

Mean of a Binary Variable

For a binary (0-1) variable, calculate the mean to get the proportion of 1's.

Mean of a Binary Variable

For a binary (0-1) variable, calculate the mean to get the proportion of 1's.

$$\text{Proportion} = \frac{\#\text{ of 1's}}{\text{Total number of Units}}$$

Mean of a Binary Variable

For a binary (0-1) variable, calculate the mean to get the proportion of 1's.

$$\text{Proportion} = \frac{\#\text{ of 1's}}{\text{Total number of Units}}$$

$$\text{mean}(y_i) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Mean of a Binary Variable

For a binary (0-1) variable, calculate the mean to get the proportion of 1's.

$$\text{Proportion} = \frac{\#\text{ of 1's}}{\text{Total number of Units}}$$

$$\begin{aligned}\text{mean}(y_i) &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{1}{n}[y_1 + y_2 + \cdots + y_n]\end{aligned}$$

Mean of a Binary Variable

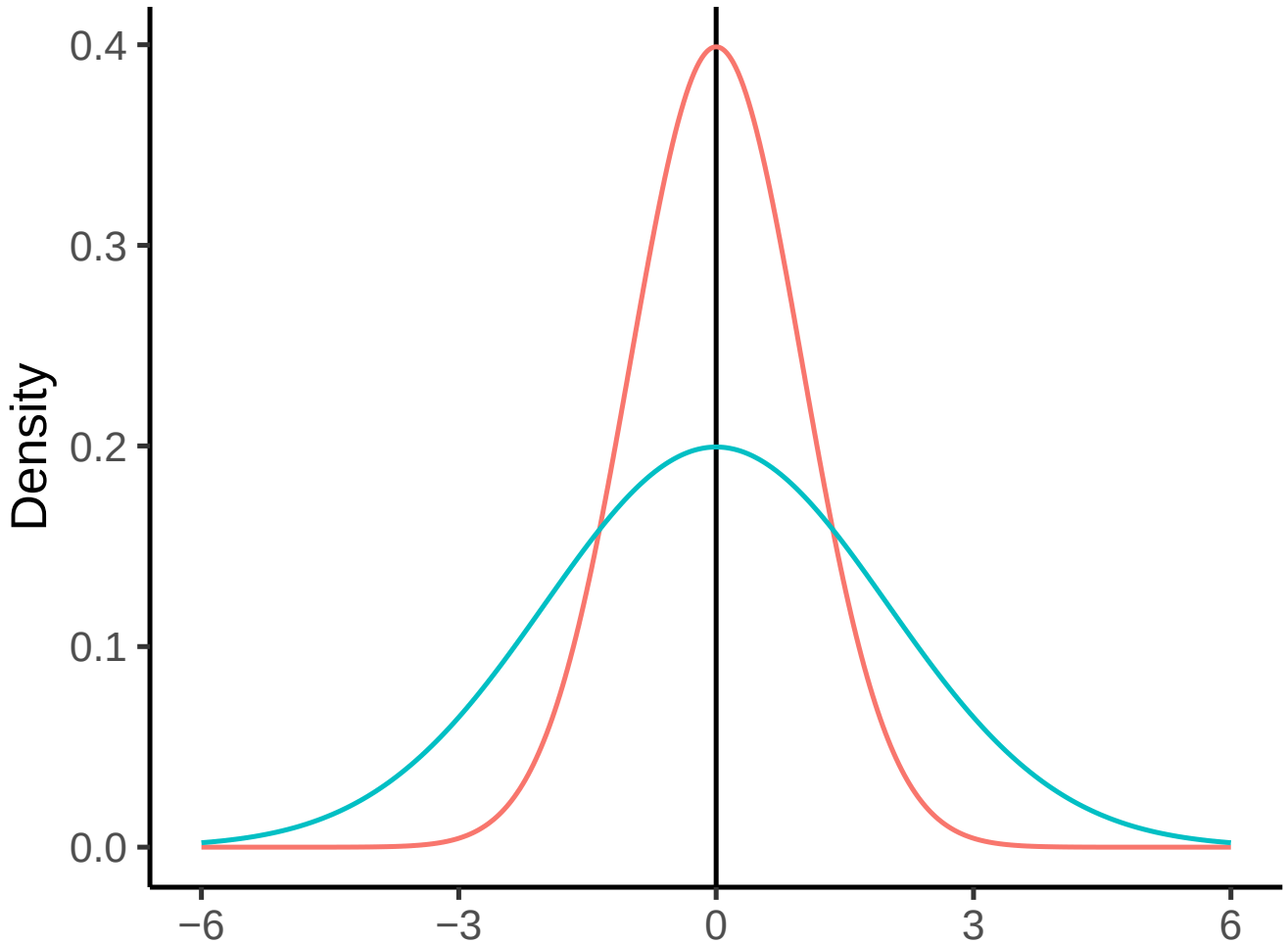
For a binary (0-1) variable, calculate the mean to get the proportion of 1's.

