STA 260: Statistics and Probability II

Al Nosedal.
University of Toronto.

Winter 2017

"If you can't explain it simply, you don't understand it well enough"

Albert Einstein.

Theorem 7.5

Let Y and $Y_1, Y_2, Y_3, \ldots$ be random variables with moment-generating functions $M(t)$ and $M_1(t), M_2(t), \ldots$, respectively. If

$$\lim_{n\to\infty} M_n(t) = M(t) \qquad \text{for all real } t,$$

then the distribution function of Y_n converges to the distribution function of Y as $n \to \infty$.

Maclaurin Series

A Maclaurin series is a Taylor series expansion of a function about 0,

$$f(x) = f(0) + f^{'}(0)x + \frac{f^{''}(0)x^2}{2!} + \frac{f^3(0)x^3}{3!} + ... + \frac{f^n(0)x^n}{n!} + ...$$

Maclaurin series are named after the Scottish mathematician Colin Maclaurin.

Useful properties of MGFs

- $M_Y(0) = E(e^{0Y}) = E(1) = 1.$
- $M'_Y(0) = E(Y).$
- $M''_Y(0) = E(Y^2).$
- $M_{aY}(t) = E(e^{t(aY)}) = E(e^{(at)Y}) = M_Y(at)$, where a is a constant.

Central Limit Theorem

Let $Y_1, Y_2, ..., Y_n$ be independent and identically distributed random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$. Define

$$U_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
Then the distribution function of U_n converges to the standard Normal distribution function as $n \to \infty$. That is,

$$\lim_{n\to\infty} P(U_n \leq u) = \int_{\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \text{ for all u.}$$

Proof

Let $Z_i = \frac{X_i - \mu}{\sigma}$. Note that $E(Z_i) = 0$ and $V(Z_i) = 1$. Let us rewrite U_n

$$\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{n\sigma}\right) = \frac{1}{\sqrt{n}}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma}\right)$$

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i.$$

Since the mfg of the sum of independent random variables is the product of their individual mfgs, if $M_{Z_i}(t)$ denotes the mgf of each random variable Z_i

$$M_{\sum Z_i}(t) = [M_{Z_1}(t)]^n$$

and

$$M_{U_n} = M_{\sum Z_i}(t/\sqrt{n}) = \left[M_{Z_1}(t/\sqrt{n})\right]^n.$$

Recall that $M_{Z_i}(0) = 1$, $M'_{Z_i}(0) = E(Z_i) = 0$, and $M''_{Z_i}(0) = E(Z_i^2) = V(Z_i^2) = 1$.

Now, let us write the Taylor's series of $M_{Z_i}(t)$ at 0

$$M_{Z_i}(t) = M_{Z_i}(0) + tM'_{Z_i}(0) + \frac{t^2}{2!}M''_{Z_i}(0) + \frac{t^3}{3!}M'''_{Z_i}(0) + ...$$

$$M_{Z_i}(t) = 1 + \frac{t^2}{2} + \frac{t^3}{3!}M'''_{Z_i}(0) + ...$$

$$M_{U_n}(t) = \left[M_{Z_1}(t/\sqrt{n})\right]^n = \left[1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}}M'''_{Z_i}(0) + ...\right]^n$$

Recall that if

$$\lim_{n\to\infty} b_n = b \qquad \lim_{n\to\infty} \left(1 + \frac{b_n}{n}\right)^n = e^b$$

But

$$\lim_{n\to\infty} \left[\frac{t^2}{2} + \frac{t^3}{3!n^{1/2}} M'''_{Z_i}(0) + ...\right] = \frac{t^2}{2}$$

Therefore,

$$\lim_{n\to\infty} M_{U_n}(t) = exp\left(\frac{t^2}{2}\right)$$

which is the moment-generating function for a standard Normal random variable. Applying Theorem 7.5 we conclude that U_n has a distribution function that converges to the distribution function of the standard Normal random variable.

Example

An anthropologist wishes to estimate the average height of men for a certain race of people. If the population standard deviation is assumed to be 2.5 inches and if she randomly samples 100 men, find the probability that the difference between the sample mean and the true population mean will not exceed 0.5 inch.

Solution

Let $\bar{Y}$ denote the mean height and $\sigma = 2.5$ inches. By the Central Limit Theorem, $\bar{Y}$ has, roughly, a Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, that is $N(\mu, 2.5/10)$.

$P(|\bar{Y} - \mu| \leq 0.5) = P(-0.5 \leq \bar{Y} - \mu \leq 0.5)$

$= P(-\frac{(0.5)(10)}{2.5} \leq \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \leq \frac{(0.5)(10)}{2.5})$

$\approx P(-2 \leq Z \leq 2) = 0.95$

Example

The acidity of soils is measured by a quantity called the pH, which range from 0 (high acidity) to 14 (high alkalinity). A soil scientist wants to estimate the average pH for a large field by randomly selecting n core samples and measuring the pH in each sample. Although the population standard deviation of pH measurements is not known, past experience indicates that most soils have a pH value between 5 and 8. Suppose that the scientist would like the sample mean to be within 0.1 of the true mean with probability 0.90. How many core samples should the scientist take?

Solution

Let $\bar{Y}$ denote the average pH. By the Central Limit Theorem, $\bar{Y}$ has, roughly, a Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. By the way, the empirical rule suggests that the standard deviation of a set of measurements may be roughly approximated by one-fourth of the range. Which means that $\sigma \approx 3/4$.

We require

$$P(|\bar{Y} - \mu| \leq 0.1) \approx P\left(|Z| \leq \frac{\sqrt{n}(0.1)}{0.75}\right) = 0.90$$

Thus, we have that $\frac{\sqrt{n}(0.1)}{0.75} = 1.65$ (Using Table 4). So, $n = 153.1406$. Therefore, 154 core samples should be taken.

Example

Twenty-five lamps are connected in a greenhouse so that when one lamp fails, another takes over immediately. (Only one lamp is turned on at any time). The lamps operate independently, and each has a mean life of 50 hours and standard deviation of 4 hours. If the greenhouse is not checked for 1300 hours after the lamp system is turned on, what is the probability that a lamp will be burning at the end of the 1300-hour period?

Solution

Let Y_i denote the lifetime of the i-th lamp, $i = 1, 2, \ldots, 25$, and $\mu = 50$ and $\sigma = 4$. The random variable of interest is $Y_1 + Y_2 + \ldots + Y_{25} = \sum_{i=1}^{25} Y_i$ which is the lifetime of the lamp system.

$P(\sum_{i=1}^{25} Y_i \geq 1300) = P(\frac{\sum_{i=1}^{25} Y_i}{25} \geq \frac{1300}{25})$

$\approx P(Z \geq \frac{(5)(52-50)}{4}) = P(Z \geq 2.5) = 0.0062$ (using Table 4).

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

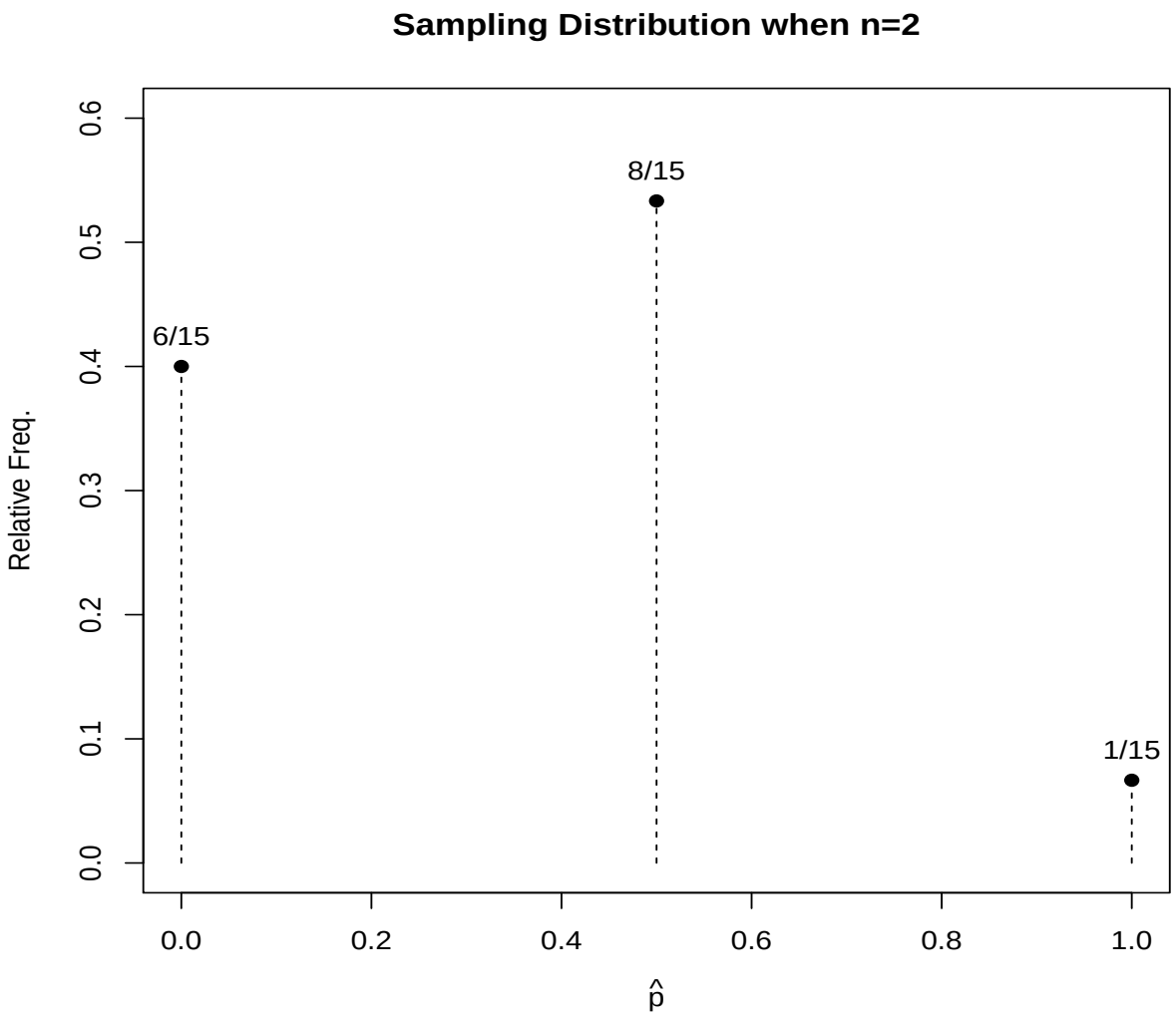
$\hat{p}_1$={A,B} = {1,1}=100%
$\hat{p}_2$={A,C} = {1,0}=50%
$\hat{p}_3$={A,D} = {1,0}=50%
$\hat{p}_4$ ={A,E}= {1,0}=50%
$\hat{p}_5$={A,F} = {1,0}=50%
$\hat{p}_6$={B,C} = {1,0}=50%
$\hat{p}_7$={B,D} = {1,0}=50%
$\hat{p}_8$ ={B,E} = {1,0}=50%
$\hat{p}_9$ ={B,F} = {1,0}=50%
$\hat{p}_{10}$ ={C,D} = {0,0}=0%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

