Sampling Distributions

Al Nosedal.
University of Toronto.
Fall 2018

June 10, 2019

Sampling Distribution

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Toy Problem

- We have a population with a total of six individuals: A, B, C, D, E and F.
- All of them voted for one of two candidates: Bert or Ernie.
- A and B voted for Bert and the remaining four people voted for Ernie.
- Proportion of voters who support Bert is $p = \frac{2}{6} = 33.33\%$. This is an example of a population parameter.

Toy Problem

- We are going to estimate the population proportion of people who voted for Bert, p, using information coming from an exit poll of size two.
- Ultimate goal is seeing if we could use this procedure to predict the outcome of this election.

List of all possible samples

{A,B}	{B,C}	{C,E}
{A,C}	{B,D}	{C,F}
{A,D}	{B,E}	{D,E}
{A,E}	{B,F}	{D,F}
{A,F}	{C,D}	{E,F}

Sample proportion

The proportion of people who voted for Bert in each of the possible random samples of size two is an example of a statistic. In this case, it is a sample proportion because it is the proportion of Bert's supporters within a sample; we use the symbol $\hat{p}$ (read "p-hat") to distinguish this sample proportion from the population proportion, p.

List of possible estimates

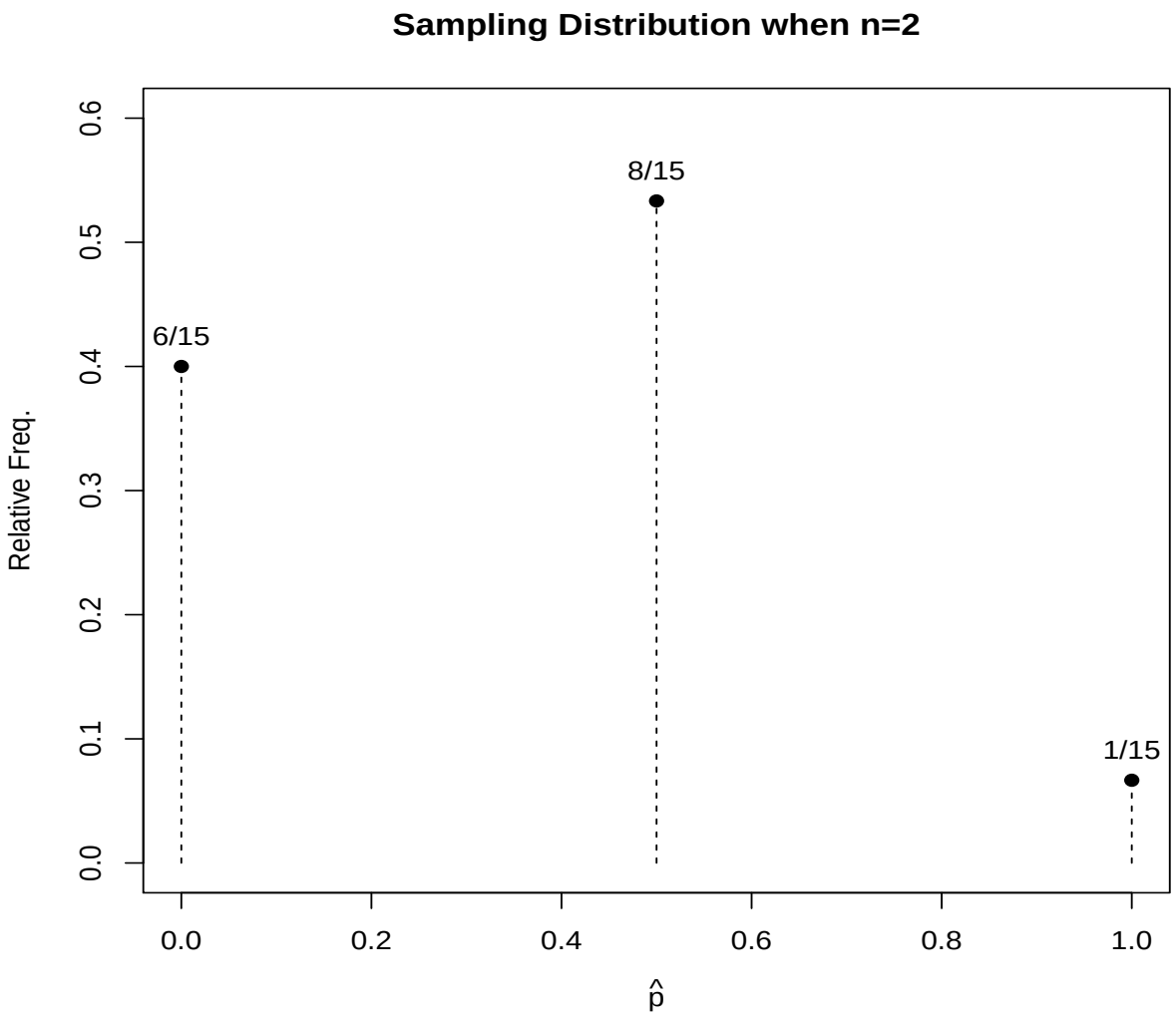
$\hat{p}_1$={A,B} = {1,1}=100%
$\hat{p}_2$={A,C} = {1,0}=50%
$\hat{p}_3$={A,D} = {1,0}=50%
$\hat{p}_4$ ={A,E}= {1,0}=50%
$\hat{p}_5$={A,F} = {1,0}=50%
$\hat{p}_6$={B,C} = {1,0}=50%
$\hat{p}_7$={B,D} = {1,0}=50%
$\hat{p}_8$ ={B,E} = {1,0}=50%
$\hat{p}_9$ ={B,F} = {1,0}=50%
$\hat{p}_{10}$ ={C,D} = {0,0}=0%
$\hat{p}_{11}$ ={C,E}= {0,0}=0%
$\hat{p}_{12}$={C,F} {0,0}=0%
$\hat{p}_{13}$ ={D,E}{0,0}=0%
$\hat{p}_{14}$={D,F}{0,0}=0%
$\hat{p}_{15}$={E,F} {0,0}=0%

mean of sample proportions = 0.3333 = 33.33%.
standard deviation of sample proportions = 0.3333 = 33.33%.

Frequency table

$\hat{p}$	Frequency	Relative Frequency
0	6	6/15
1/2	8	8/15
1	1	1/15

Sampling distribution of $\hat{p}$ when $n = 2$.



Predicting outcome of the election

Proportion of times we would declare Bert lost the election using this procedure$= \frac{6}{15} = 40\%$.

Problem (revisited)

Next, we are going to explore what happens if we increase our sample size. Now, instead of taking samples of size 2 we are going to draw samples of size 3.

List of all possible samples

{A,B,C}	{A,C,E}	{B,C,D}	{B,E,F}
{A,B,D}	{A,C,F}	{B,C,E}	{C,D,E}
{A,B,E}	{A,D,E}	{B,C,F}	{C,D,F}
{A,B,F}	{A,D,F}	{B,D,E}	{C,E,F}
{A,C,D}	{A,E,F}	{B,D,F}	{D,E,F}

List of all possible estimates

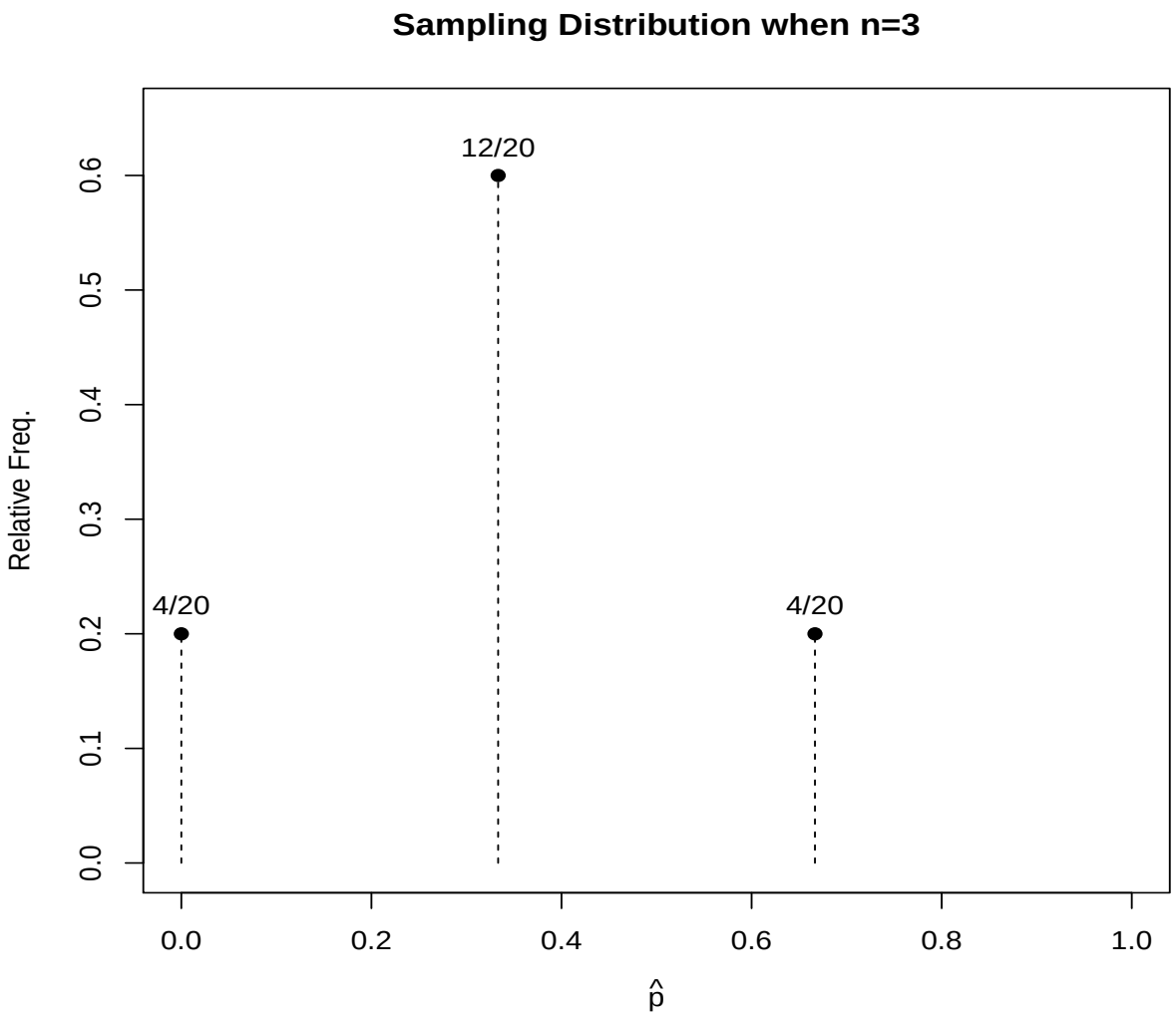
$$
\begin{array}{cccc}
\hat{p}_1 = 2/3 & \hat{p}_6 = 1/3 & \hat{p}_{11} = 1/3 & \hat{p}_{16} = 1/3 \\
\hat{p}_2 = 2/3 & \hat{p}_7 = 1/3 & \hat{p}_{12} = 1/3 & \hat{p}_{17} = 0 \\
\hat{p}_3 = 2/3 & \hat{p}_8 = 1/3 & \hat{p}_{13} = 1/3 & \hat{p}_{18} = 0 \\
\hat{p}_4 = 2/3 & \hat{p}_9 = 1/3 & \hat{p}_{14} = 1/3 & \hat{p}_{19} = 0 \\
\hat{p}_5 = 1/3 & \hat{p}_{10} = 1/3 & \hat{p}_{15} = 1/3 & \hat{p}_{20} = 0
\end{array}
$$

mean of sample proportions $= 0.3333 = 33.33\%$.
standard deviation of sample proportions $= 0.2163 = 21.63\%$.