$$\text{Proportion} = \frac{\#\text{ of 1's}}{\text{Total number of Units}}$$

$$\begin{aligned}
\text{mean}(y_i) &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \frac{1}{n}[y_1 + y_2 + \cdots + y_n] \\
&= \frac{1}{n}[\#\text{ of 1's}] \\
&= \frac{\#\text{ of 1's}}{\text{Total number of Units}}
\end{aligned}$$

Variance

The variance measures how dispersed a distribution is



Variance and Standard Deviation

Variance quantifies how far any given unit is from the mean, on average.

Variance and Standard Deviation

Variance quantifies how far any given unit is from the mean, on average.

$$\text{var}(y_i) := \sigma^2$$

Variance and Standard Deviation

Variance quantifies how far any given unit is from the mean, on average.

$$\begin{aligned} \text{var}(y_i) &:= \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \end{aligned}$$

Variance and Standard Deviation

Variance quantifies how far any given unit is from the mean, on average.

$$\begin{aligned}\text{var}(y_i) &:= \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \text{mean}((y_i - \bar{y})^2)\end{aligned}$$

Variance and Standard Deviation

Variance quantifies how far any given unit is from the mean, on average.

$$\begin{aligned}\text{var}(y_i) &:= \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \text{mean}((y_i - \bar{y})^2)\end{aligned}$$

Standard deviation is the square root of the variance:

$$\text{sd}(y_i) = \sigma = \sqrt{\sigma^2}$$

Variance and Standard Deviation

Variance quantifies how far any given unit is from the mean, on average.

