STA258H5

Al Nosedal
and Alison Weir

Winter 2017

SAMPLING DISTRIBUTIONS

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

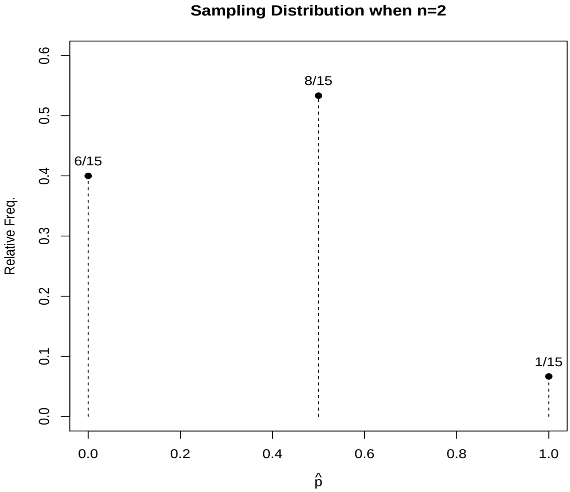
$\hat{p}_1$={A,B} = {1,1}=100%	$\hat{p}_9$ ={B,F} = {1,0}=50%
$\hat{p}_2$={A,C} = {1,0}=50%	$\hat{p}_{10}$ ={C,D} = {0,0}=0%
$\hat{p}_3$={A,D} = {1,0}=50%	$\hat{p}_{11}$ ={C,E}= {0,0}=0%
$\hat{p}_4$ ={A,E}= {1,0}=50%	$\hat{p}_{12}$={C,F} {0,0}=0%
$\hat{p}_5$={A,F} = {1,0}=50%	$\hat{p}_{13}$ ={D,E}{0,0}=0%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

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 $n = 2$.



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.

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

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

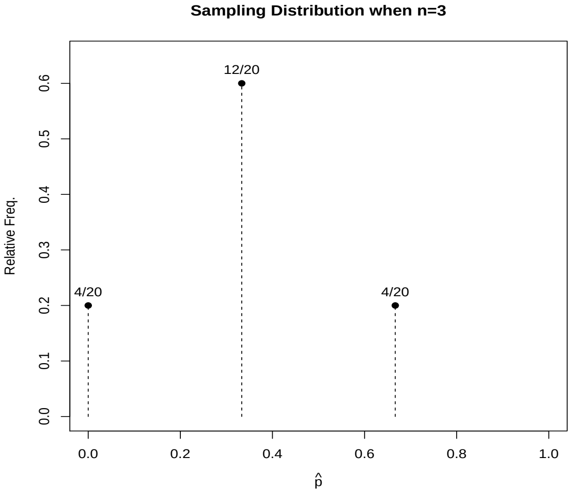
$$
\begin{array}{llll}
\hat{p}_1 = 2/3 & \hat{p}_6 = 1/3 & \hat{p}_{11} = 1/3 & \hat{p}_{16} = 1/3 \\
\hat{p}_2 = 2/3 & \hat{p}_7 = 1/3 & \hat{p}_{12} = 1/3 & \hat{p}_{17} = 0 \\
\hat{p}_3 = 2/3 & \hat{p}_8 = 1/3 & \hat{p}_{13} = 1/3 & \hat{p}_{18} = 0 \\
\hat{p}_4 = 2/3 & \hat{p}_9 = 1/3 & \hat{p}_{14}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

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 $n = 3$.