Frequency table

$\hat{p}$	Frequency	Relative Frequency
0	4	4/20
1/3	12	12/20
2/3	4	4/20

Sampling distribution of $\hat{p}$ when $n = 3$.



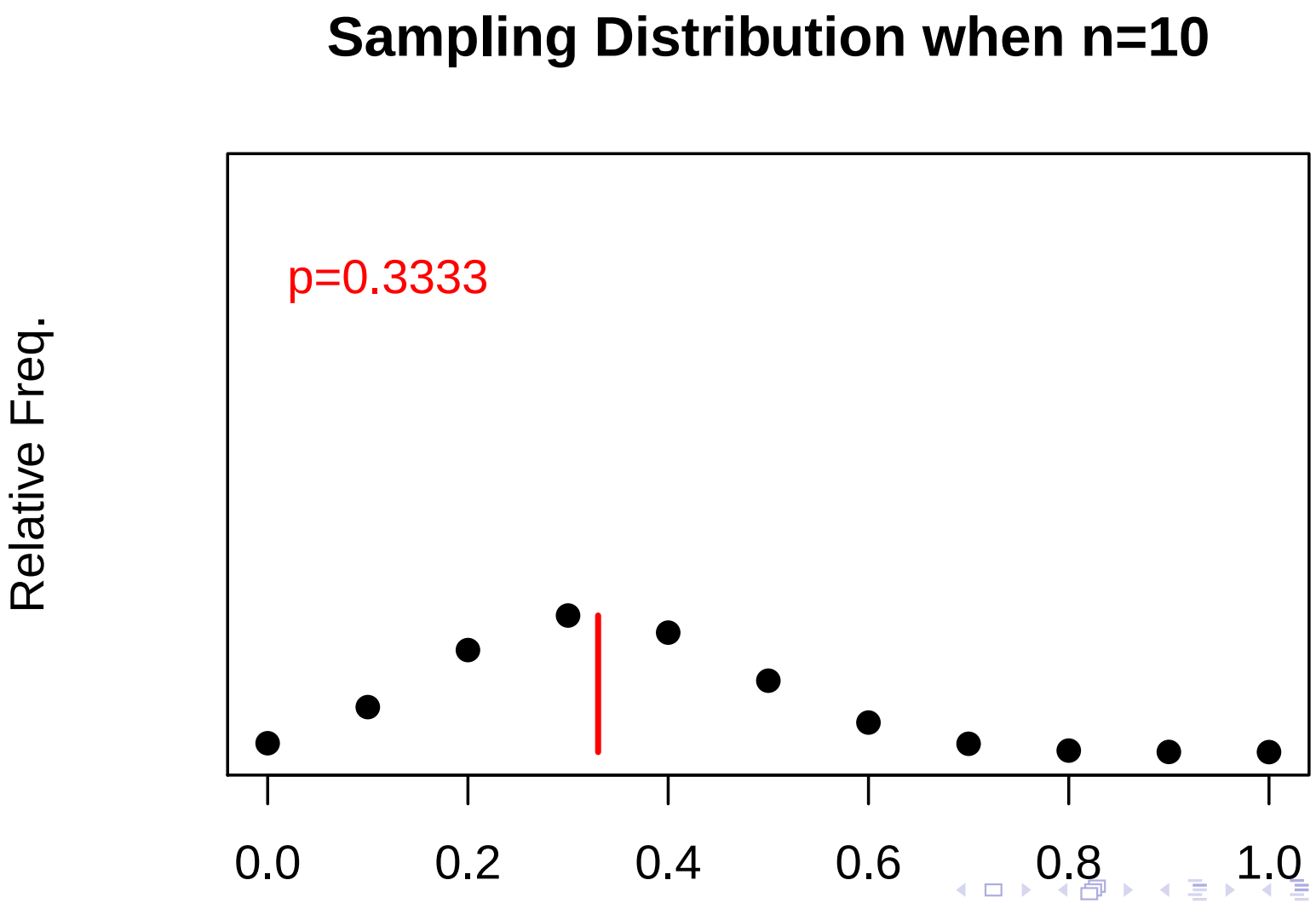
Prediction outcome of the election

Proportion of times we would declare Bert lost the election using this procedure$= \frac{16}{20} = 80\%$.

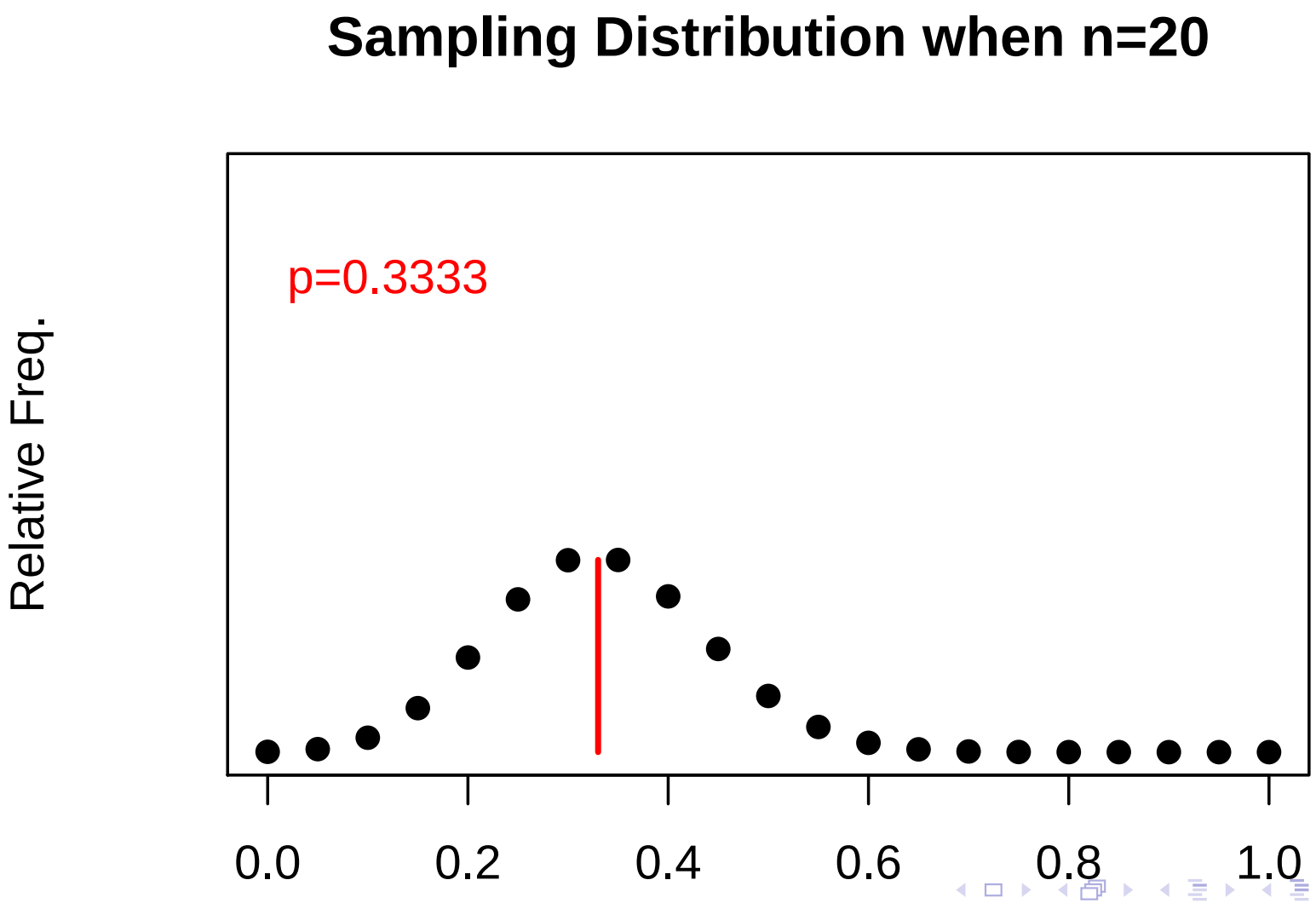
More realistic example

Assume we have a population with a total of 1200 individuals. All of them voted for one of two candidates: Bert or Ernie. Four hundred of them voted for Bert and the remaining 800 people voted for Ernie. Thus, the proportion of votes for Bert, which we will denote with p, is $p = \frac{400}{1200} = 33.33\%$. We are interested in estimating the proportion of people who voted for Bert, that is p, using information coming from an exit poll. Our ultimate goal is to see if we could use this procedure to predict the outcome of this election.

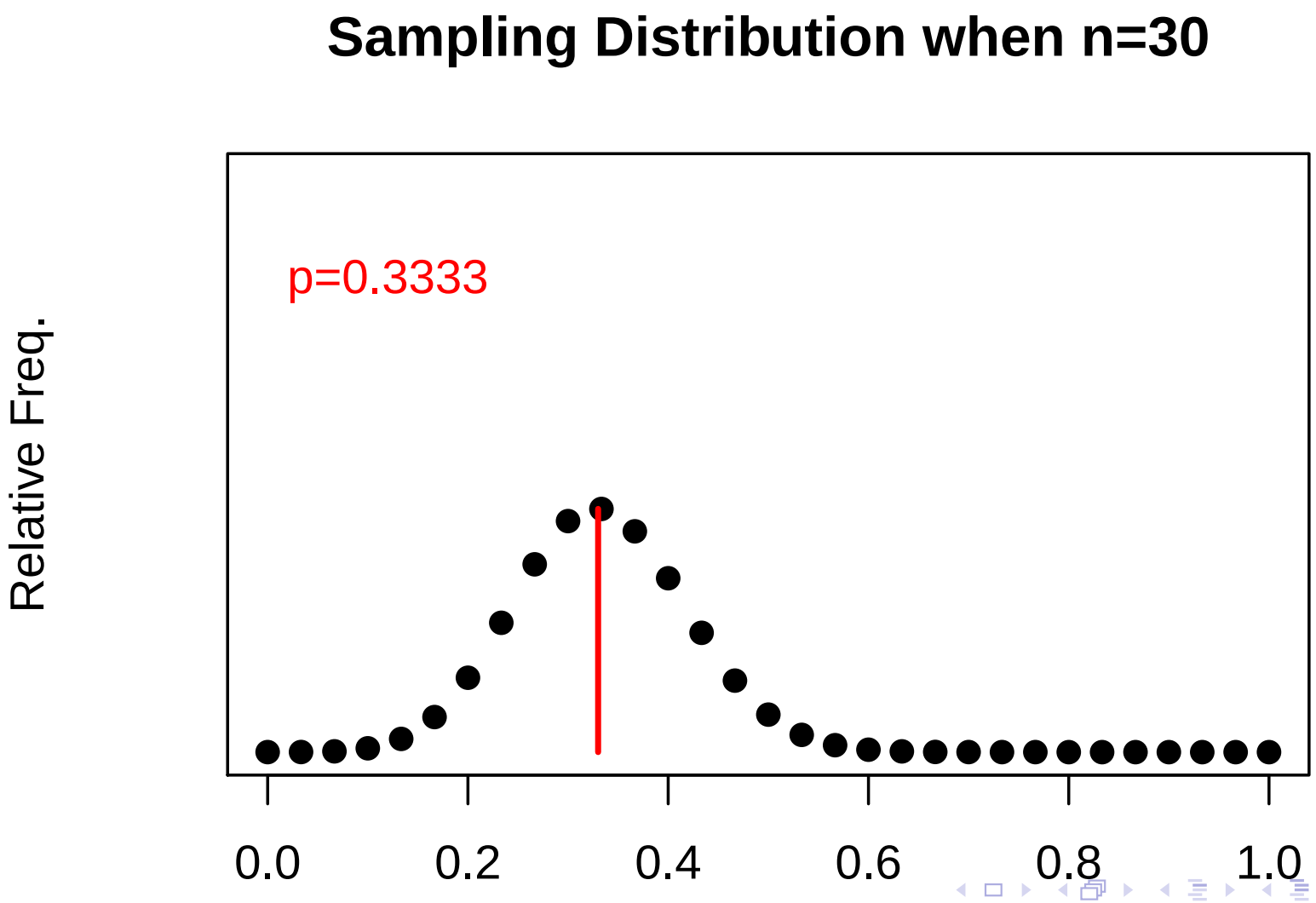
Sampling distribution of $\hat{p}$ when $n = 10$.