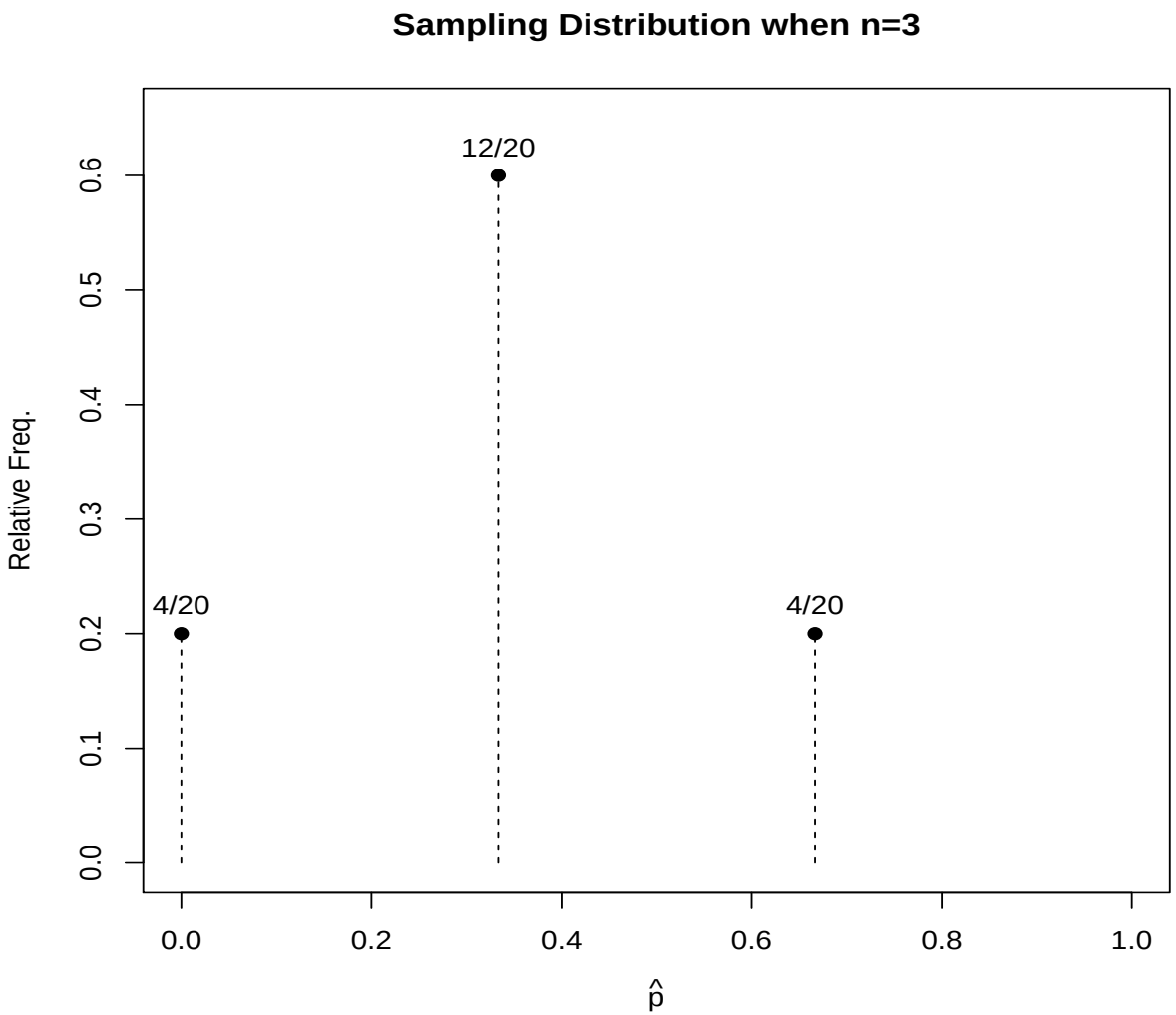
$$
\begin{array}{cccc}
\hat{p}_1 = 2/3 & \hat{p}_6 = 1/3 & \hat{p}_{11} = 1/3 & \hat{p}_{16} = 1/3 \\
\hat{p}_2 = 2/3 & \hat{p}_7 = 1/3 & \hat{p}_{12} = 1/3 & \hat{p}_{17} = 0 \\
\hat{p}_3 = 2/3 & \hat{p}_8 = 1/3 & \hat{p}_{13} = 1/3 & \hat{p}_{18} = 0 \\
\hat{p}_4 = 2/3 & \hat{p}_9 = 1/3 & \hat{p}_{14}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

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 $n = 3$.



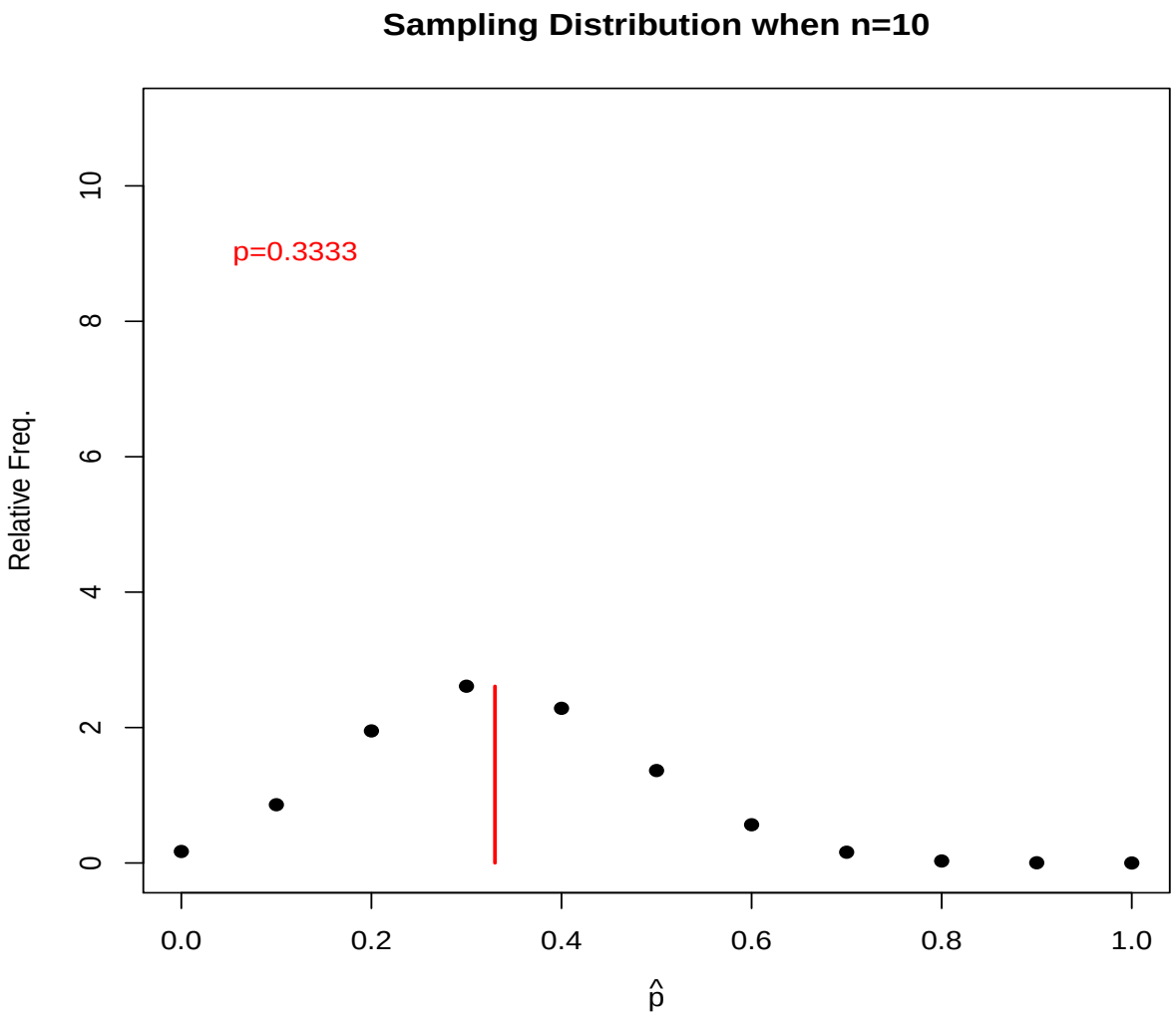
Prediction outcome of the election

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

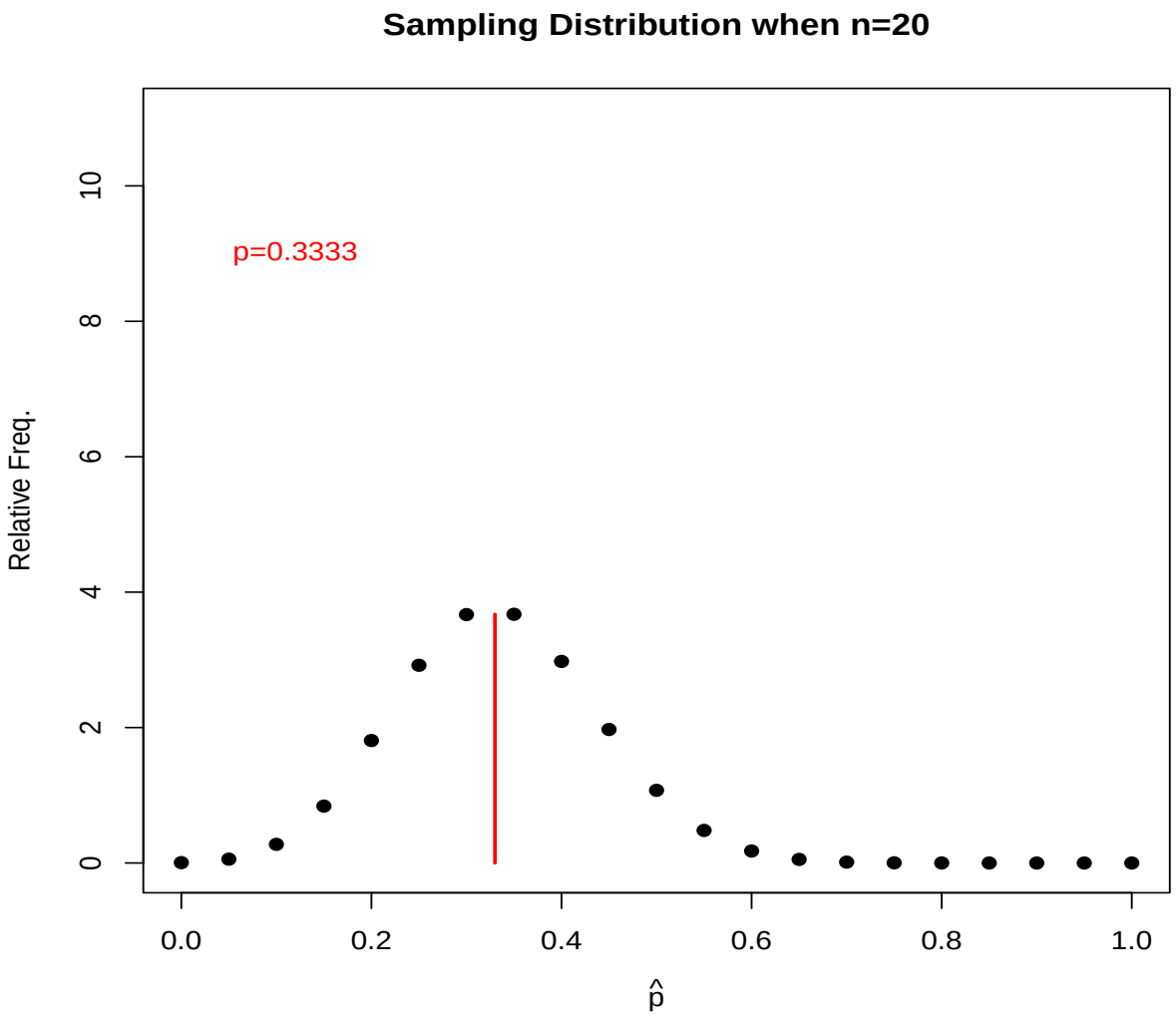
More realistic example

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.