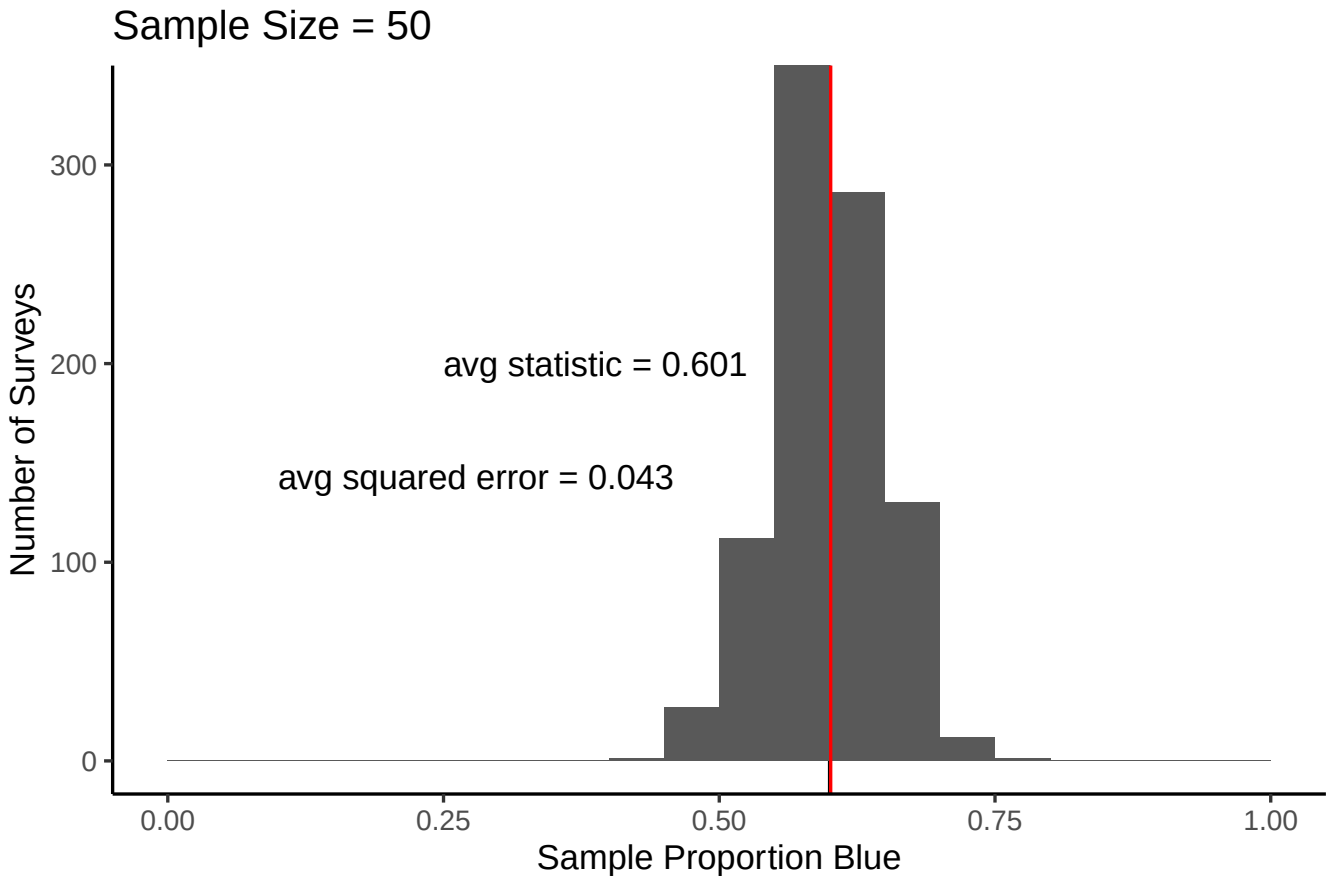
$$\begin{aligned}\text{var}(y_i) &:= \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \text{mean}((y_i - \bar{y})^2)\end{aligned}$$

Standard deviation is the square root of the variance:

$$\text{sd}(y_i) = \sigma = \sqrt{\sigma^2}$$

In `R`: `var()` and `sd()`

Repeated sampling generates a *sampling distribution* of the statistic



Bias and Standard Error

The mean and variance of this distribution tell us whether the statistic is correct, on average, and how far away we can expect any single *realization* to be from the truth.

- Under SRS, the mean of sampling distribution is the true value in the population: SRS is *unbiased*
- The standard deviation of the sampling distribution is the *standard error*

Unbiasedness of Random Sampling

- The mean is *unbiased* under simple random sampling
- We might have error in one survey, but on average we get the right answer
- If selection into sample were related to the outcome variable, we would have *bias*

Standard Error Under Simple Random Sampling

We'll call the estimate of the mean $\hat{y}$. The variance and standard error of the mean under SRS are given by:

$$\begin{aligned}
\text{var}(\hat{y}) &= \text{var}\left(\frac{1}{n}\sum y_i\right) \\
&= \frac{1}{n^2}\sum \text{var}(y_i) \\
&= \frac{n\ \text{var}(y_i)}{n^2} \\
&= \frac{\text{var}(y_i)}{n} \\
\text{se}(\hat{y}) &= \frac{\text{sd}(y_i)}{\sqrt{n}}
\end{aligned}$$