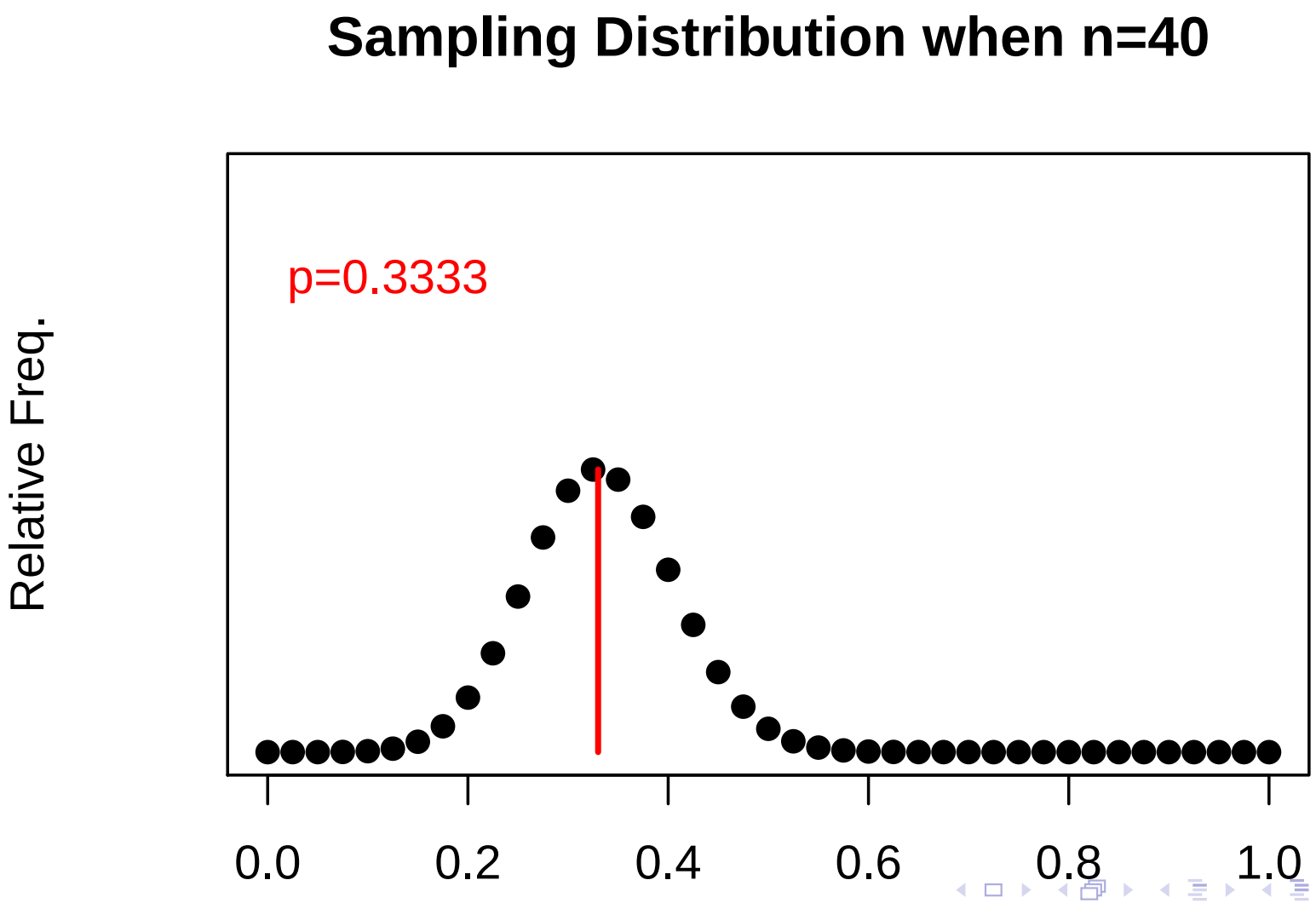
Sampling distribution of $\hat{p}$ when $n = 20$.



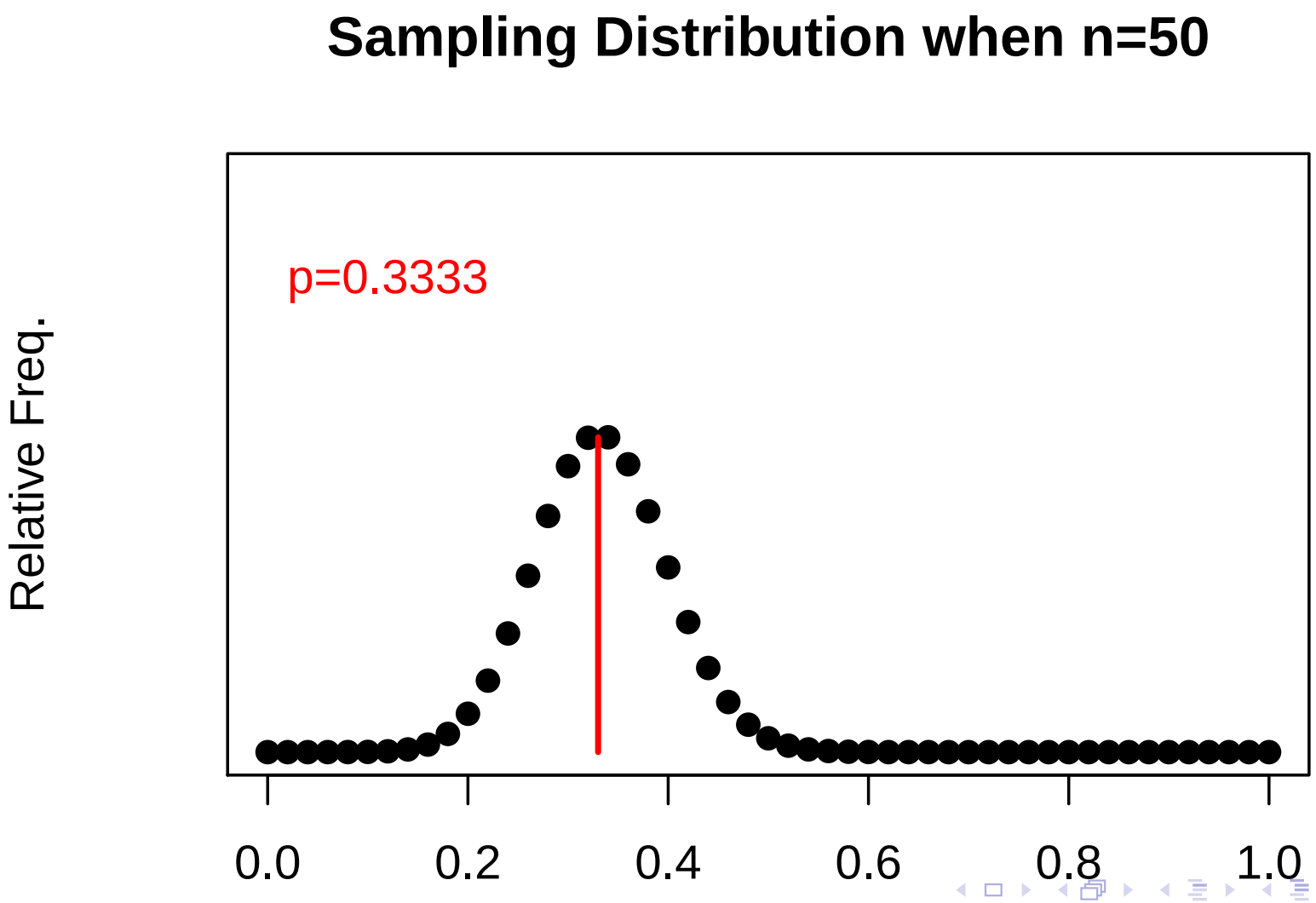
Sampling distribution of $\hat{p}$ when $n = 30$.



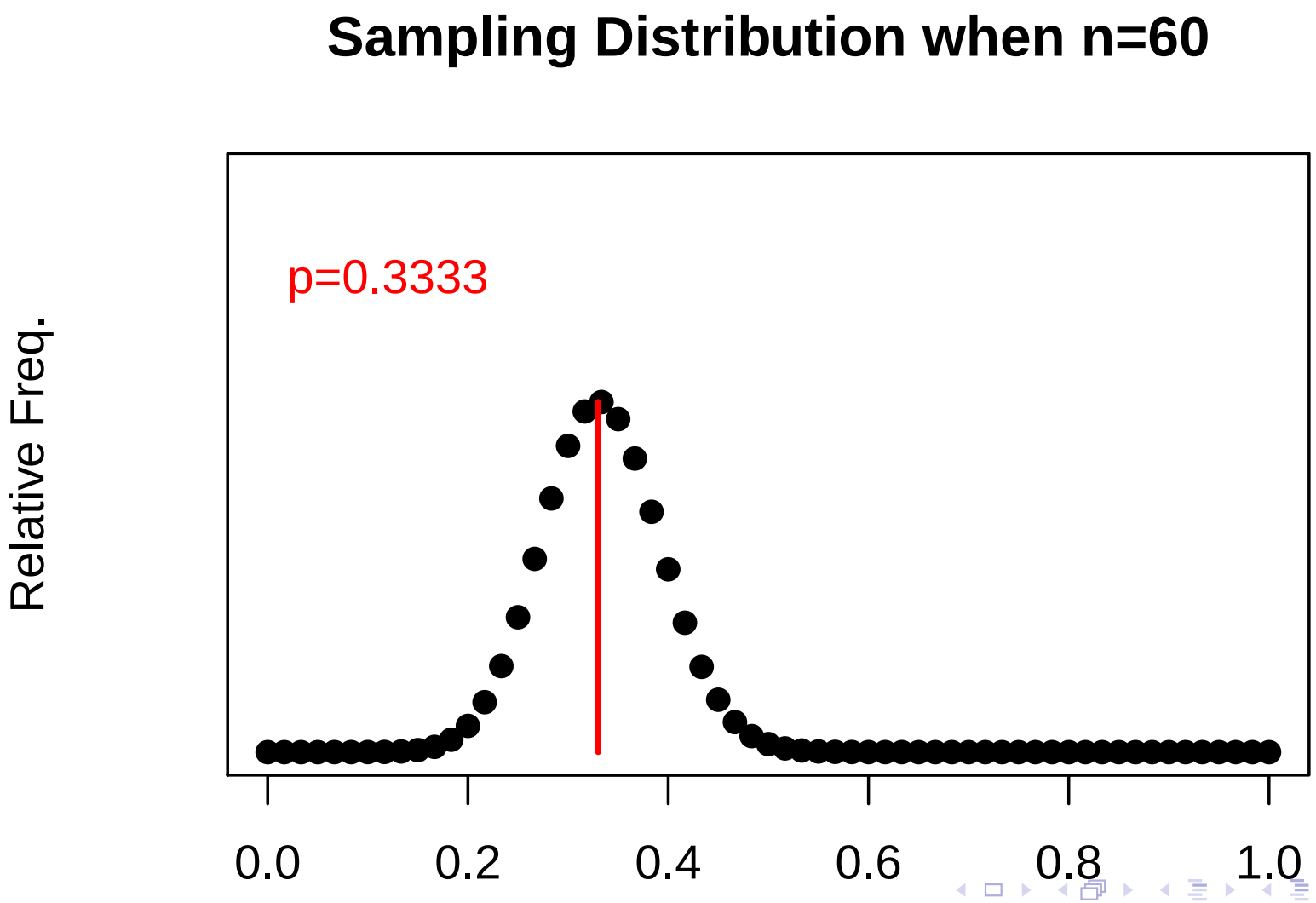
Sampling distribution of $\hat{p}$ when $n = 40$.