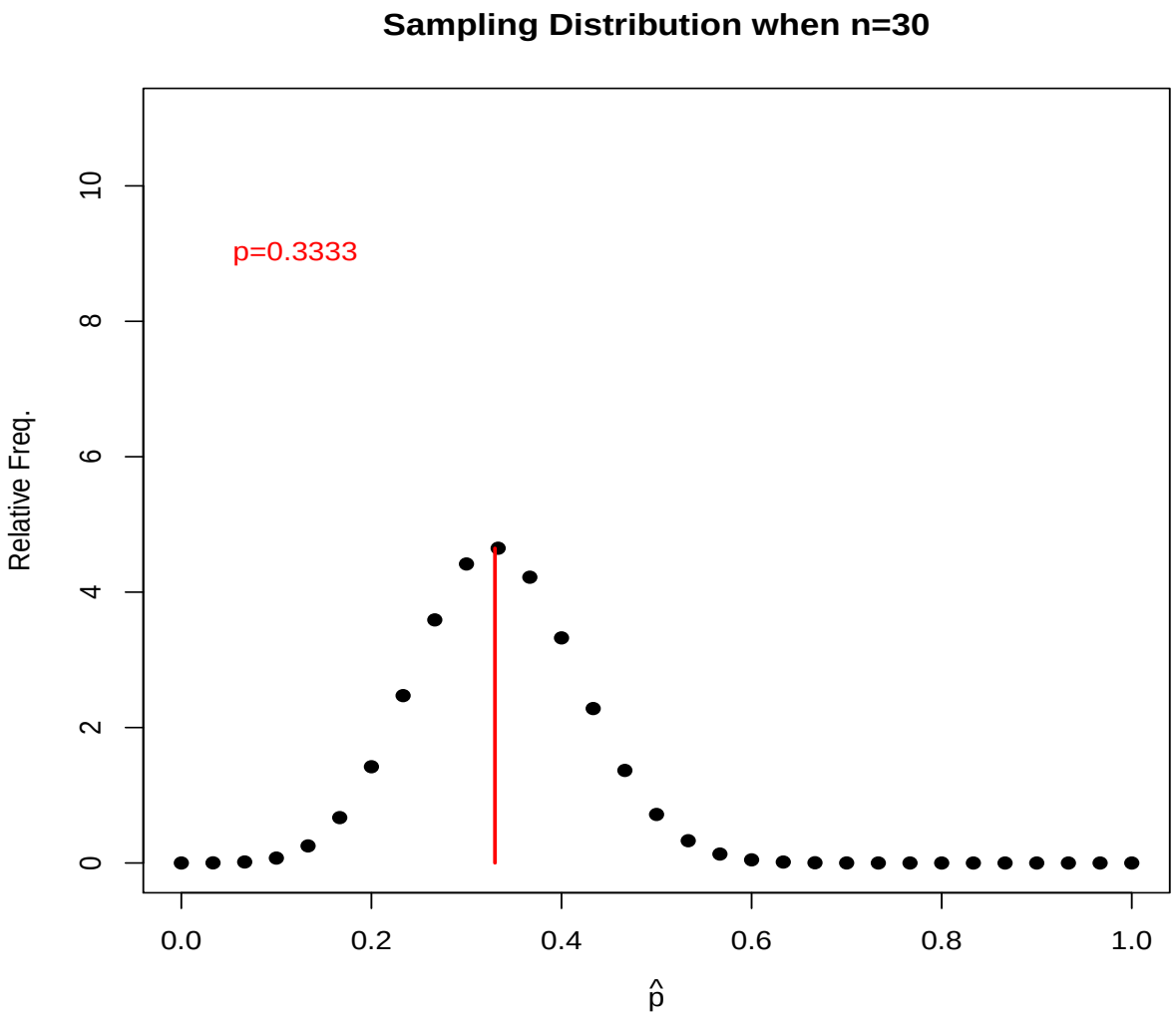
Sampling distribution of $\hat{p}$ when $n = 10$.



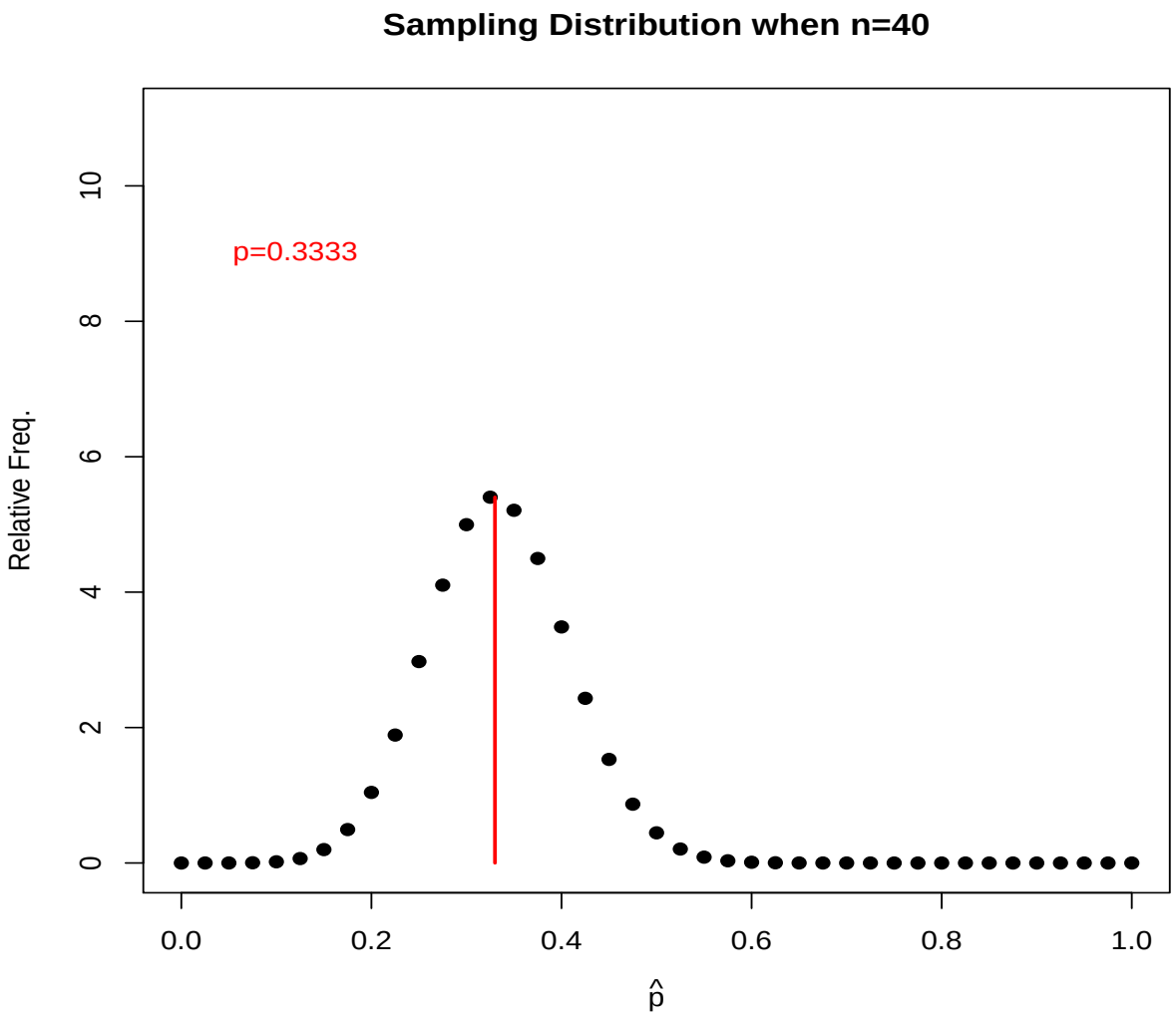
Sampling distribution of $\hat{p}$ when $n = 20$.



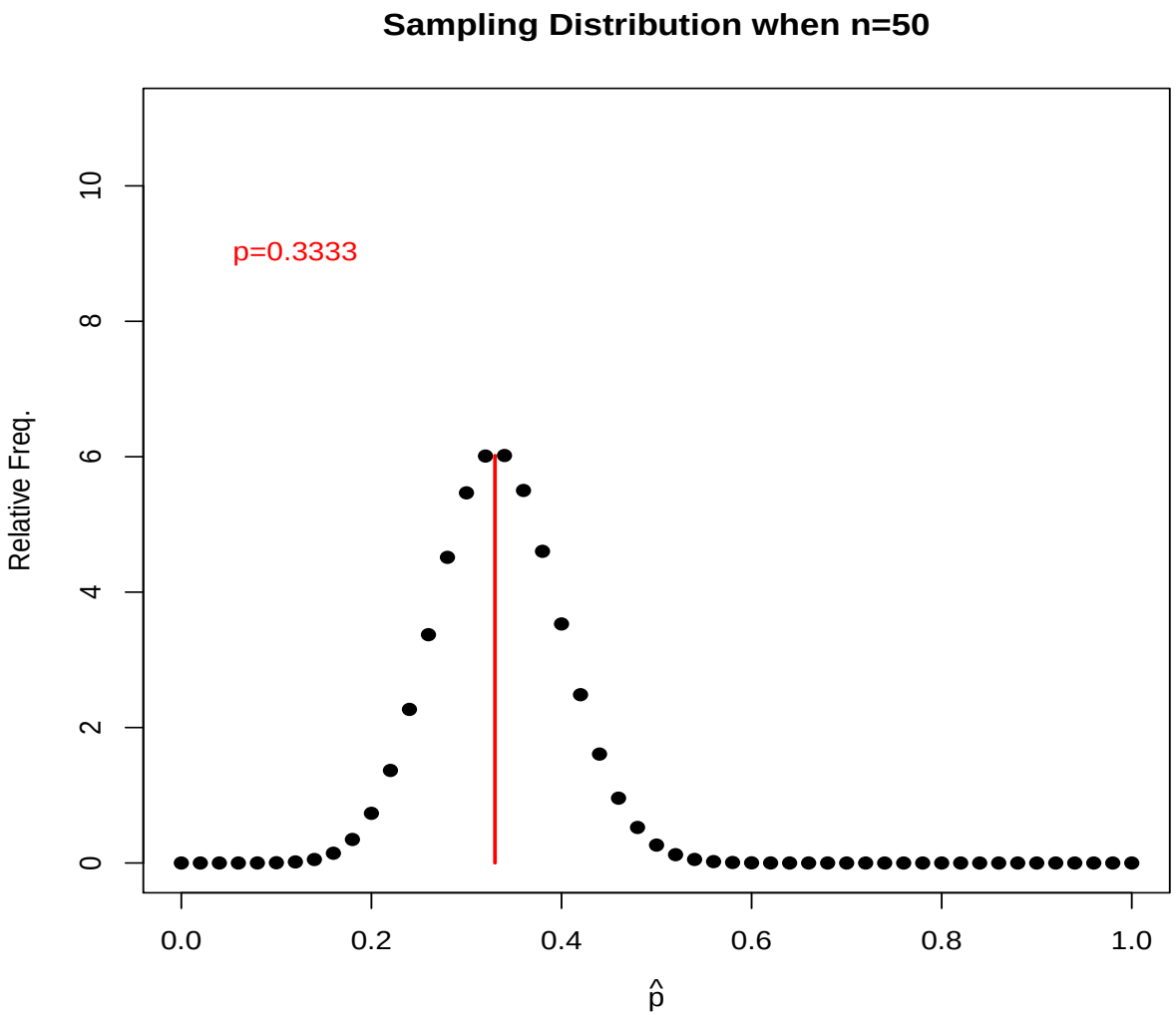
Sampling distribution of $\hat{p}$ when $n = 30$.