Standard Error Under Simple Random Sampling

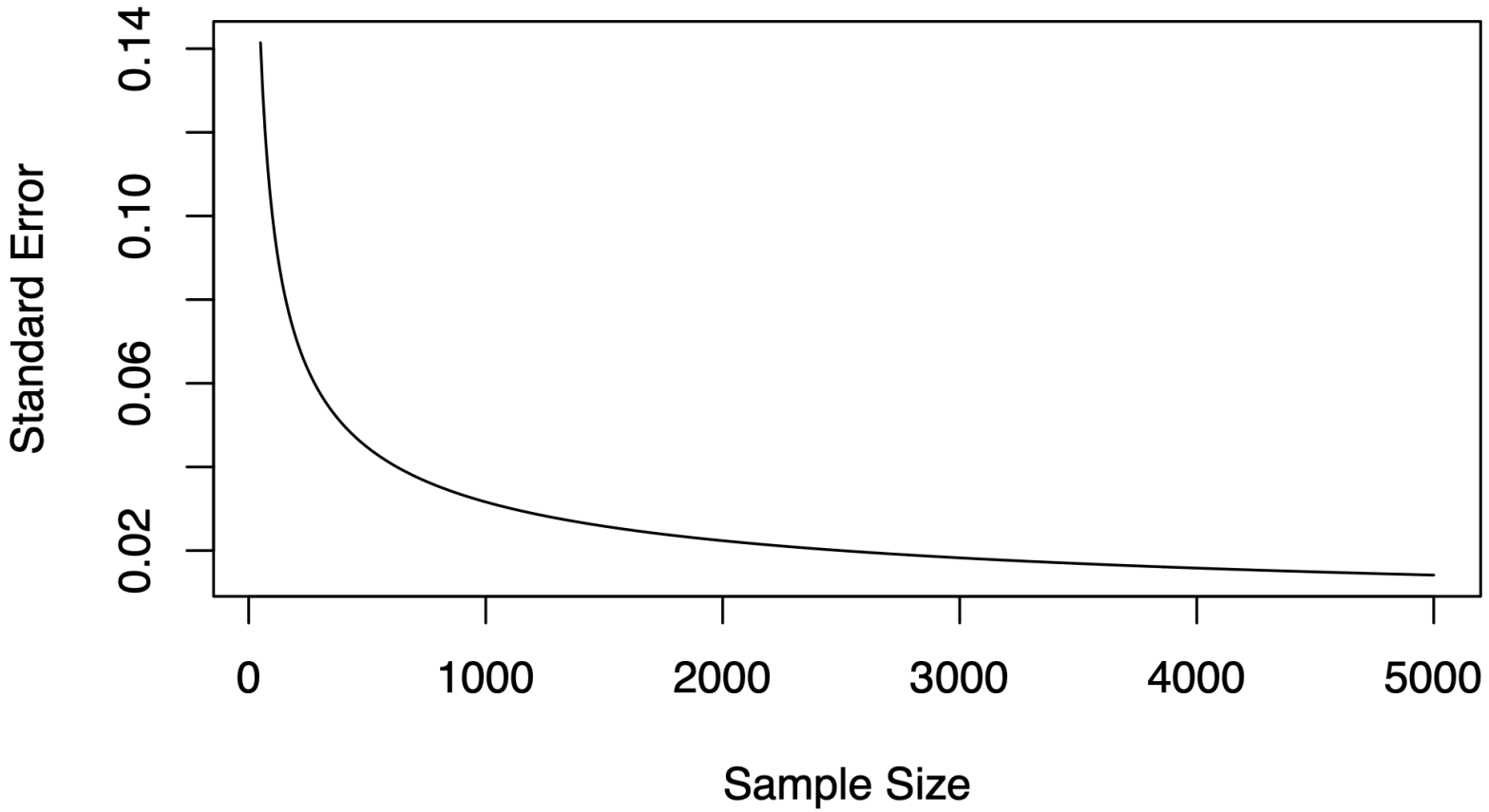
We'll call the estimate of the mean $\hat{y}$. The variance and standard error of the mean under SRS are given by:

$$\begin{aligned}
\text{var}(\hat{y}) &= \text{var}\left(\frac{1}{n}\sum y_i\right) \\
&= \frac{1}{n^2}\sum \text{var}(y_i) \\
&= \frac{n\ \text{var}(y_i)}{n^2} \\
&= \frac{\text{var}(y_i)}{n} \\
\text{se}(\hat{y}) &= \frac{\text{sd}(y_i)}{\sqrt{n}}
\end{aligned}$$

Sample size n is in the denominator. Larger sample size $\rightsquigarrow$ smaller standard errorb

Standard Error Decreases as Sample Size Increases

...but there are diminishing returns to increasing sample size!