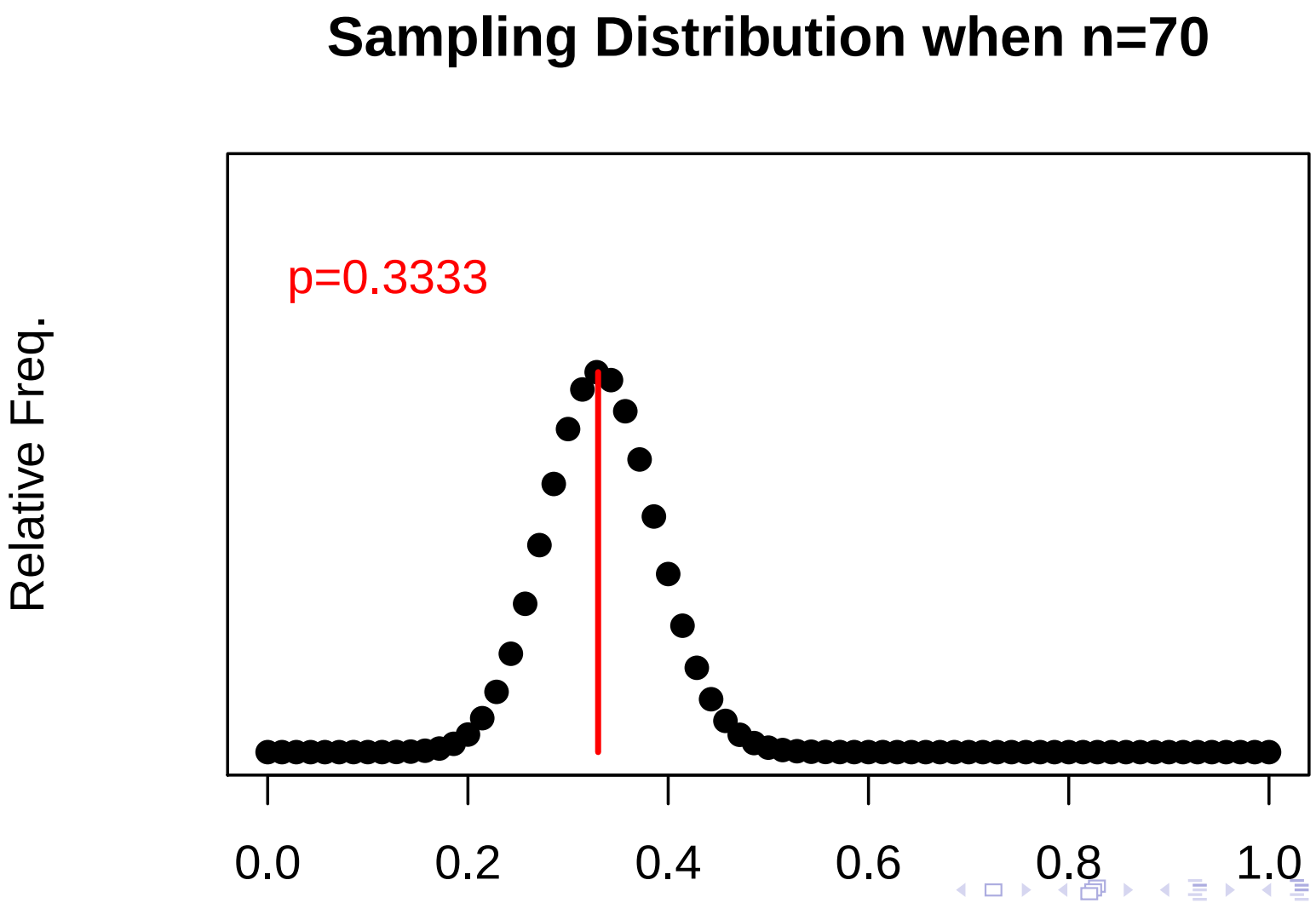
Sampling distribution of $\hat{p}$ when $n = 50$.



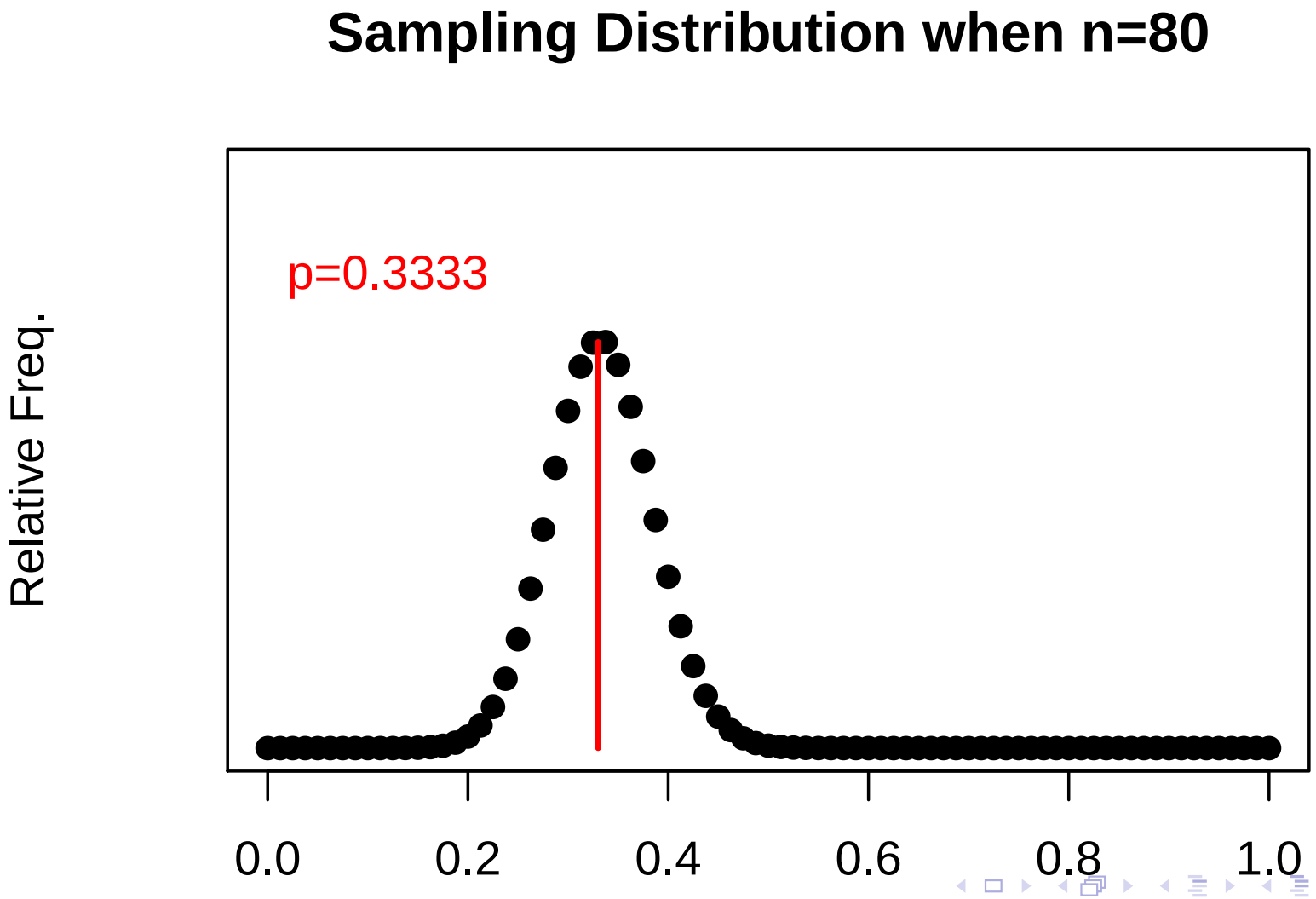
Sampling distribution of $\hat{p}$ when $n = 60$.



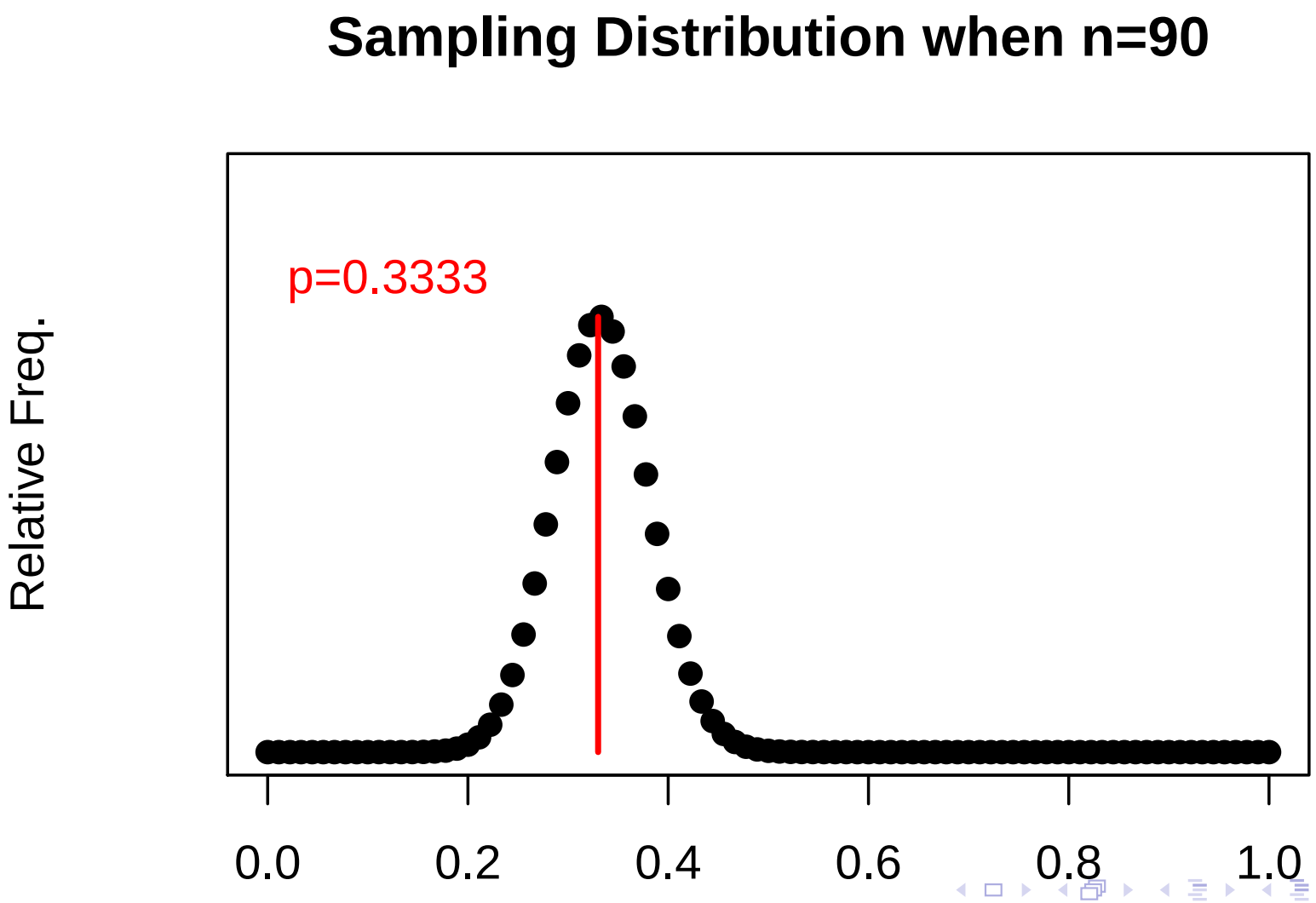
Sampling distribution of $\hat{p}$ when $n = 70$.