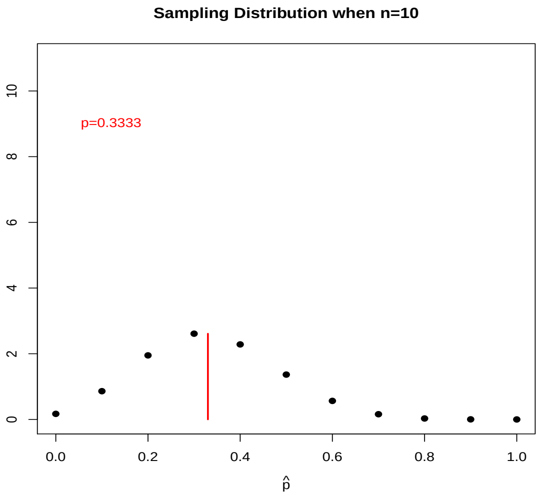
Prediction outcome of the election

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

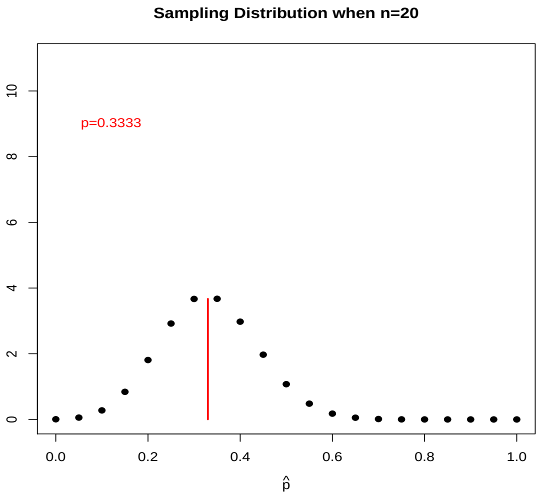
More realistic example

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.

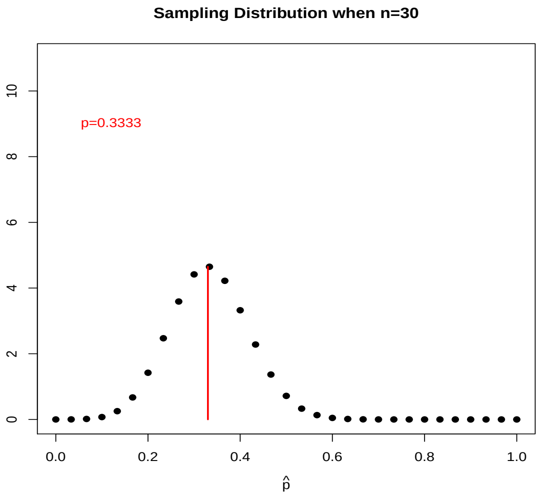
Sampling distribution of $\hat{p}$ when $n = 10$.



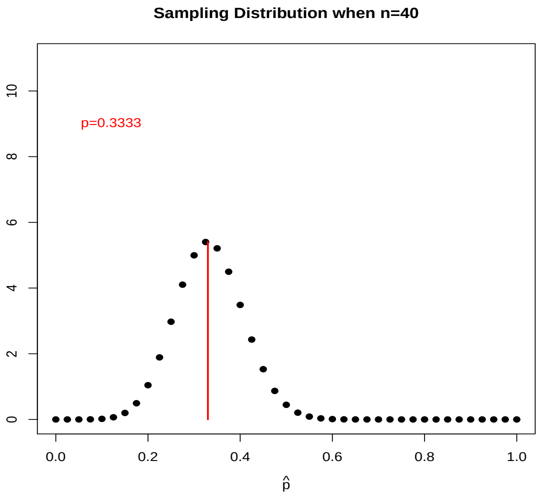
Sampling distribution of $\hat{p}$ when $n = 20$.



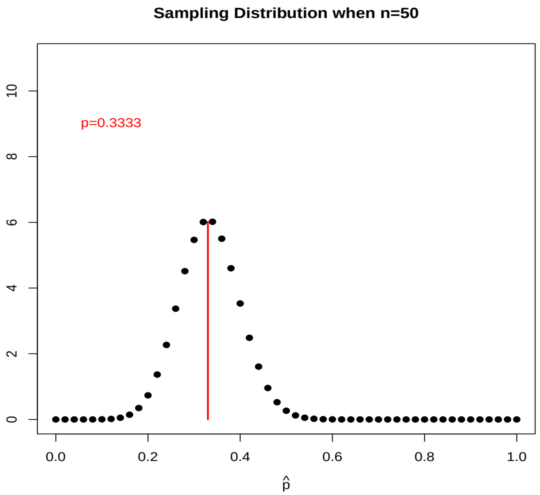
Sampling distribution of $\hat{p}$ when $n = 30$.



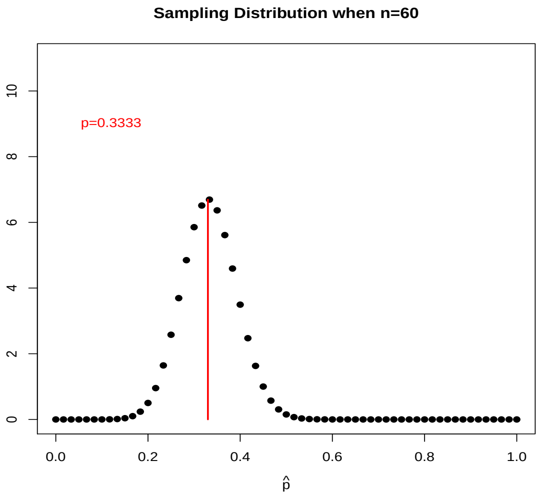
Sampling distribution of $\hat{p}$ when $n = 40$.