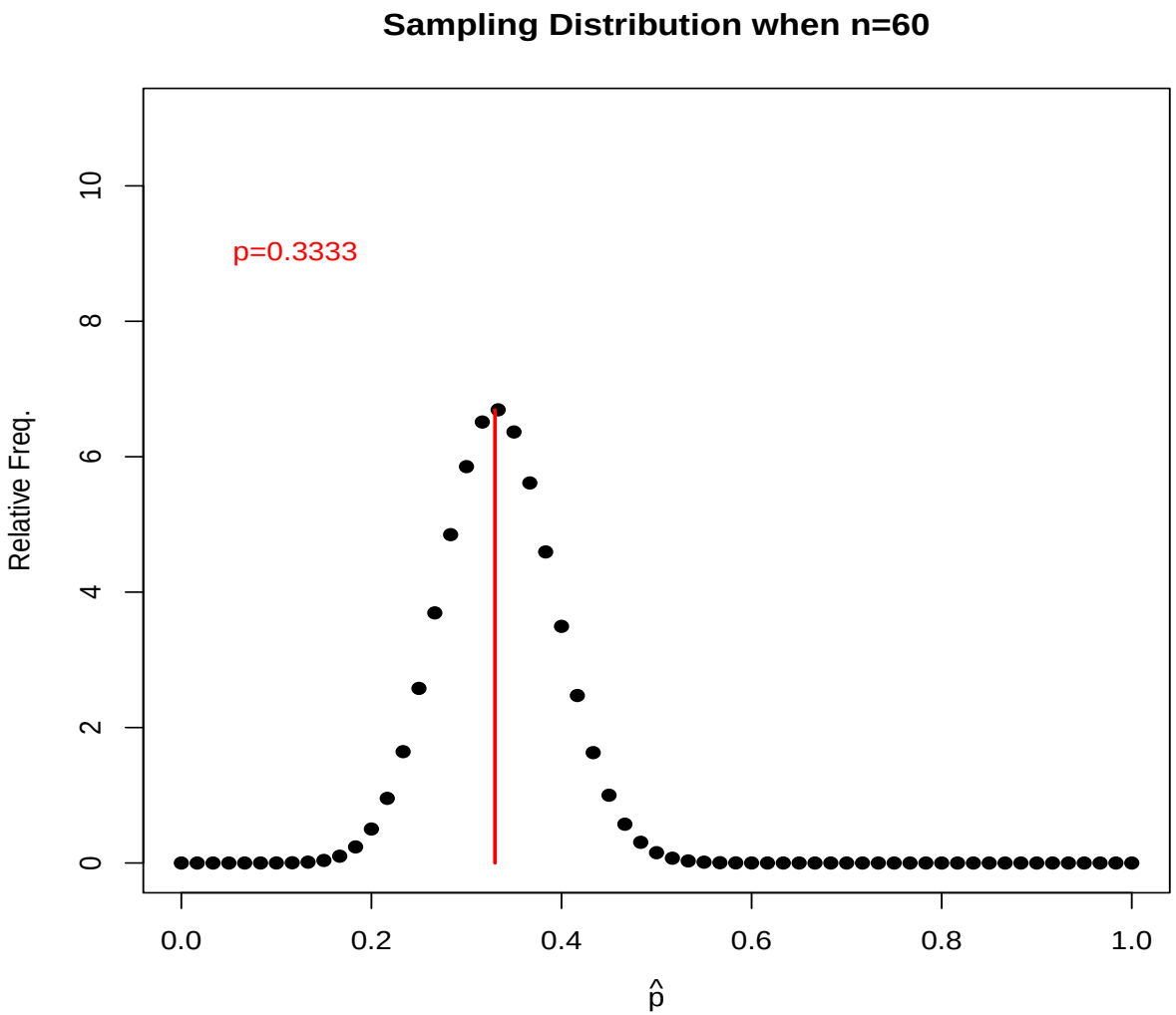
Sampling distribution of $\hat{p}$ when $n = 40$.



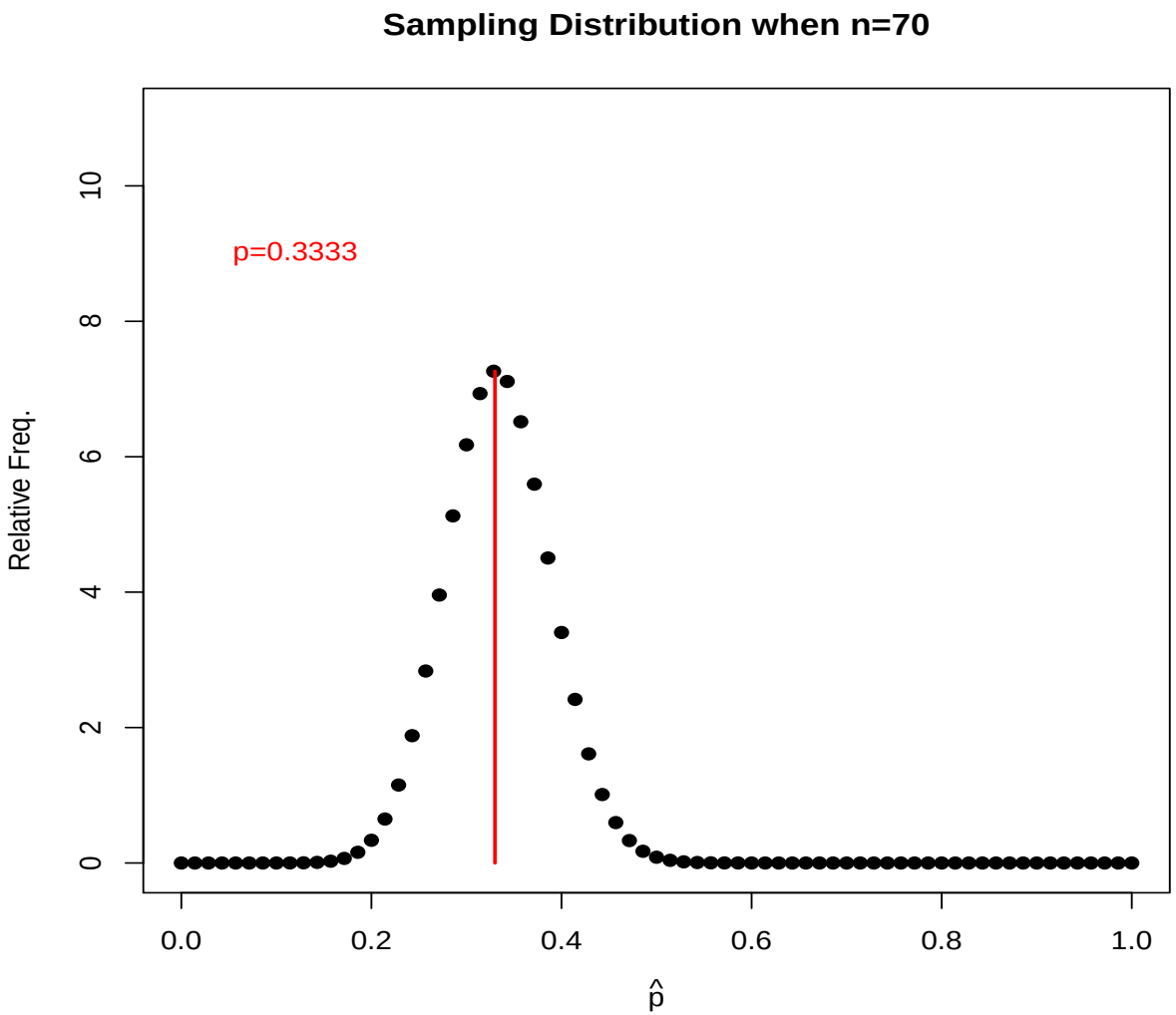
Sampling distribution of $\hat{p}$ when $n = 50$.



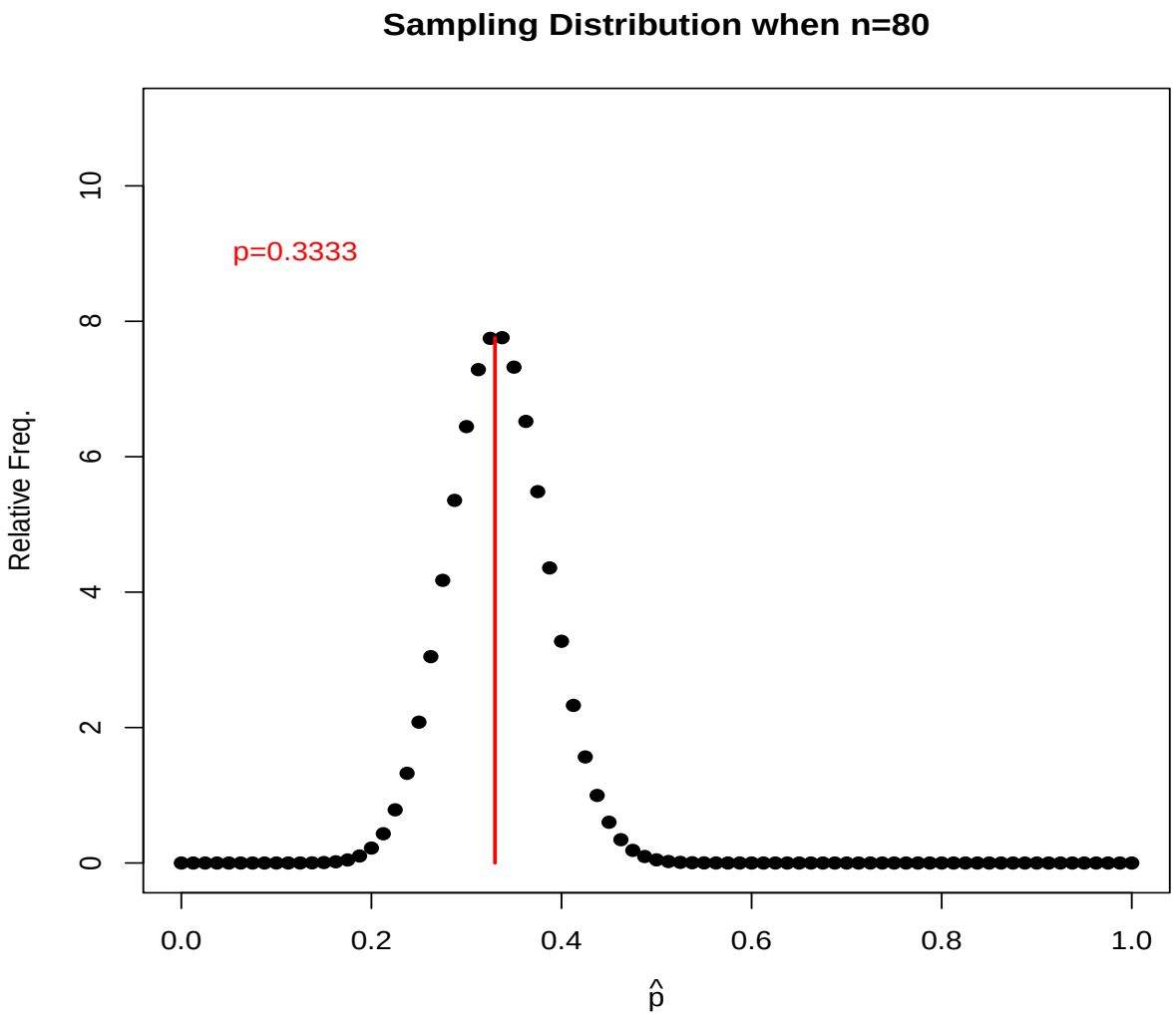
Sampling distribution of $\hat{p}$ when $n = 60$.