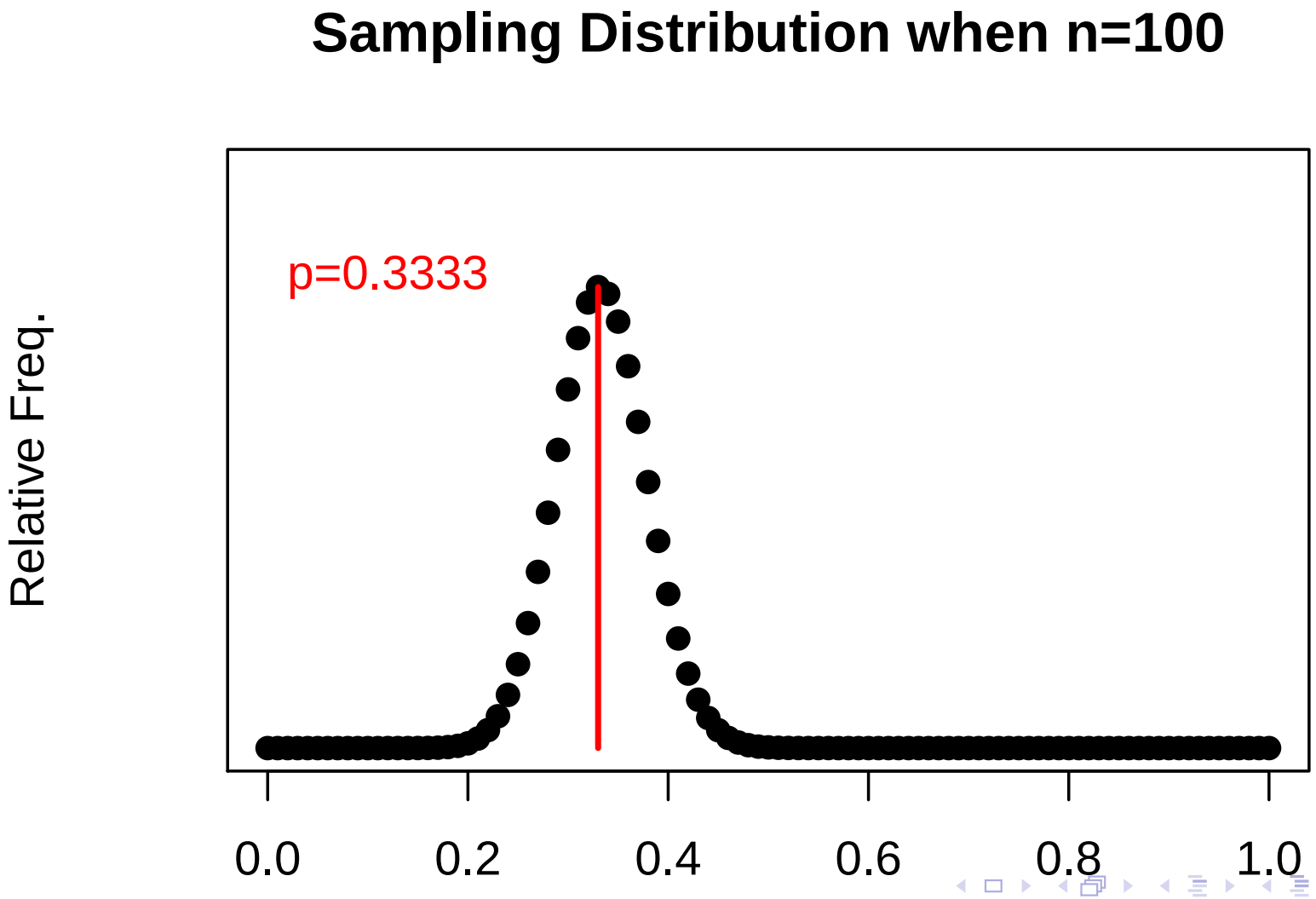
Sampling distribution of $\hat{p}$ when $n = 80$.



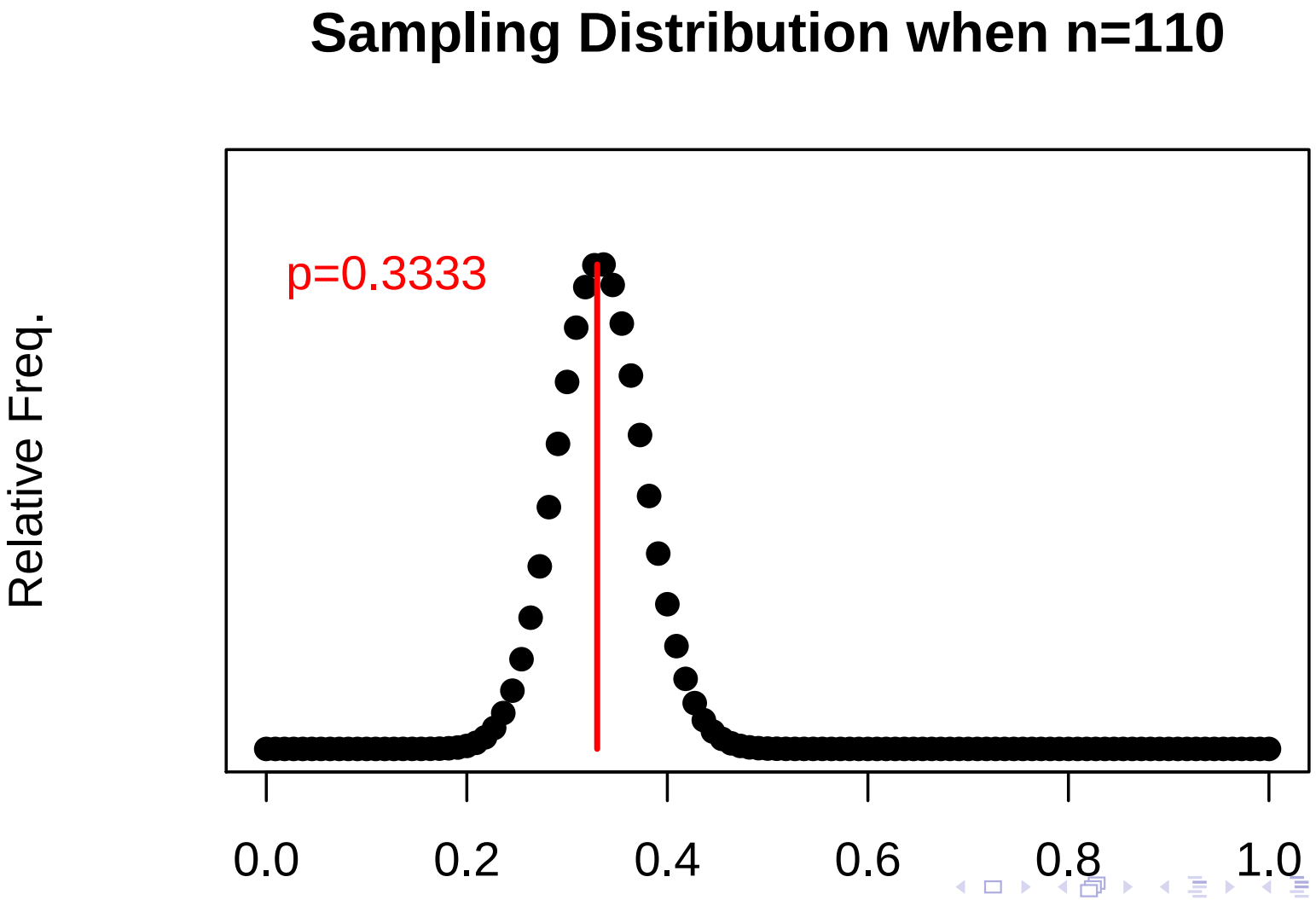
Sampling distribution of $\hat{p}$ when $n = 90$.