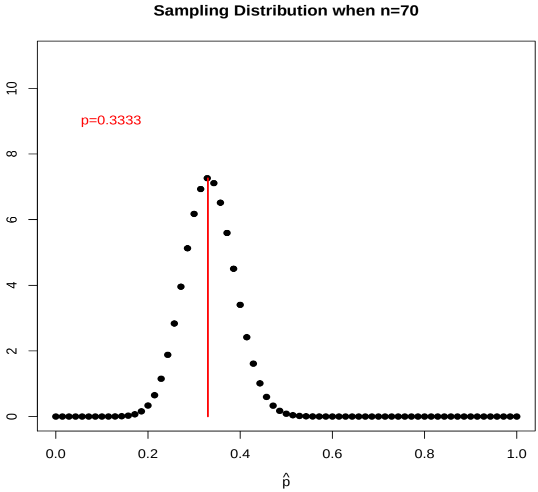
Sampling distribution of $\hat{p}$ when $n = 50$.



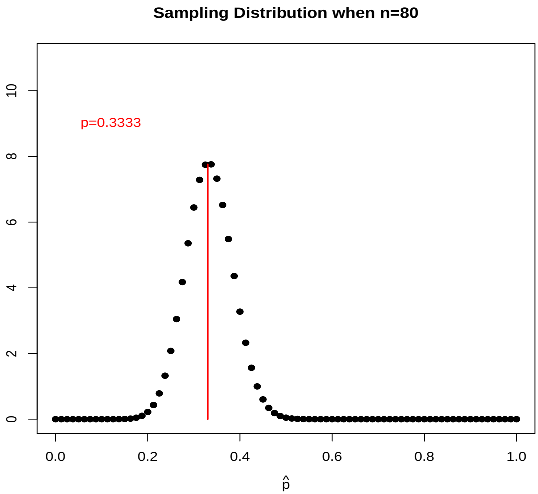
Sampling distribution of $\hat{p}$ when $n = 60$.



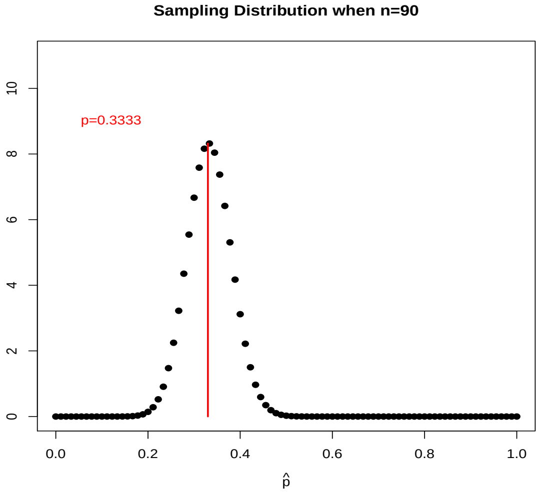
Sampling distribution of $\hat{p}$ when $n = 70$.



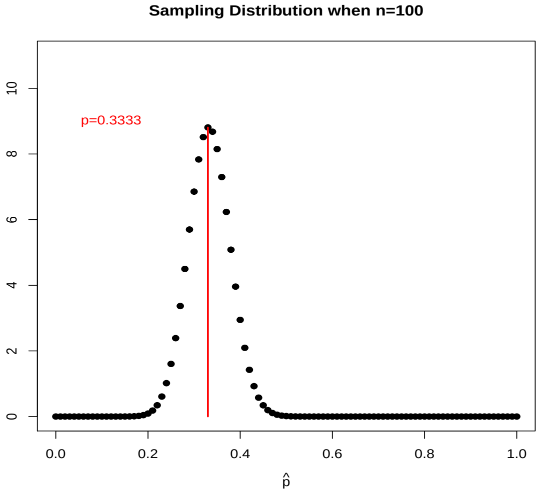
Sampling distribution of $\hat{p}$ when $n = 80$.