Confidence Intervals

- We can use the standard error to construct a *confidence interval* (statistics term) or *margin of error* (survey term)
- The CI gives a measure of uncertainty
- We construct the 95% CI by adding and subtracting 1.96 times the standard error to our estimate:

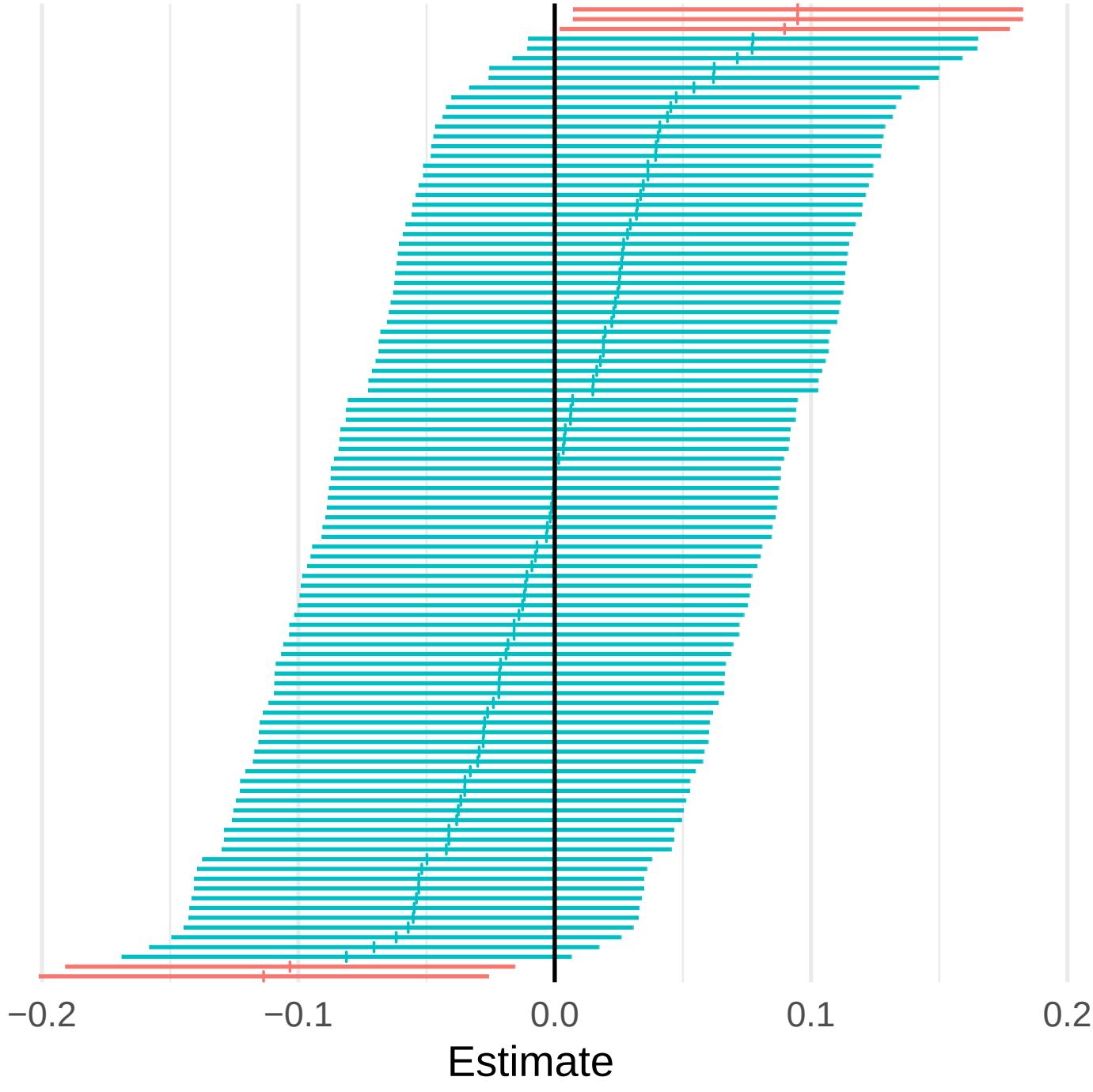
$$CI = \hat{y} \pm 1.96 \times \text{se}(\hat{y})$$