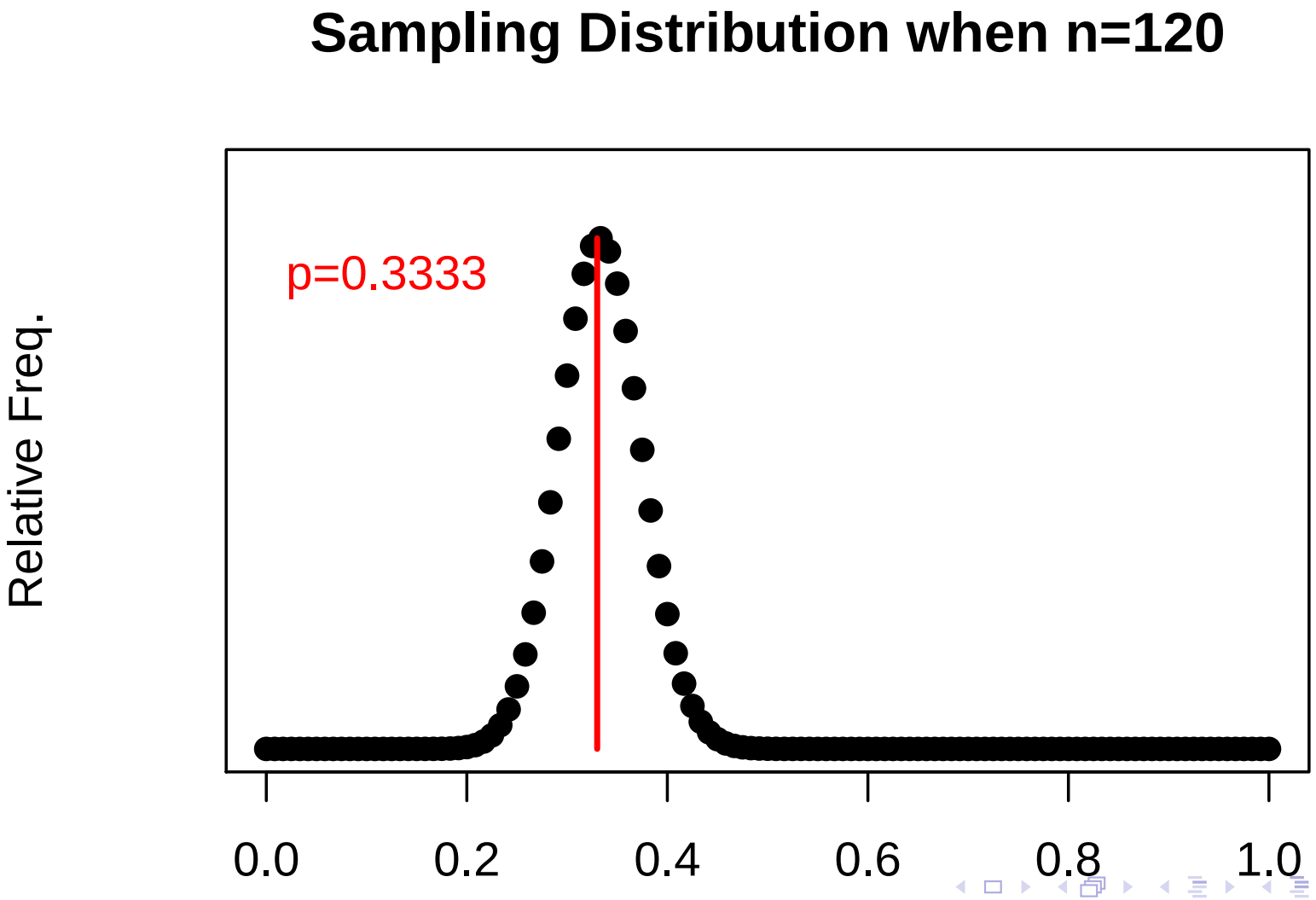
Sampling distribution of $\hat{p}$ when $n = 100$.



Sampling distribution of $\hat{p}$ when $n = 110$.



Sampling distribution of $\hat{p}$ when $n = 120$.



Observation

The larger the sample size, the more closely the distribution of sample proportions approximates a Normal distribution.

The question is: Which Normal distribution?

Sampling Distribution of a sample proportion

Draw an SRS of size n from a large population that contains proportion p of "successes". Let $\hat{p}$ be the **sample proportion** of successes,

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

Then:

- The **mean** of the sampling distribution of $\hat{p}$ is p.
- The **standard deviation** of the sampling distribution is

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

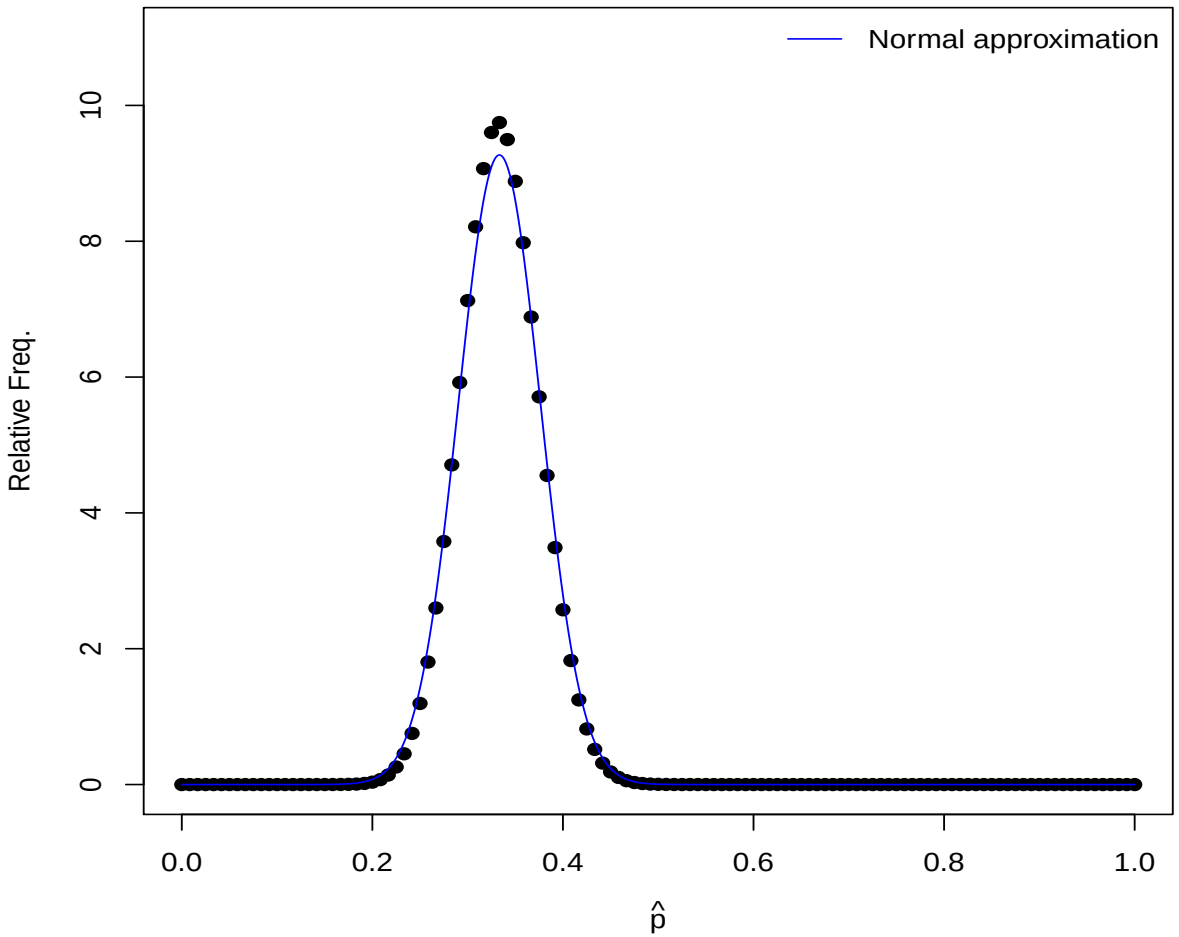
- As the sample size increases, the sampling distribution of $\hat{p}$ becomes **approximately Normal.** That is, for large n, $\hat{p}$ has approximately the $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ distribution.

Approximating Sampling Distribution of $\hat{p}$

If the proportion of **all** voters that supports Bert is $p = \frac{1}{3} = 33.33\%$ and we are taking a random sample of size 120, the Normal distribution that approximates the sampling distribution of $\hat{p}$ is:

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \text{ that is } N\left(\mu = 0.3333, \sigma = 0.0430\right) \qquad (1)$$

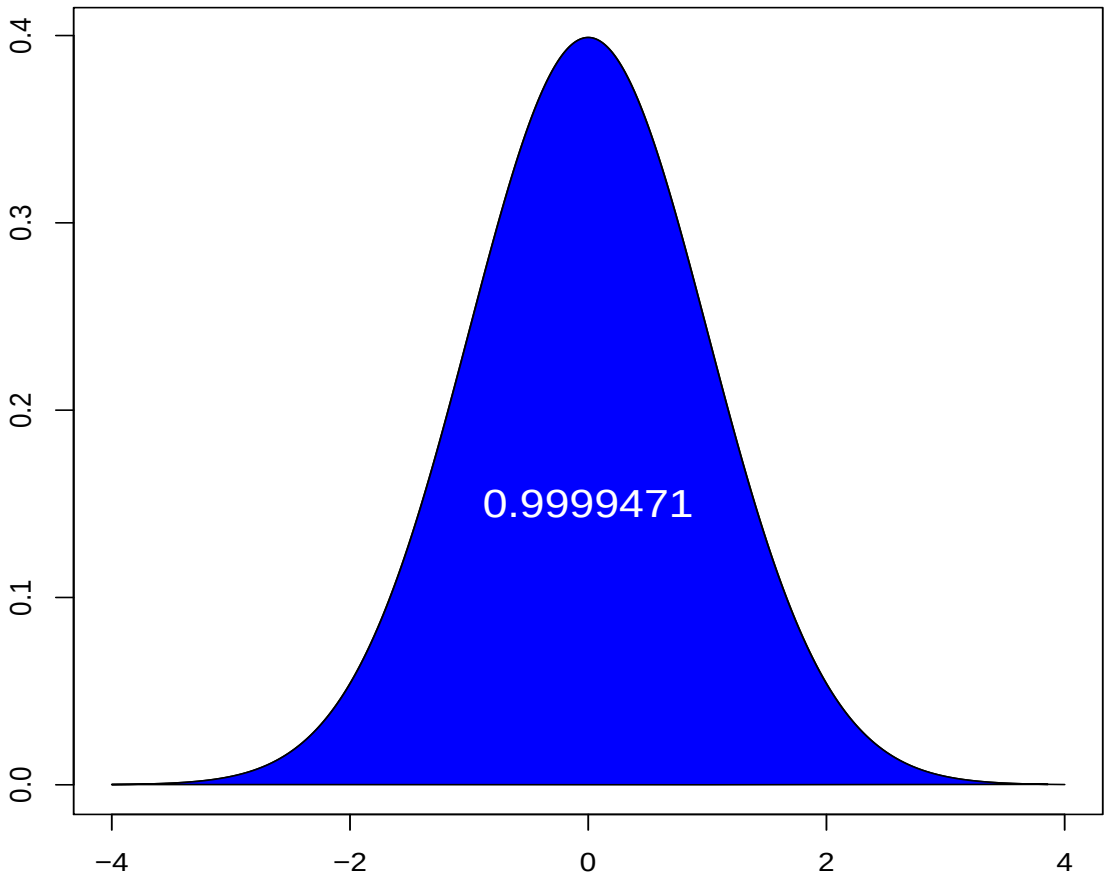
Sampling Distribution of $\hat{p}$ vs Normal Approximation



Predicting outcome of the election with our approximation

Proportion of times we would declare Bert lost the election using this procedure = Proportion of samples that yield a $\hat{p} < 0.50$.
Let $Y = \hat{p}$, then Y has a Normal Distribution with $\mu = 0.3333$ and $\sigma = 0.0430$.
Proportion of samples that yield a $\hat{p} < 0.50$=
$P(Y < 0.50) = P\left(\frac{Y-\mu}{\sigma} < \frac{0.5-0.3333}{0.0430}\right) = P(Z < 3.8767)$.

$P(Z < 3.8767)$



Predicting outcome of the election with our approximation

This implies that roughly 99.99% of the time taking a random exit poll of size 120 from a population of size 1200 will predict the outcome of the election correctly, when $p = 33.33\%$.

A few remarks

What is a Sampling Distribution?
It is the distribution that results when we find the proportions ($\hat{p}$) in **all** possible samples of a given size.
How do we construct a Sampling Distribution?

- Finding **all** possible samples of the size selected.
- Computing **statistic** of interest (sample proportion, for instance).
- Making a table of relative frequencies (or a graphical representation of it).

A few remarks

Why do we care about Sampling Distributions?

- It is impractical or too expensive to survey every individual in the population.
- It is reasonable to consider the idea of using a random sample to estimate a parameter.
- Sampling distributions help us to understand the behavior of a statistic when random sampling is used.

