

# 7 Sampling and Sampling Distributions

## 7.1 Types of Sampling Methods

### A Survey and a Census

**Example:** Suppose we are interested in the population proportion of American adults who support the President's foreign policy in Afghanistan?

Ideally, take a \_\_\_\_\_.

What would be the *disadvantages*?

What would be the *advantages*?

Instead, take a \_\_\_\_\_.

**Definition:** A **sample survey** selects a sample of people from a population and interviews them to collect data.

What would be the *advantages*?

What would be the *disadvantages*?

### Probability vs. Nonprobability samples

The **frame** is the list of subjects in the population from which the sample is taken.

Types of **probability sampling** include simple random sampling, systematic sampling, stratified sampling and cluster sampling.

An example of **nonprobabilistic sampling** occurred during the Presidential Election of 1948, when the major polls were predicting that Thomas Dewey would defeat Harry Truman.

## Simple Random Samples

A **simple random sample** of  $n$  subjects from a population is one in which each possible sample of that size has the same chance of being selected.

Sampling **WITH** replacement vs. sampling **withOUT** replacement.

*If  $n$  is a small percentage of the population size, then sampling **without** replacement is similar to sampling **with** replacement, since sampling the same person more than once would be quite unlikely.*

**Systematic Samples** - Sampling is done on every  $k$ th item in a list.

Most of the time, a *systematic sample* gives results comparable to those of a *simple random sample*.

When might a *systematic sample* not be as satisfactory as a *simple random sample*?

When might a *systematic sample* improve over a *simple random sample*?

*Caution:* In systematic sampling, we still must have a sampling frame and be careful when defining the target population.

Is sampling every 20th student who enters the library a representative sample of the entire student body?

Is sampling every 10th passenger who exits an airplane a representative sample of all the persons on that flight?

**Stratified random sampling** occurs when the population is divided into groups called *strata*, such that a simple random sample is taken within each *stratum*.

For example: Poll elementary school children in Harrisonburg to inquire about whether or not they like school, or get enough food on a daily basis.

How should stratification be done?

**Cluster sampling** occurs when the elements (or observation units) are aggregated into large sampling units, called *clusters*, such that some clusters are sampled, and then additional sampling occurs within the clusters themselves.

For example, suppose you want to survey Lutheran church members in Minneapolis, but you do not have a list of all church members in the city, so you cannot take a simple random sample of church members.

Does a *cluster sample* of 500 Lutherans provide as much information as a *simple random sample* of 500 Lutherans?

Why should *cluster sampling* be performed?

**Example:** Suppose one wants to estimate the average amount of time that professors at JMU spend grading homework in a specific week.

How can one perform a *simple random sample* of size  $n$ ?

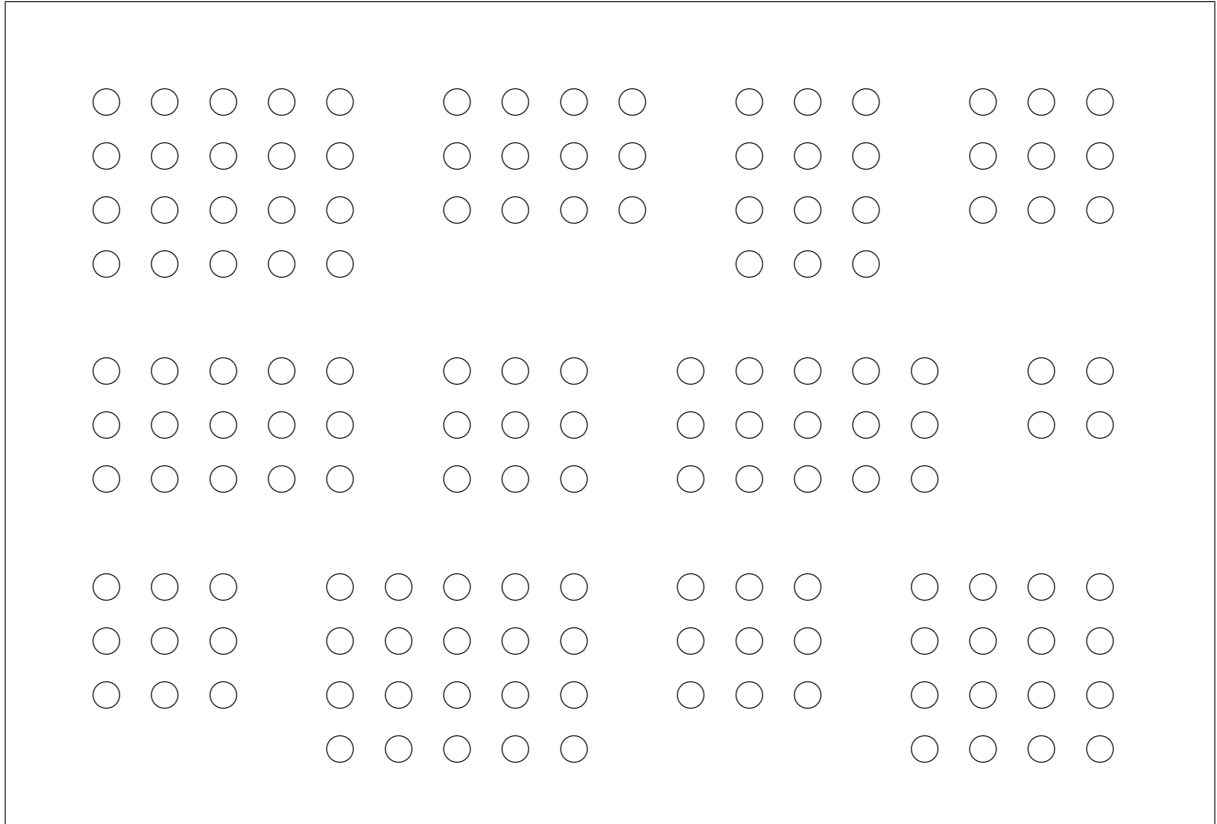
How can one perform a *stratified sample*?