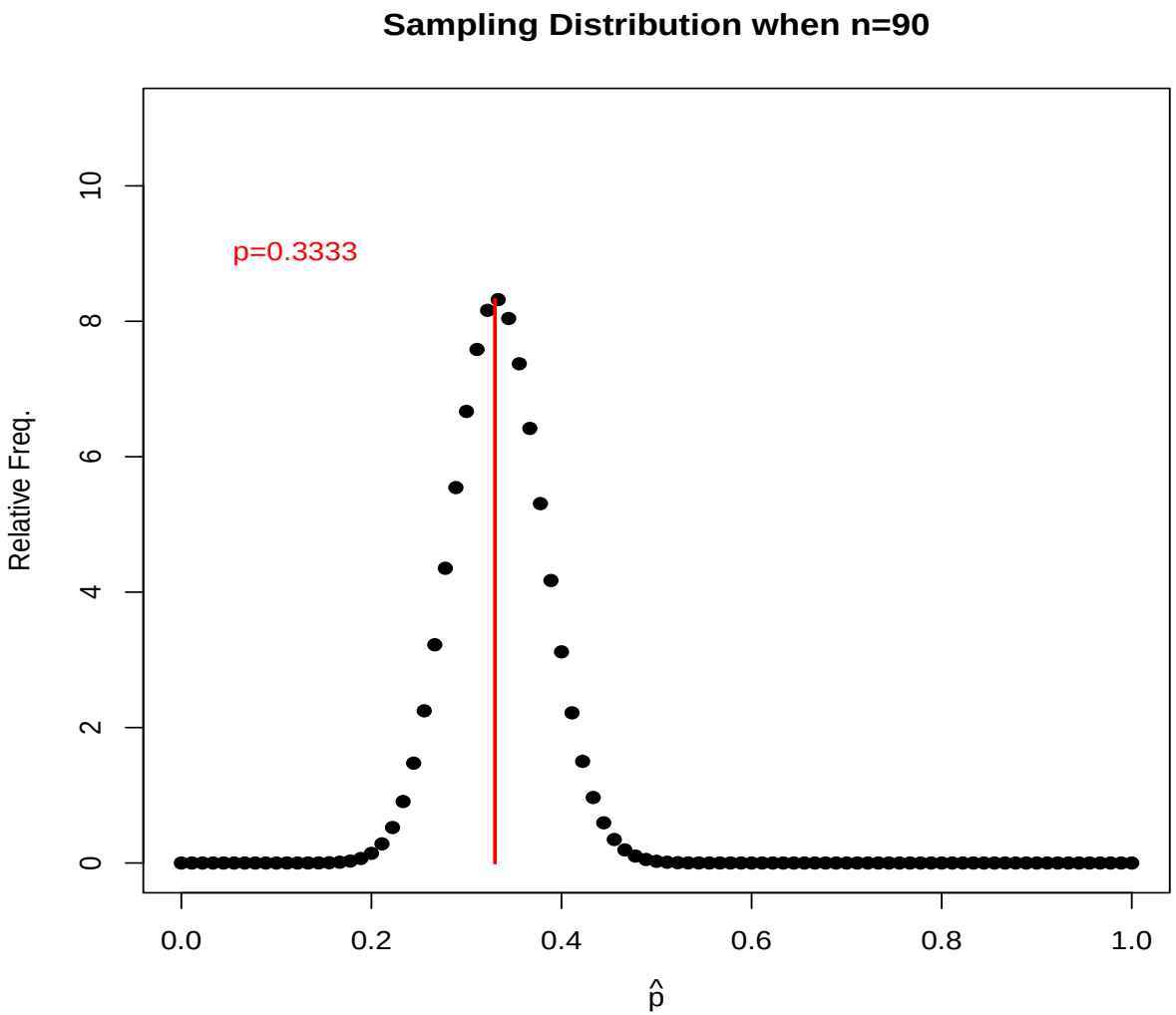
Sampling distribution of $\hat{p}$ when $n = 70$.



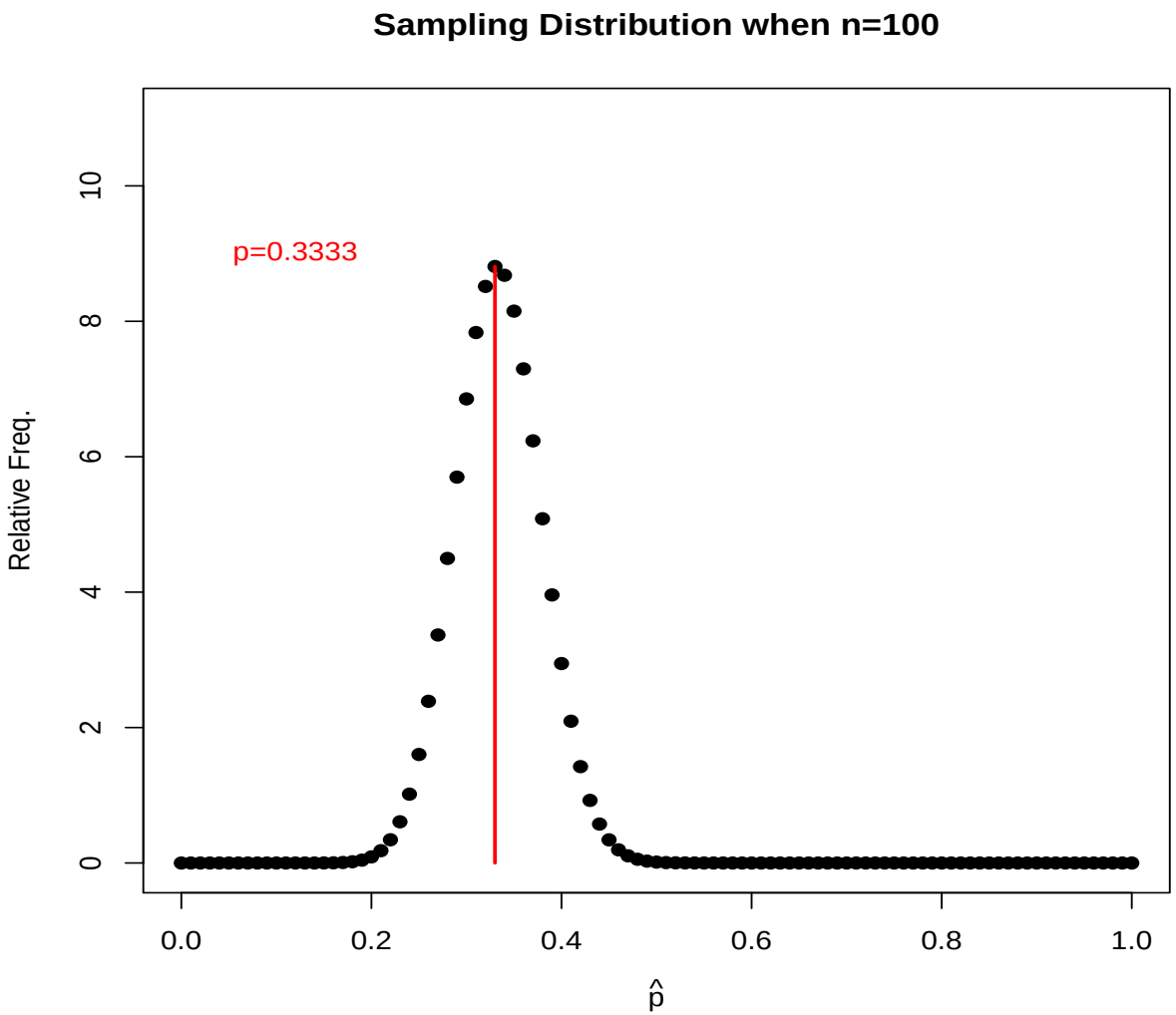
Sampling distribution of $\hat{p}$ when $n = 80$.