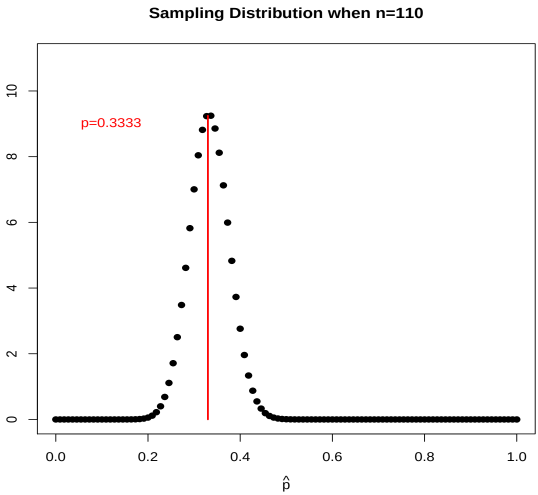
Sampling distribution of $\hat{p}$ when $n = 90$.



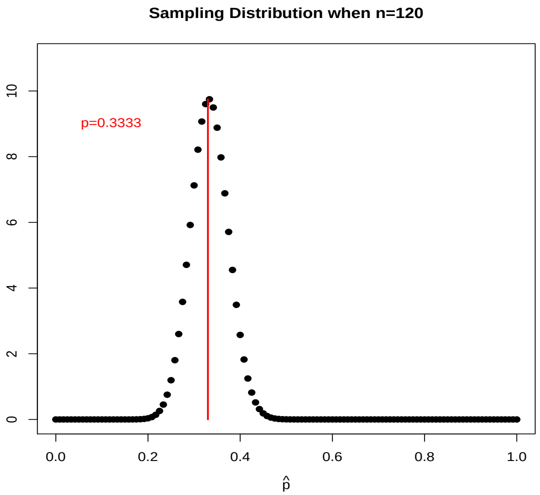
Sampling distribution of $\hat{p}$ when $n = 100$.



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

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

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:

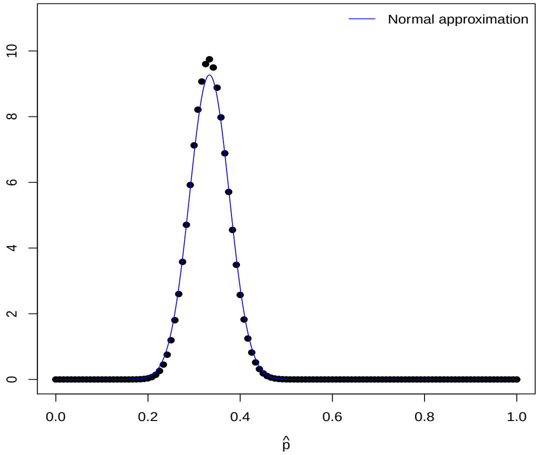
- As the sample size increases, the sampling distribution of $\hat{p}$ becomes **approximately Normal.** That is, for large n, $\hat{p}$ has approximately the $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ distribution.

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:

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

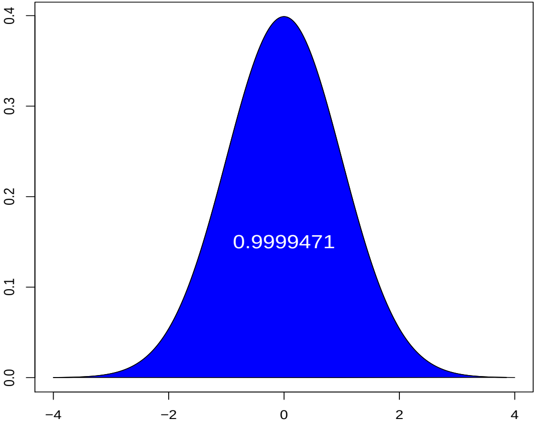
Sampling Distribution of $\hat{p}$ vs Normal Approximation



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 $\hat{p} < 0.50$=
$P(Y < 0.50) = P\left(\frac{Y-\mu}{\sigma} < \frac{0.5-0.3333}{0.0430}\right) = P(Z < 3.8767)$.

$P(Z < 3.8767)$



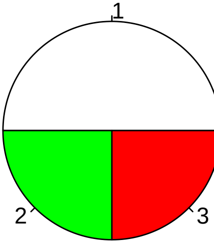
Predicting outcome of the election with our approximation

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 $p = 33.33\%$.

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.

Another example



Parent population: Children's game spinner.

X = number of moves.

x	1	2	3
$p(x)$	$1/2$	$1/4$	$1/4$

Another example (cont.)

x	1	2	3	Sum
$p(x)$	$1/2$	$1/4$	$1/4$	1
$xp(x)$	$2/4$	$2/4$	$3/4$	$7/4 = 1.75$
$x^2p(x)$	$2/4$	$4/4$	$9/4$	$15/4 = 3.75$

Mean $= \mu = 1.75$
Variance $= \sigma^2 = 3.75 - (1.75)^2 = 0.6875$

Another example (cont.)