Example

In the last election, a state representative received 52% of the votes cast. One year after the election, the representative organized a survey that asked a random sample of 300 people whether they would vote for him in the next election. If we assume that his popularity has not changed, what is the probability that more than half of the sample would vote for him?

Solution

We want to determine the probability that the sample proportion is greater than 50%. In other words, we want to find $P(\hat{p} > 0.50)$.
We know that the sample proportion $\hat{p}$ is roughly Normally distributed with mean $p = 0.52$ and standard deviation $\sqrt{p(1-p)/n} = \sqrt{(0.52)(0.48)/300} = 0.0288$.
Thus, we calculate

$$P(\hat{p} > 0.50) = P\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} > \frac{0.50-0.52}{0.0288}\right)$$
$$= P(Z > -0.69) = 1 - P(Z < -0.69)$$
$$= 1 - 0.2451 = 0.7549.$$

If we assume that the level of support remains at 52%, the probability that more than half the sample of 300 people would vote for the representative is 0.7549.

R code

Just type the following:

```
1- pnorm(0.50, mean = 0.52, sd = 0.0288);

## [1] 0.7562982
```

In this case, `pnorm` will give you the area to the left of 0.50, for a Normal distribution with mean 0.52 and standard deviation 0.0288.

Mean and Standard Deviation of a Sample Mean

Suppose that $\bar{x}$ is the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ. Then the sampling distribution of $\bar{X}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Application

Many variables important to the real estate market are skewed, limited to only a few values or considered as categorical variables. Yet, marketing and business decisions are often made on means and proportions calculated over many homes. One reason these statistics are useful is the Central Limit Theorem.
Data on 1063 houses sold recently in the Saratoga, New York area are available at

```
"https://mcs.utm.utoronto.ca/~nosedal/datasets/real_estate
```

Let's investigate how the CLT guarantees that the sampling distribution of means of a quantitative variable approaches the Normal distribution (even when samples are drawn from populations that are far from Normal).

Application

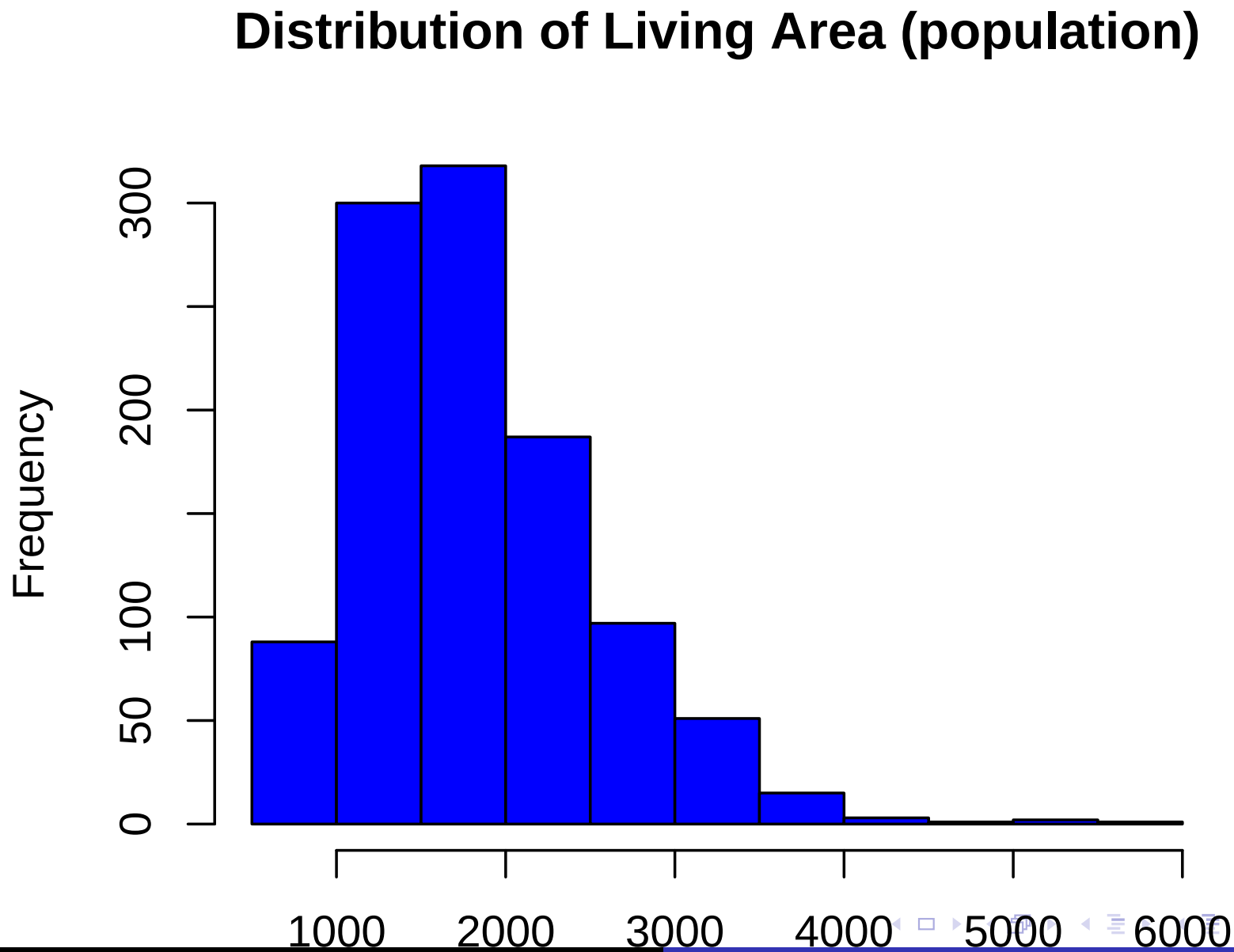
a) **Using R**, create an object (vector) called *areas* using the **entire population** of 1063 homes for the quantitative variable *Living.Area*. Then make a histogram for this quantitative variable *areas*. Describe the distribution (including its mean and standard deviation).

Application

```
#Step 1. Entering data;

# import data in R;

# url of real_estate;
real_estate_url=
"https://mcs.utm.utoronto.ca/~nosedal/datasets/
real_estate.txt"

real_estate= read.table(real_estate_url,header=TRUE);

names(real_estate)

areas=real_estate$Living.Area;
```

Application

```
# Step 2. Making histogram;

hist(areas,
main="Distribution of Living Area (population)",
col="blue");

# Step 3. Numerical summaries;

fivenum(areas);

mean(areas);

sd(areas);
```


Distribution of Living Area (population)
Frequency
0
50
100
200
300
1000
2000
3000
4000
5000
6000

Numerical summaries (population)

```
## [1]  672.0 1343.5 1680.0 2242.0 5632.0
## [1] 1833.49
## [1] 689.605
```

Application

b) **Using R**, do the following:

- Draw 500 samples of size 100 from this population of homes and find the means of these samples. To do so, type the following commands in R:

```
vec.means=rep(NA,500);
for (i in 1:500){
        vec.means[i]=mean(sample(areas,100))
                }
```

- Find the mean and standard deviation of this vector of means.
- Make a histogram of these 500 means.

Solution b)

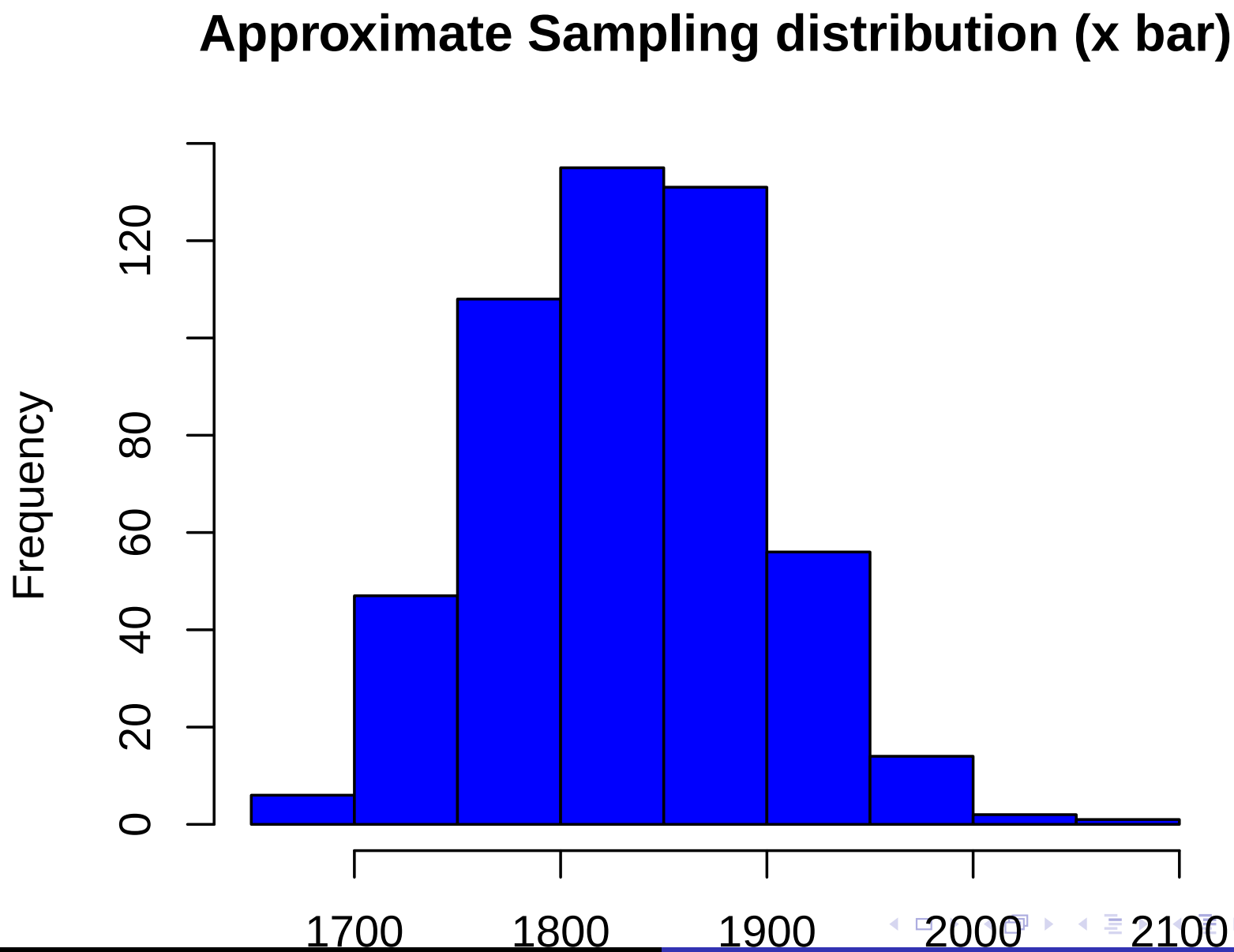
```
vec.means=rep(NA,500);
# we are creating a blank vector of means;
# we we will fill in this blank vector;

for (i in 1:500){
        vec.means[i]=mean(sample(areas,100))
                }

mean(vec.means);

sd(vec.means);
```

Solution b)

```
## [1] 1833.528
## [1] 64.12982
```

Histogram (vector of means)

```
hist(vec.means,
main="Approximate Sampling distribution (x bar)",
col="blue");
```


Approximate Sampling distribution (x bar)
Frequency
0
20
40
60
80
120
1700
1800
1900
2000
2100

Solution b) Again...