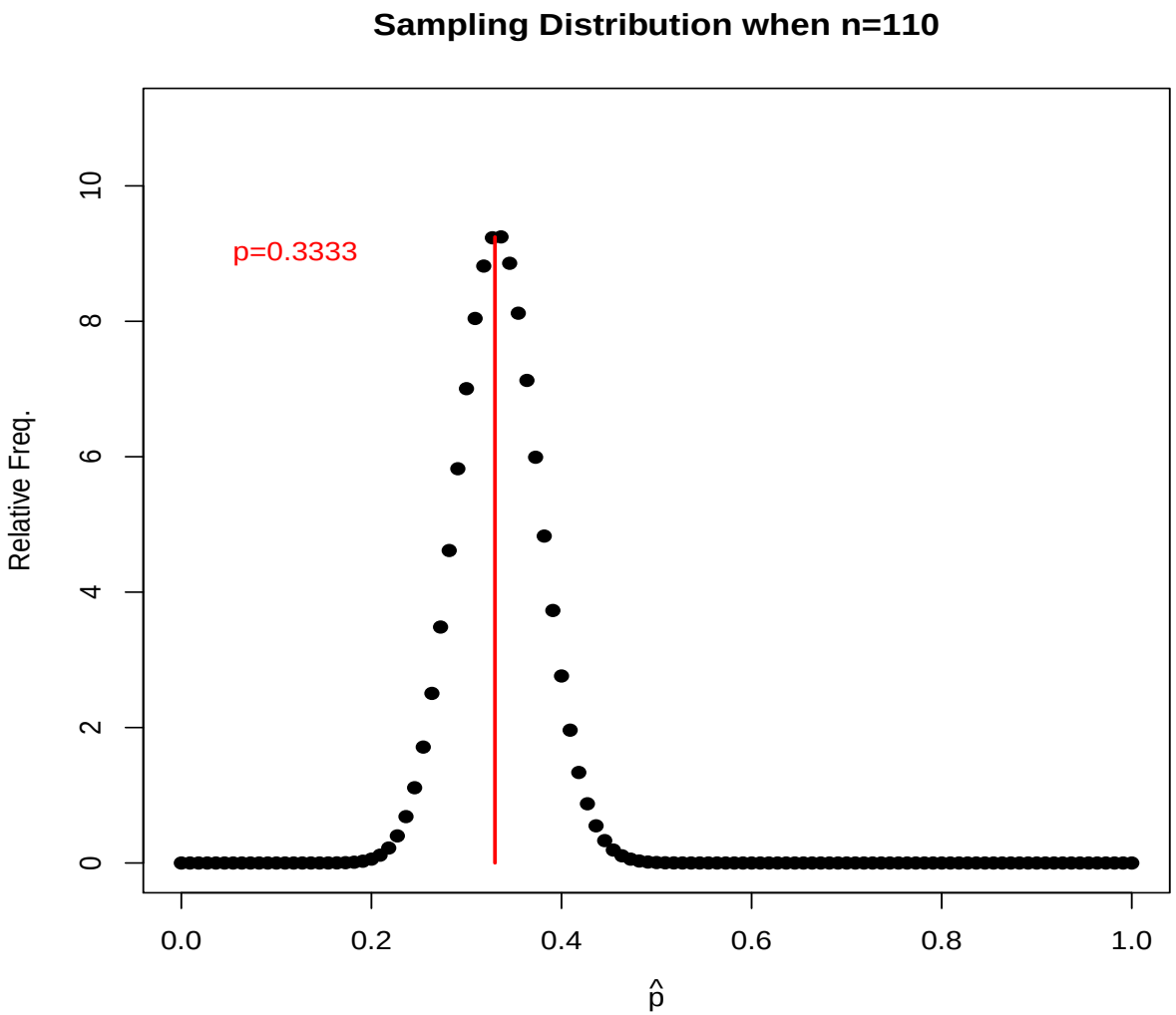
Sampling distribution of $\hat{p}$ when $n = 90$.



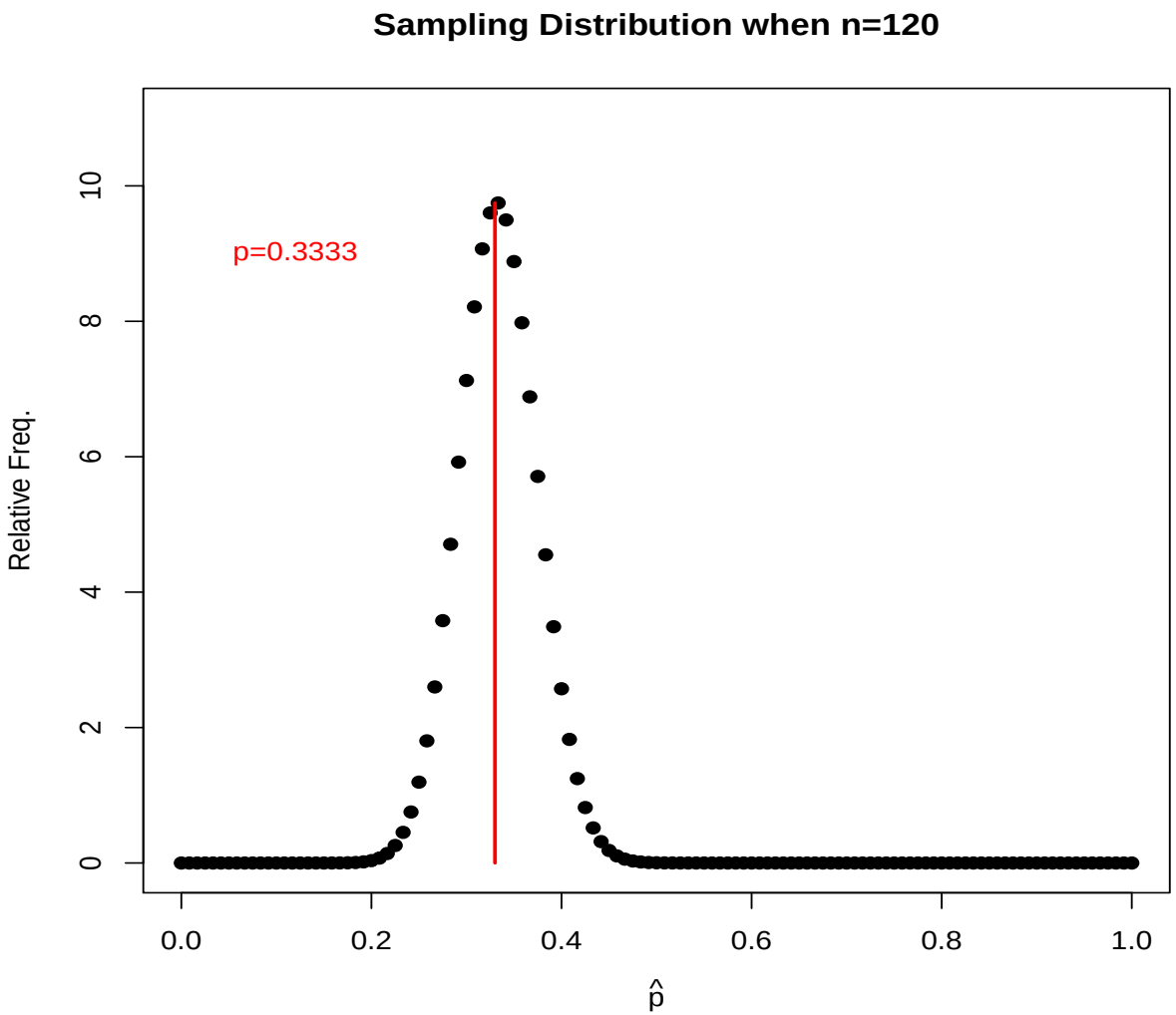
Sampling distribution of $\hat{p}$ when $n = 100$.



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

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

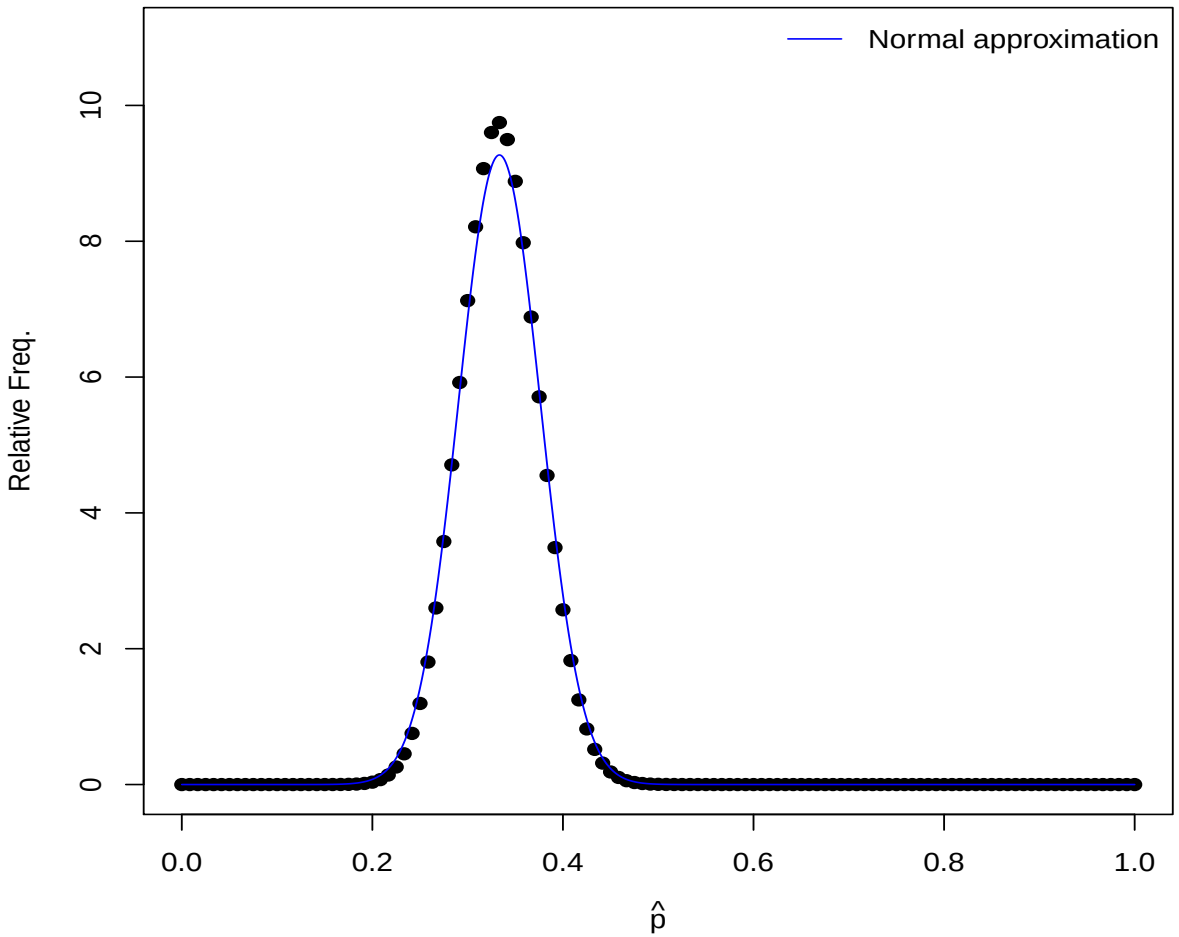
- As the sample size increases, the sampling distribution of $\hat{p}$ becomes **approximately Normal.** That is, for large n, $\hat{p}$ has approximately the $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ distribution.