Spin twice and get distribution of sample mean ($n = 2$).

sample	probability	sample mean
(1,1)	$(1/2)(1/2) = 1/4$	1
(1,2)	$(1/2)(1/4) = 1/8$	1.5
(1,3)	$(1/2)(1/4) = 1/8$	2
(2,1)	$1/8$	1.5
(2,2)	$1/16$	2
(2,3)	$1/16$	2.5
(3,1)	$1/8$	2
(3,2)	$1/16$	2.5
(3,3)	$1/16$	3

$\bar{x}$	1	1.5	2	2.5	3	Sum
$p(\bar{x})$	4/16	4/16	5/16	2/16	1/16	1
$\bar{x}p(\bar{x})$	4/16	6/16	10/16	5/16	3/16	28/16
$\bar{x}^2 p(\bar{x})$	8/32	18/32	40/32	25/32	18/32	109/32

Mean $= \mu = 1.75$
Variance $= Var(\bar{x}) = \sigma^2_{\bar{x}} = 109/32 - (28/16)^2 = 0.34375$

Statistics vs Parameters

- **Statistic** is a numerical value computed from a sample. The sample mean is a statistic.
- **Parameter** is a numerical value associated with a population. The population mean is a parameter.
- We often want to know about a parameter. But we can?t since the population is too large. But we can estimate the parameter using statistics computed from the sample.

Sample Mean from Normal Population

Let $X_1, X_2, ..., X_n$ be iid $N(\mu, \sigma)$.
Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
$\bar{X}$ is a random variable with the $N(\mu, \sigma/n)$ distribution.

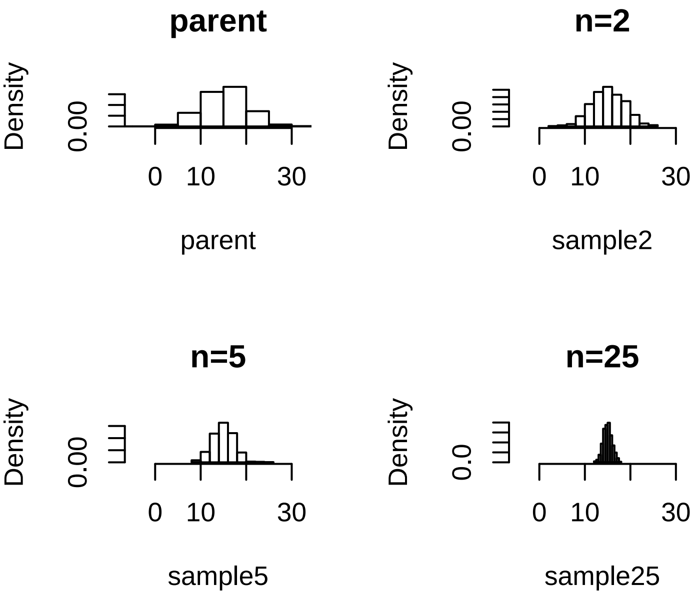
Proof (using MGFs)

$M_{\bar{X}}(t) = E(e^{t\bar{X}}) = E\left(exp\left(\frac{t}{n}X_1 + \frac{t}{n}X_2 + ... + \frac{t}{n}X_n\right)\right)$
$= E\left(exp\left(\frac{t}{n}X_1\right)\right) E\left(exp\left(\frac{t}{n}X_2\right)\right) ... E\left(exp\left(\frac{t}{n}X_n\right)\right)$ (by independence)
$= \left[E\left(exp\left(\frac{t}{n}X\right)\right)\right]^n$ (identical distributions)
$= \left[M_X\left(\frac{t}{n}\right)\right]^n$
$= \left[exp\left(\mu\frac{t}{n} + \sigma^2\frac{t^2}{2n^2}\right)\right]^n$
$= exp\left(\mu t + \frac{\sigma^2}{n}\frac{t^2}{2}\right)$
which is the MGF for the $N(\mu, \sigma^2/n)$ distribution.