- "Margin of error" usually refers to $1.96 \times \text{se}(\hat{y})$

Confidence Intervals

A 95% confidence interval means the following:

- If we were to conduct the survey 100 times, in 95 of those surveys, the true value would be contained in the 95% confidence interval


−0.2
−0.1
0.0
0.1
0.2
Estimate

Uncertainty

- Always consider sampling uncertainty in polls
- Easy to over-interpret small changes in poll results: could be due to random chance!
- When reporting results, always include confidence intervals or margins of error

Calculating Margin of Error

to R $\rightsquigarrow$

Beyond Simple Random Sampling

Cluster Sampling

- Group units into groups, then sample groups
- Common for household surveys, face-to-face surveys, surveys of schools, etc.
- Advantage: Often easier and cheaper
- Drawback: Increased sampling variance, more uncertainty

Illustration in R $\rightsquigarrow$

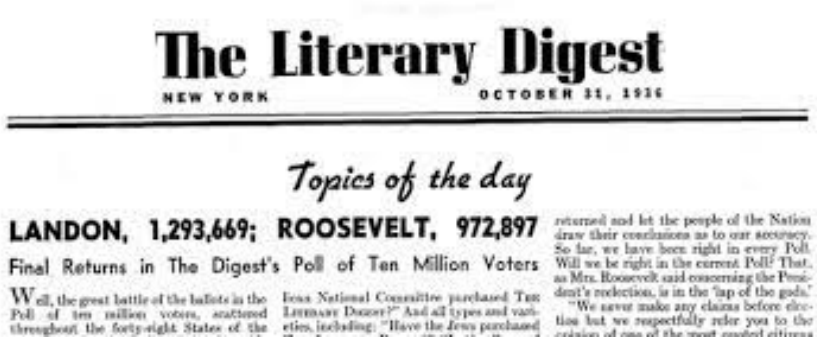
Stratified Sampling

- Separate units in "strata"
- Select strata for sampling in first stage, then units from within the strata in the second stage
- Advantage: can be used to ensure adequate sample sizes in subgroups
- Drawback: need data on units to put them into strata
- Commonly used for political surveys: target geographic locations (e.g. states), racial groups, etc.