What advantage does *stratified sampling* have over *simple random sampling*?

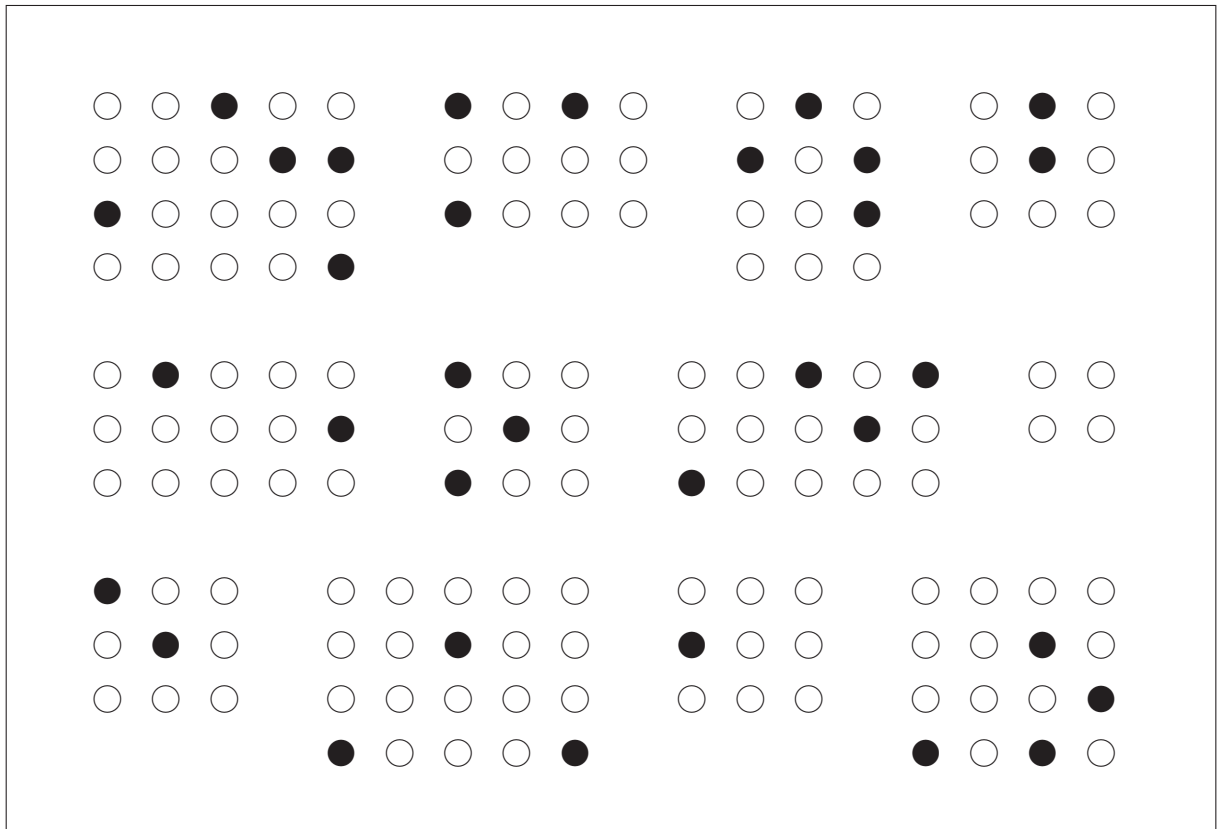
How can one perform a *cluster sample*?