Effect of Sample Size

```
parent<- rnorm(1000,mean=15, sd=5);

sample2<-rnorm(1000,mean=15,sd=5/sqrt(2));

sample5<-rnorm(1000,mean=15,sd=5/sqrt(5));

sample25<-rnorm(1000,mean=15,sd=5/sqrt(25));
```

```r
MIN<-min(parent);

MAX<-max(parent);

par(mfrow=c(2,2))

hist(parent,freq=FALSE,main="parent", xlim=c(MIN,MAX));

hist(sample2,freq=FALSE,main="n=2", xlim=c(MIN,MAX));

hist(sample5,freq=FALSE,main="n=5", xlim=c(MIN,MAX));

hist(sample25,freq=FALSE,main="n=25", xlim=c(MIN,MAX));
```



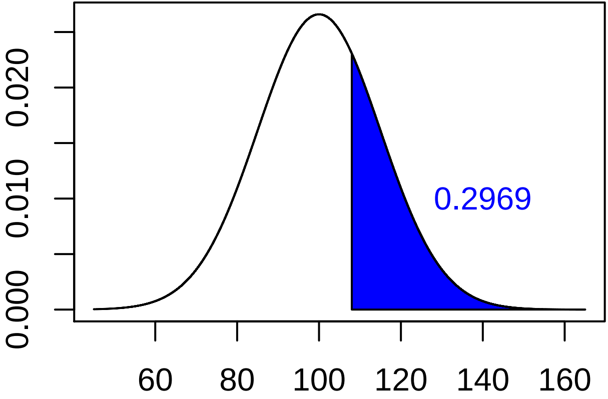
parent
Density
0.00
0
10
30
parent
n=2
Density
0.00
0
10
30
sample2
n=5
Density
0.00
0
10
30
sample5
n=25
Density
0.0
0
10
30
sample25

Example

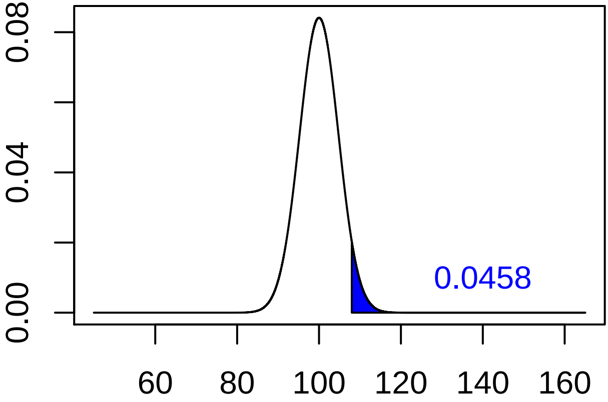
Adult IQ scores are Normally distributed with a mean of 100 and standard deviation 15.
a. What is the probability a randomly selected adult has an IQ that is at least 108?
b. If 10 (independent) adults are randomly selected, what is the probability their average IQ is at least 108?
c. If 100 (independent) adults are randomly selected, what is the probability their average IQ is at least 108?

```r
## Solution a;

1-pnorm(108,mean=100,sd=15);

## [1] 0.2969014

## Solution b;

1-pnorm(108,mean=100,sd=15/sqrt(10));

## [1] 0.04584514

## Solution c;

1-pnorm(108,mean=100,sd=15/sqrt(100));

## [1] 4.821303e-08
```

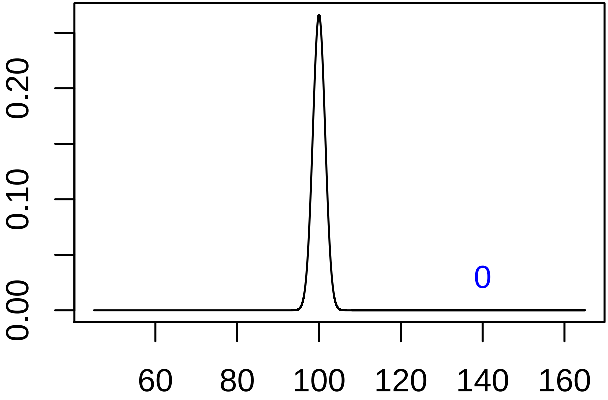
Solution a



Solution b



Solution c



Solution b (using table)

$\bar{X}$ = average IQ of ten randomly selected adults, we want to find $P(\bar{X} > 108)$, where $\bar{X}$ is Normally distributed, $\mu^* = 100$ and $\sigma^* = \frac{\sigma}{\sqrt{10}} = \frac{15}{\sqrt{10}} = 4.7434$. Hence,

$$P(\bar{X} > 108) = P\left(\frac{\bar{X}-\mu^*}{\sigma^*} > \frac{108-100}{4.7434}\right)$$
$$= P\left(Z > 1.6865\right)$$
$$\approx P\left(Z > 1.69\right)$$
$$= 0.0455$$