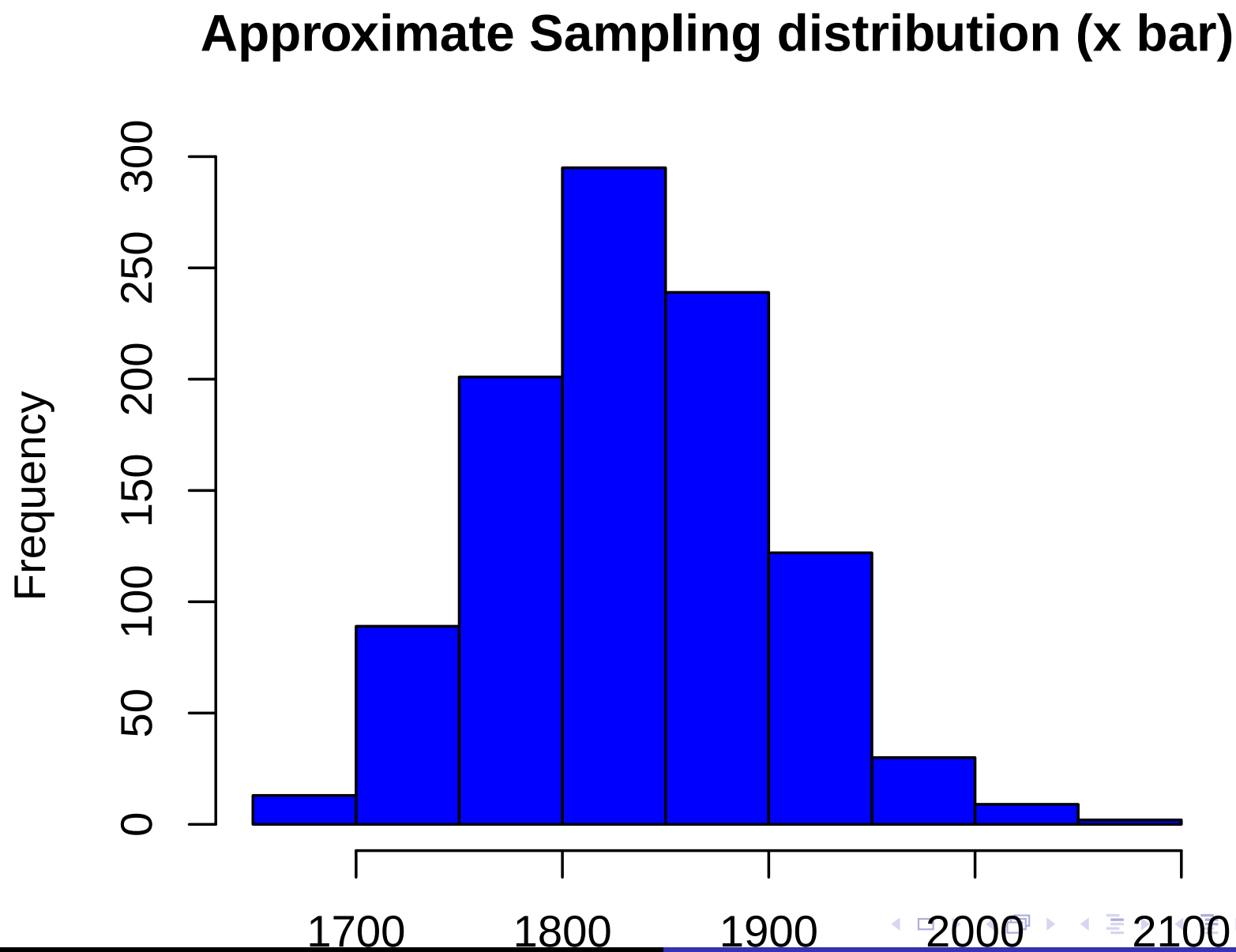
```
vec.means=rep(NA,1000);
# we are creating a blank vector of means;
# we we will fill in this blank vector;

for (i in 1:1000){
        vec.means[i]=mean(sample(areas,100))
                  }

mean(vec.means);

sd(vec.means);
```

Solution b)

```
## [1] 1834.942
## [1] 65.86645
```

Histogram (vector of means)

```
hist(vec.means,
main="Approximate Sampling distribution (x bar)",
col="blue");
```

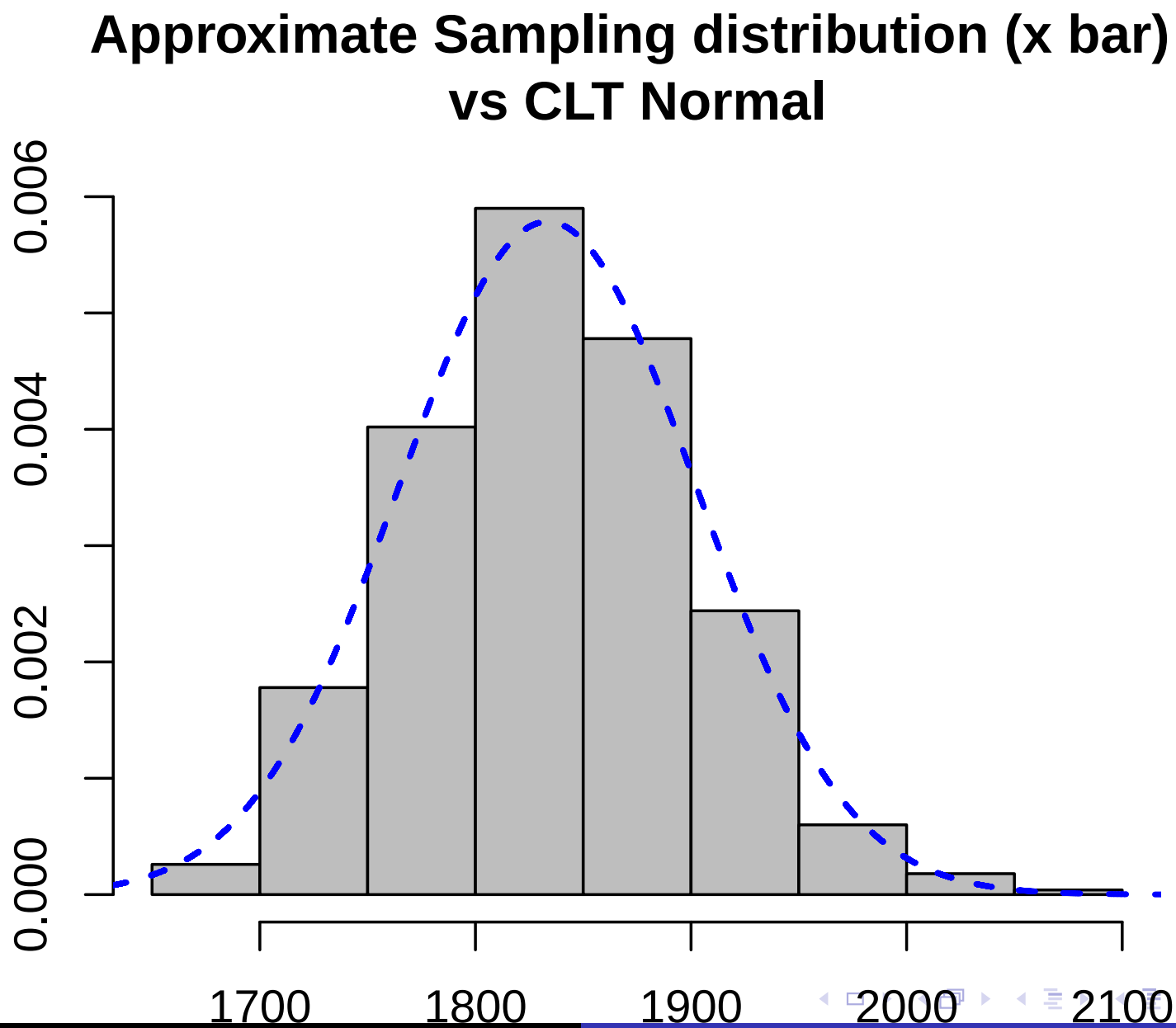

Approximate Sampling distribution (x bar)
Frequency
0
50
100
150
200
250
300
1700
1800
1900
2000
2100

Central Limit Theorem

Draw an SRS of size n from any population with mean μ and standard deviation σ. The Central Limit Theorem (CLT) says that when n is large the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$\bar{X}$ is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

The Central Limit Theorem allows us to use Normal probability calculations to answer questions about sample means from many observations.


Approximate Sampling distribution (x bar)
vs CLT Normal
0.000
0.002
0.004
0.006
1700
1800
1900
2000
2100

Example

A manufacturer of automobile batteries claims that the distribution of the lengths of life of its best battery has a mean of 54 months and a standard deviation of 6 months. Suppose a consumer group decides to check the claim by purchasing a sample of 50 of the batteries and subjecting them to tests that estimate the battery's life.

a) Assuming that the manufacturer's claim is true, describe the sampling distribution of the mean lifetime of a sample of 50 batteries.

b) Assuming that the manufacturer's claim is true, what is the probability that the consumer group's sample has a mean life of 52 or fewer months?

Solution a)

We can use the Central Limit Theorem to deduce that the sampling distribution for a sample mean lifetime of 50 batteries is approximately Normally distributed. Furthermore, the mean of this sampling distribution ($\mu_{\bar{X}}$) is 54 months according to the manufacturer's claim. Finally, the standard deviation of the sampling distribution is given by

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.8485 \text{ month}$$

Solution b)

If the manufacturer's claim is true, the probability that the consumer group observes a mean battery life of 52 or fewer months for its sample of 50 batteries is given by
$P(\bar{X} \leq 52)$, where $\bar{X}$ is Normally distributed, $\mu_{\bar{X}} = 54$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.8485$. Hence,

$$\begin{aligned} P(\bar{X} \leq 52) &= P\left(\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{52-54}{0.8485}\right) \\ &= P\left(Z \leq -2.3571008\right) \\ &\approx P\left(Z \leq -2.36\right) \end{aligned}$$

(from Table 3)

$$= 0.0091$$

Example

The number of accidents per week at a hazardous intersection varies with mean 2.2 and standard deviation 1.4. This distribution takes only whole-number values, so it is certainly not Normal.
a) Let $\bar{x}$ be the mean number of accidents per week at the intersection during a year (52 weeks). What is the approximate distribution of $\bar{x}$ according to the Central Limit Theorem?
b) What is the approximate probability that $\bar{x}$ is less than 2?
c) What is the approximate probability that there are fewer than 100 accidents at the intersection in a year?

Solution

a) By the Central Limit Theorem, $\bar{X}$ is roughly Normal with mean $\mu^* = 2.2$ and standard deviation $\sigma^* = \sigma/\sqrt{n} = 1.4/\sqrt{52} = 0.1941$.

b) $P(\bar{X} < 2) = P\left(\frac{\bar{X}-\mu^*}{\sigma^*} < \frac{2-2.2}{0.1941}\right)$

$= P(Z < -1.0303) = 0.1515$

Solution

Let X_i be the number of accidents during week i.

c) $P(Total < 100) = P\left(\sum_{i=1}^{52} X_i < 100\right)$

$$= P\left(\frac{\sum_{i=1}^{52} X_i}{52} < \frac{100}{52}\right)$$

$$= P\left(\bar{X} < 1.9230\right)$$

$$= P(Z < -1.4270) = 0.0768$$