For this example, is *cluster sampling* better than *stratified sampling*?

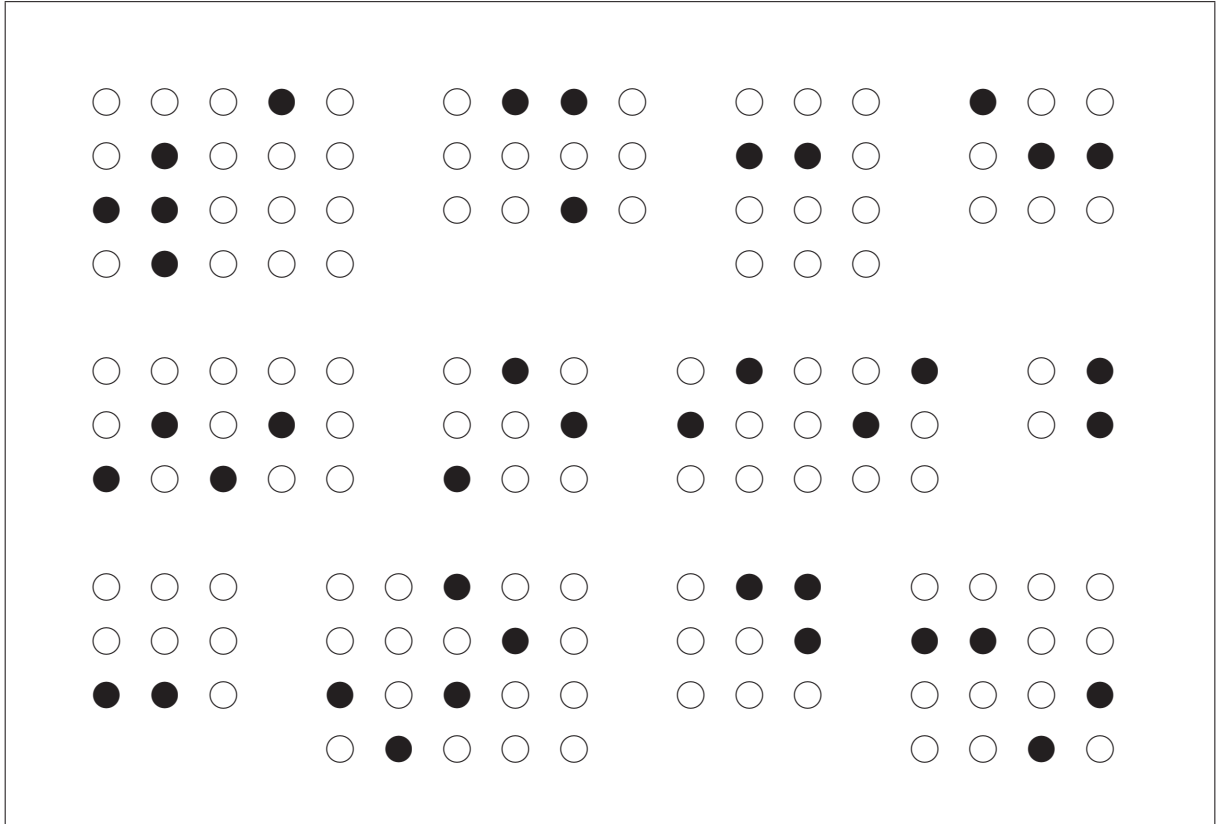
Consider the following target population.