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:

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

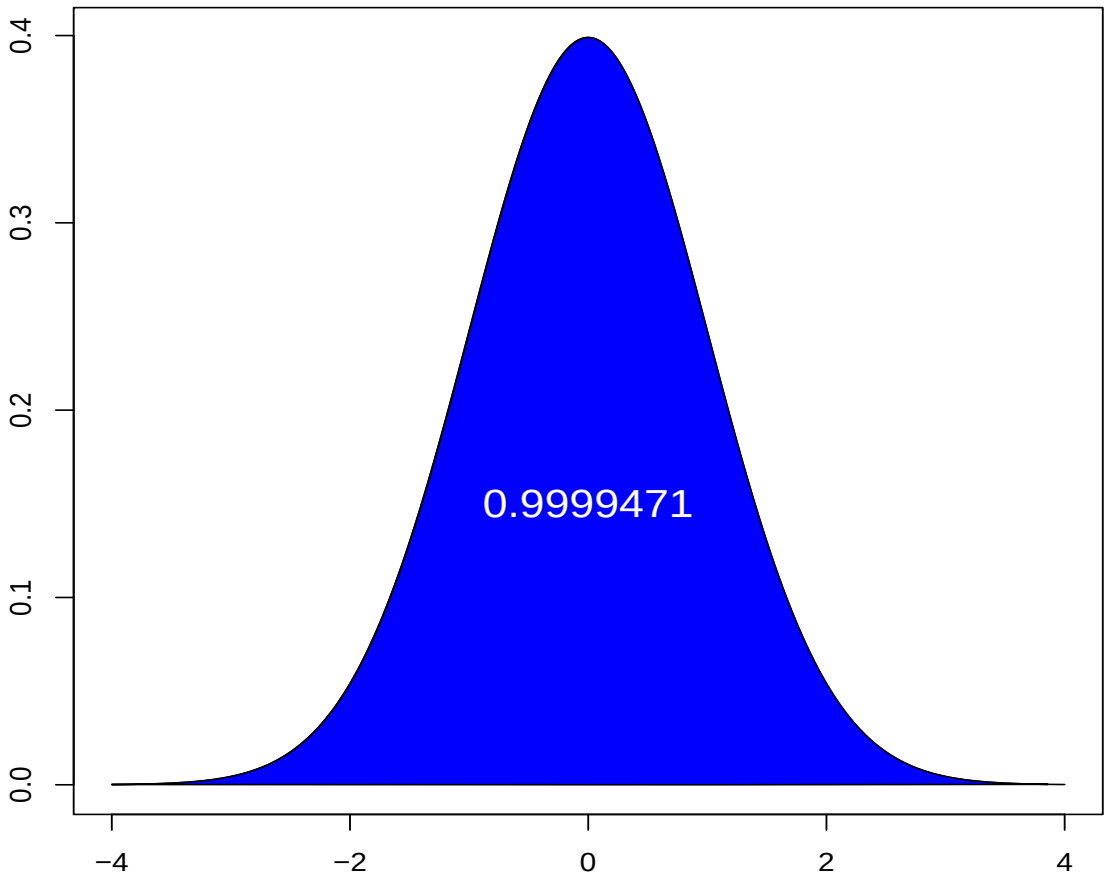
Sampling Distribution of $\hat{p}$ vs Normal Approximation



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 $p = 33.33\%$.

Binomial Distribution

A random variable Y is said to have a **binomial distribution** based on n trials with success probability p if and only if

$$p(y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \ y = 0, 1, 2, ..., n \text{ and } 0 \leq p \leq 1.$$

$$E(Y) = np \text{ and } V(Y) = np(1-p).$$

Bernoulli Distribution (Binomial with $n = 1$)

$$x_i = \begin{cases} 1 & \text{i-th trial is a "success"} \\ 0 & \text{otherwise} \end{cases}$$

$\mu = E(x_i) = p$
$\sigma^2 = V(x_i) = p(1-p)$
Let $\hat{p}$ be our estimate of p. Note that $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$. If n is "large", by the Central Limit Theorem, we know that:
$\bar{x}$ is roughly $N(\mu, \frac{\sigma}{\sqrt{n}})$, that is,
$\hat{p}$ is roughly $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Example

A machine is shut down for repairs if a random sample of 100 items selected from the daily output of the machine reveals at least 15% defectives. (Assume that the daily output is a large number of items). If on a given day the machine is producing only 10% defective items, what is the probability that it will be shut down?

Solution

Note that Y_i has a Bernoulli distribution with mean $1/10$ and variance $(1/10)(9/10) = 9/100$. By the Central Limit Theorem, $\bar{Y}$ has, roughly, a Normal distribution with mean $1/10$ and standard deviation $(3/10)/\sqrt{100} = \frac{3}{100}$.

We want to find $P(\sum_{i=1}^{100} Y_i \geq 15)$.

$P(\sum_{i=1}^{100} Y_i \geq 15) = P(\bar{Y} \geq 15/100) = P(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \geq \frac{15/100-10/100}{3/100})$

$\approx P(Z \geq 5/3) = P(Z \geq 1.66) = 0.0485$

Example

The manager of a supermarket wants to obtain information about the proportion of customers who dislike a new policy on cashing checks. How many customers should he sample if the wants the sample fraction to be within 0.15 of the true fraction, with probability 0.98?

Solution

$\hat{p}$ = sample proportion.

$P(|p - \hat{p}| \leq 0.15) = 0.98$

$P(-0.15 \leq p - \hat{p} \leq 0.15) = 0.98$

$$P\left(-\frac{-0.15}{\sqrt{p(1-p)/n}} \leq Z \leq \frac{-0.15}{\sqrt{p(1-p)/n}}\right) = 0.98$$

Let $z^* = \frac{-0.15}{\sqrt{p(1-p)/n}}$, then we want to find z^* such that

$P(-z^* \leq Z \leq z^*) = 0.98$

Using our table, $z^* = 2.33$.

Solution

Now, we have to solve the following equation for n

$$\frac{0.15}{\sqrt{p(1-p)/n}} = 2.33$$

$$n = \left(\frac{2.33}{0.15}\right)^2 p(1-p)$$

Note that the sample size, n, depends on p. We set $p = 0.5$, "to play it safe".

$$n = \left(\frac{2.33}{0.15}\right)^2 (0.5)^2 = 60.32$$

Therefore, 61 customers should be included in this sample.

Example

A lot acceptance sampling plan for large lots specifies that 50 items be randomly selected and that the lot be accepted if no more than 5 of the items selected do not conform to specifications. What is the approximate probability that a lot will be accepted if the true proportion of nonconforming items in the lot is 0.10?

Solution

Y_i has a Bernoulli distribution with mean $\mu = 1/10$ and standard deviation $\sigma = \sqrt{(1/10)(9/10)} = \sqrt{9/100} = 3/10$. By the Central Limit Theorem, we know that $\bar{Y}$ has, roughly, a Normal distribution with mean $1/10$ and standard deviation $\sigma/\sqrt{n} = 0.3/\sqrt{50} = 0.04242641$

$P(\sum_{i=1}^{50} Y_i \leq 5) = P(\sum_{i=1}^{50} Y_i \leq 5.5)$ (Using continuity correction, see page 382).

$= P(\sum_{i=1}^{50} Y_i \leq 5.5) = P(\bar{Y} \leq \frac{5.5}{50})$

$= P(\bar{Y} \leq 0.11) = P(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \leq \frac{0.11-0.10}{0.0424})$

$\approx P(Z < 0.24) = 1 - 0.4052 = 0.5948$ (Using our table).

Continuity Correction

Suppose that Y has a Binomial distribution with $n = 20$ and $p = 0.4$. We will find the exact probabilities that $Y \leq y$ and compare these to the corresponding values found by using two Normal approximations. One of them, when X is Normally distributed with $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$. The other one, W, a shifted version of X.

Continuity Correction (cont.)

For example,

$$P(Y \leq 8) = 0.5955987$$

As previously stated, we can think of Y as having approximately the same distribution as X.

$$P(Y \leq 8) \approx P(X \leq 8)$$

$$= P\left[\frac{X-np}{\sqrt{np(1-p)}} \leq \frac{8-8}{\sqrt{20(0.4)(0.6)}}\right]$$