Illstration in R $\rightsquigarrow$

When Non-Random Sampling Goes Wrong

Literary Digest Poll



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

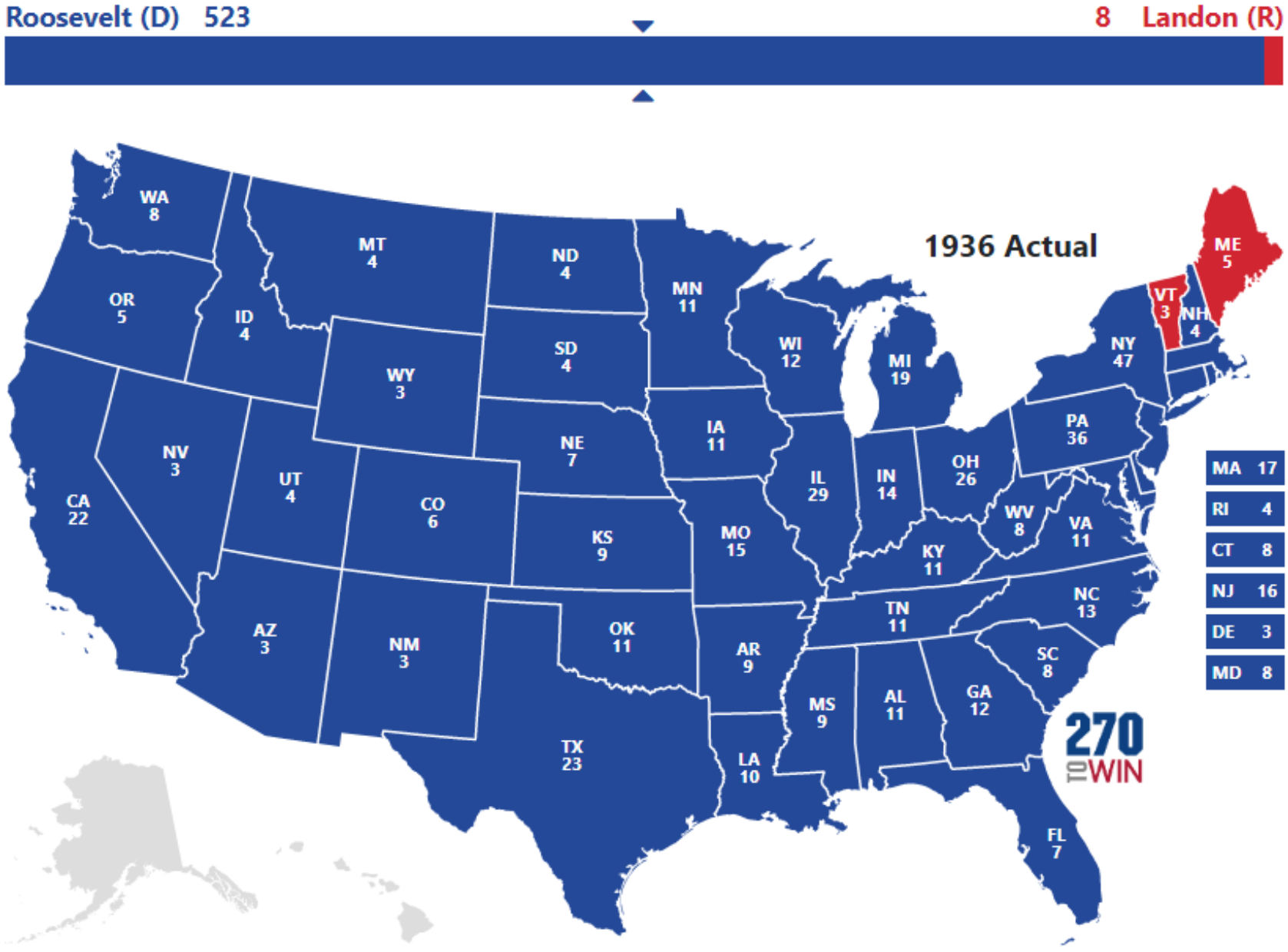
Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Iowa National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

- 1936 election
- Sent 10 million ballots, got $>$ 2.3 million responses
- Got addresses from car and phone registration records


Roosevelt (D) 523
8 Landon (R)
1936 Actual
WA 8
OR 5
CA 22
ID 4
NV 3
MT 4
WY 3
UT 4
AZ 3
CO 6
NM 3
ND 4
SD 4
NE 7
KS 9
OK 11
TX 23
MN 11
IA 11
MO 15
AR 9
LA 10
WI 12
IL 29
MS 9
MI 19
IN 14
KY 11
TN 11
AL 11
OH 26
WV 8
GA 12
FL 7
SC 8
NC 13
VA 11
PA 36
NY 47
VT 3
NH 4
ME 5
MA 17
RI 4
CT 8
NJ 16
DE 3
MD 8
270 TO WIN

Literary Digest Poll

Table 1. 1936 Presidential Vote by Car and Telephone Ownership (in Percent)

Presidential Vote	Car & Phone	Car, No Phone	Phone, No Car	Neither
Roosevelt	55	68	69	79
Landon	45	30	30	19
Other	1	2	0	2
Total *N*	946	447	236	657

SOURCE: American Institute of Public Opinion, 28 May 1937.

Literary Digest Poll

Table 2. Presidential Vote by Receiving *Literary Digest* Straw Vote Ballot or Not (in Percent)

Presidential Vote	Received Poll	Not Receive Poll	Do Not Know
Roosevelt	55	71	73
Landon	44	27	25
Other	1	1	3
Total *N*	780	1339	149

SOURCE: American Institute of Public Opinion, 28 May 1937.