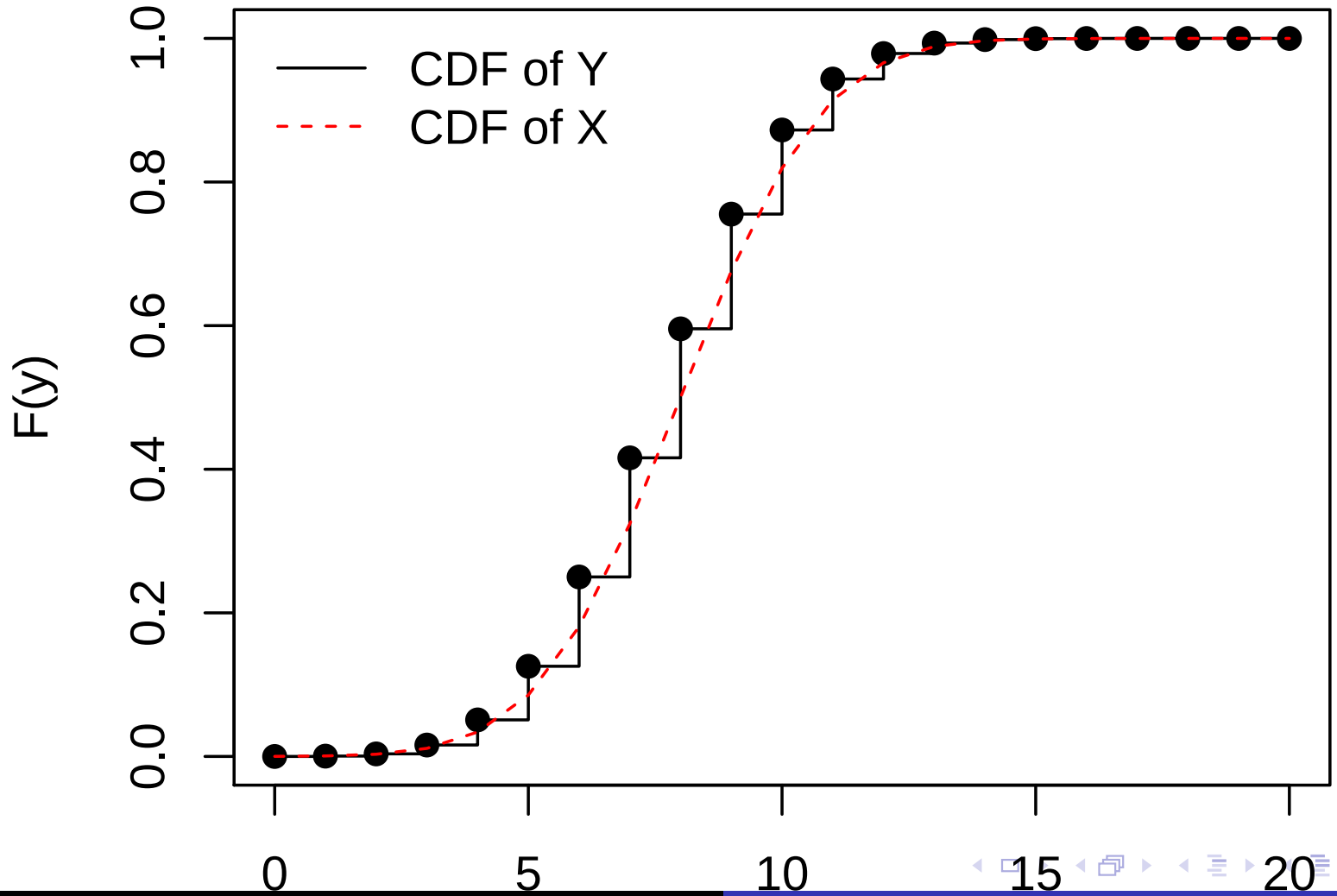
$$= P(Z \leq 0) = 0.5$$

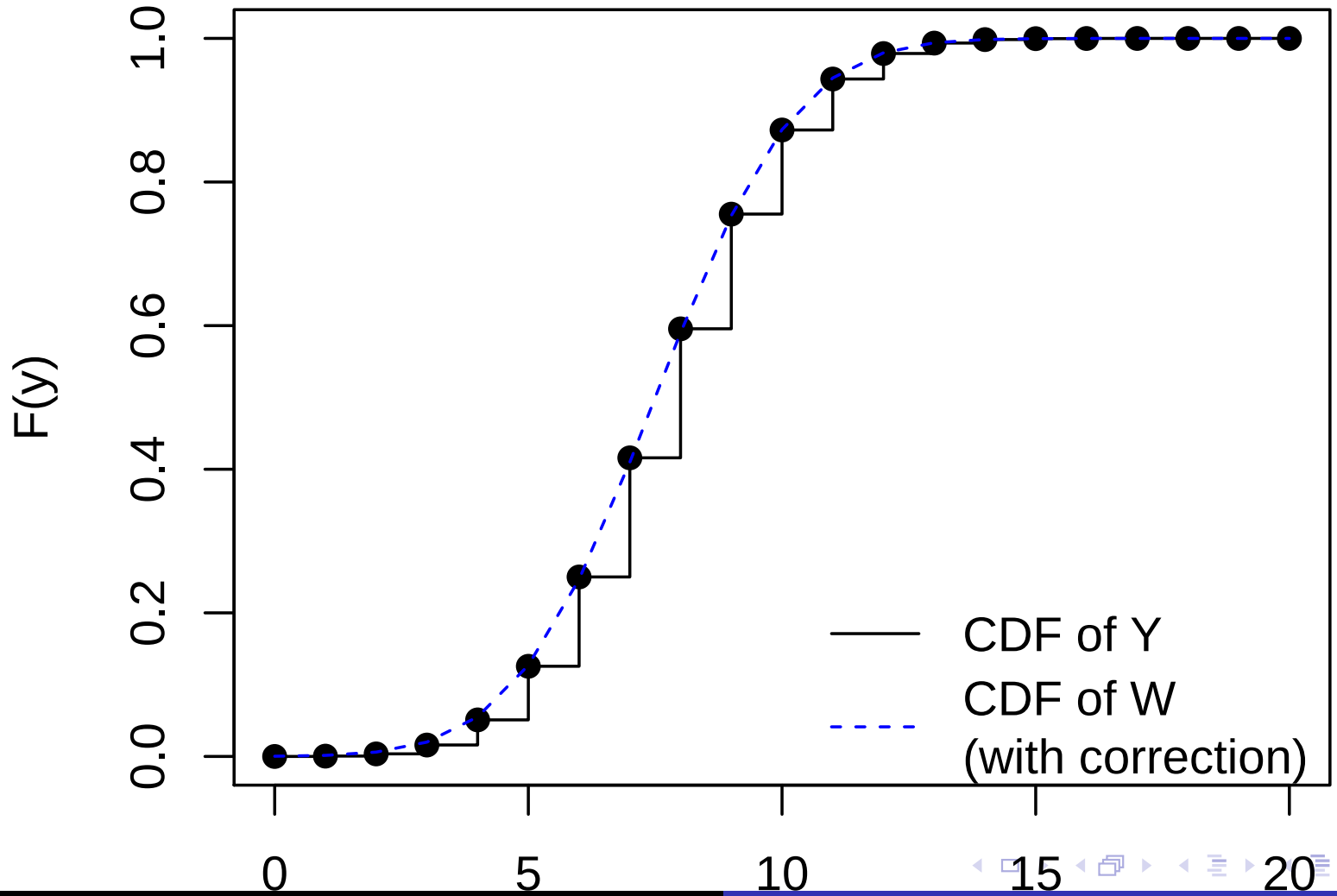
Continuity Correction (cont.)

$$P(Y \leq 8) \approx P(W \leq 8.5)$$

$$= P\left[\frac{W-np}{\sqrt{np(1-p)}} \leq \frac{8.5-8}{\sqrt{20(0.4)(0.6)}}\right]$$

$$= P(Z \leq 0.2282) = 0.5902615$$


CDF of Y
CDF of X
F(y)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


F(y)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
CDF of Y
CDF of W
(with correction)

Normal Approximation to Binomial

Let $X = \sum_{i=1}^{n} Y_i$ where $Y_1, Y_2, ..., Y_n$ are iid Bernoulli random variables. Note that $X = n\hat{p}$.

1. $n\hat{p}$ is approximately Normally distributed provided that np and $n(1-p)$ are greater than 5.
2. The expected value: $E(n\hat{p}) = np$.
3. The variance: $V(\hat{np}) = np(1-p) = npq$.

Theorem 7.1

Let $Y_1, Y_2, ..., Y_n$ be a random sample of size n from a Normal distribution with mean μ and variance σ^2. Then

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is Normally distributed with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$.

Theorem 7.2

Let $Y_1, Y_2, ..., Y_n$ be defined as in Theorem 7.1. Then $Z_i = \frac{Y_i - \mu}{\sigma}$ are independent, standard Normal random variables, $i = 1, 2, ..., n$, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

has a χ^2 distribution with n degrees of freedom (df).

Theorem 7.3

Let $Y_1, Y_2, ..., Y_n$ be a random sample from a Normal distribution with mean μ and variance σ^2. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

has a χ^2 distribution with $(n-1)$ df. Also, $\bar{Y}$ and S^2 are independent random variables.

Definition 7.2

Let Z be a standard Normal random variable and let W be a χ^2-distributed variable with ν df. Then, if Z and W are independent,

$$T = \frac{Z}{\sqrt{W/\nu}}$$

is said to have a t distribution with ν df.

Definition 7.3

Let W_1 and W_2 be independent χ^2-distributed random variables with ν_1 and ν_2 df, respectively. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

is said to have an F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom.

Example

Suppose that $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are independent random samples, with the variables X_i Normally distributed with mean μ_1 and variance σ_1^2 and variables Y_i Normally distributed with mean μ_2 and variance σ_2^2. The difference between the sample means, $\bar{X} - \bar{Y}$, is then a linear combination of $m + n$ Normally distributed random variables, and, by Theorem 6.3, is itself Normally distributed.

a. Find $E(\bar{X} - \bar{Y})$.

b. Find $V(\bar{X} - \bar{Y})$.

c. Suppose that $\sigma_1^2 = 2$, $\sigma_2^2 = 2.5$, and $m = n$. Find the sample sizes so that $(\bar{X} - \bar{Y})$ will be within 1 unit of $(\mu_1 - \mu_2)$ with probability 0.95.

Solution

a. First, recall that
$\bar{X} = \frac{X_1+X_2+...+X_m}{m}$ and $\bar{Y} = \frac{Y_1+Y_2+...+Y_n}{n}$.
$E(\bar{X}) = E\left(\frac{X_1+X_2+...+X_m}{m}\right) = \frac{E(X_1)+E(X_2)+...+E(X_m)}{m}$
$= \frac{\mu_1+\mu_1+...+\mu_1}{m} = \frac{m\mu_1}{m} = \mu_1$.
Similarly,
$E(\bar{Y}) = \frac{\mu_2+\mu_2+...+\mu_2}{m} = \frac{m\mu_2}{m} = \mu_2$.
Hence,
$E(\bar{X}-\bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$.

Solution

b. $V(\bar{X}) = V\left(\frac{X_1+X_2+...+X_m}{m}\right) = \frac{V(X_1)+V(X_2)+...+V(X_m)}{m^2}$

$V(\bar{X}) = \frac{\sigma_1^2+\sigma_1^2+...+\sigma_1^2}{m^2} = \frac{\sigma_1^2}{m}.$

$V(\bar{Y}) = V\left(\frac{Y_1+Y_2+...+Y_n}{n}\right) = \frac{V(Y_1)+V(Y_2)+...+V(Y_n)}{n^2}$

$V(\bar{Y}) = \frac{\sigma_2^2+\sigma_2^2+...+\sigma_2^2}{n^2} = \frac{\sigma_2^2}{n}.$

Solution

c. Let $U = \bar{X} - \bar{Y}$. By theorem 6.3, we know that U has a Normal distribution with mean $\mu_U = \mu_1 - \mu_2$ and variance $\sigma_U^2 = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$.
$P(|U - \mu_U| \leq 1) = 0.95$
$P(-1 \leq U - \mu_U \leq 1) = 0.95$
Now, we just have to divide by the standard deviation of U.
$P(-\frac{1}{\sqrt{4.5/n}} \leq \frac{U-\mu_U}{\sigma_U} \leq \frac{1}{\sqrt{4.5/n}}) = 0.95$
$P(-\frac{\sqrt{n}}{\sqrt{4.5}} \leq \frac{U-\mu_U}{\sigma_U} \leq \frac{\sqrt{n}}{\sqrt{4.5}}) = 0.95$
We need n to satisfy $\sqrt{\frac{n}{4.5}} = 1.96$, then $n = (1.96)^2(4.5) = 17.28$
Our final answer is $n = 18$ (always round up when finding a sample size).

Example

The Environmental Protection Agency is concerned with the problem of setting criteria for the amounts of certain toxic chemicals to be allowed in freshwater lakes and rivers. A common measure of toxicity for any pollutant is the concentration of the pollutant that will kill half of the test species in a given amount of time (usually 96 hours for fish species). This measure is called LC50 (lethal concentration killing 50% of test species). In many studies, the values contained in the natural logarithm of LC50 measurements are Normally distributed, and, hence, the analysis is based on ln(LC50) data.

Example

Suppose that $n = 20$ observations are to be taken on ln(LC50) measurements and that $\sigma^2 = 1.4$. Let S^2 denote the sample variance of the 20 measurements.

a. Find a number b such that $P(S^2 \leq b) = 0.975$.

b. Find a number a such that $P(a \leq S^2) = 0.975$.

c. If a and b are as in parts a) and b), what is $P(a \leq S^2 \leq b)$?

Solution

These values can be found by using percentiles from the chi-square distribution.

With $\sigma^2 = 1.4$ and $n = 20$,

$\frac{n-1}{\sigma^2}S^2 = \frac{19}{1.4}S^2$ has a chi-square distribution with 19 degrees of freedom.

a. $P(S^2 \leq b) = P\left(\frac{n-1}{\sigma^2}S^2 \leq \frac{(n-1)b}{\sigma^2}\right) = P\left(\frac{19}{1.4}S^2 \leq \frac{19b}{1.4}\right) = 0.975$

$\frac{19b}{1.4}$ must be equal to the 97.5%-tile of a chi-square with 19 df, thus $\frac{19b}{1.4} = 32.8523$ (using Table 6). An so, $b = 2.42$

Solution

b. Similarly, $P(S^2 \geq a) = P\left(\frac{n-1}{\sigma^2}S^2 \geq \frac{(n-1)a}{\sigma^2}\right) = 0.975$. Thus, $\frac{19a}{1.4} = 8.90655$, the 2.5%-tile of this chi-square distribution, and so $a = 0.656$.

c. $P(a \leq S^2 \leq b) = P(0.656 \leq S^2 \leq 2.42) = 0.95$.

Example

Use the structures of T and F given in Definitions 7.2 and 7.3, respectively, to argue that if T has a t distribution with ν df, then $U = T^2$ has an F distribution with 1 numerator degree of freedom and ν denominator degrees of freedom.

Solution

Define $T = \frac{Z}{\sqrt{W/\nu}}$ as in Definition 7.2. Then, $T^2 = \frac{Z^2}{W/\nu}$. Since Z^2 has a chi-square distribution with 1 degree of freedom, and Z and W are independent, T^2 has an F distribution with 1 numerator and ν denominator degrees of freedom.