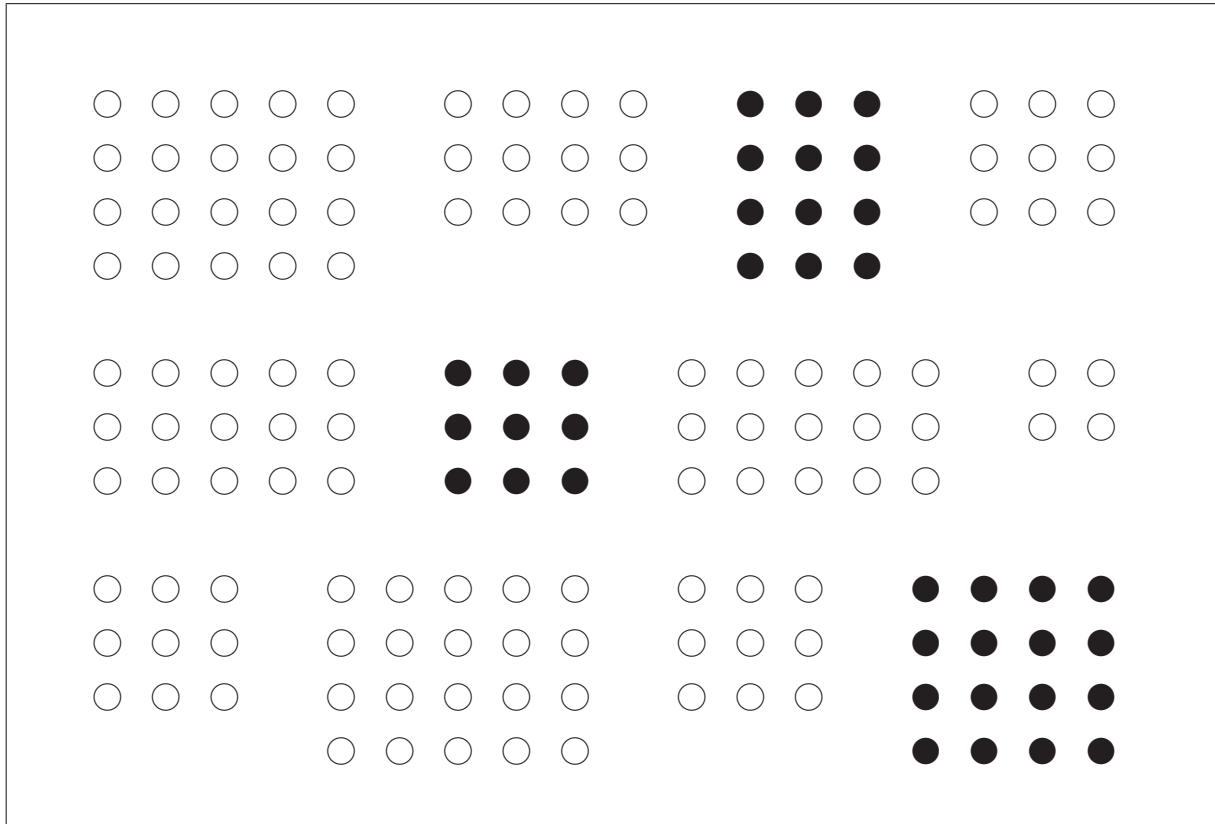
What type of sampling is represented here?



What type of sampling is represented here?



What type of sampling is represented here?



## 7.2 Evaluating Survey Worthiness

Survey error is caused by (1) coverage, (2) nonresponse, (3) sampling, and (4) measurement errors.

Recall: The **frame** is the list of subjects in the population from which the sample is taken.

### (1) Coverage error

**Example:** Sample only American senior citizens when estimating how American adults feel on Social Security issues.

What are the *population* and *sampling frame*?



Would the results of this poll be valid?

□

**(2) Nonresponse error** occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.

**(3) Sampling error** occurs from using a sample rather than a census, producing a **margin of error**.

**(4) Measurement error** occurs when the subject gives an incorrect response (perhaps lying), or the question wording or the way the interviewer asks the questions is confusing or misleading.

**Example:** In 2000 for a JMU student research project, the researchers asked questions similar to the following.

“Do you smoke marijuana?”

“Do you think JMU students smoke marijuana?”

□

**Example:** The *Literary Digest* Poll (*Know this example in detail, although you need not memorize the numbers.*)

Franklin Roosevelt vs. Alfred Landon, Election of 1936.

Since 1916, the *Literary Digest* correctly picked the Presidents.

*Digest* mailed questionnaires to 10 million people, whose names were from country club membership lists, phone books, and automobile registrations.

George Gallup, polling 50,000 people, predicted *Digest's* results in advance.

*3rd party candidates were excluded in the numbers below.*

**Sampling error** was small due to large sample.

---

	Roosevelt's percentage
The election result	62
<i>Digest's</i> prediction	43
Gallup's prediction of <i>Digest</i>	44
Gallup's prediction of election	56

---

□

## 7.3 Sampling Distributions

**Definition:** The *probability distribution* of a **statistic** is called its **sampling distribution**.

Hence, the **sampling distribution** of a *statistic* consists of the possible values of the *statistic* along with their associated probabilities.

Herein, we focus on the sampling distributions of  $\bar{X}$  and  $p$ .

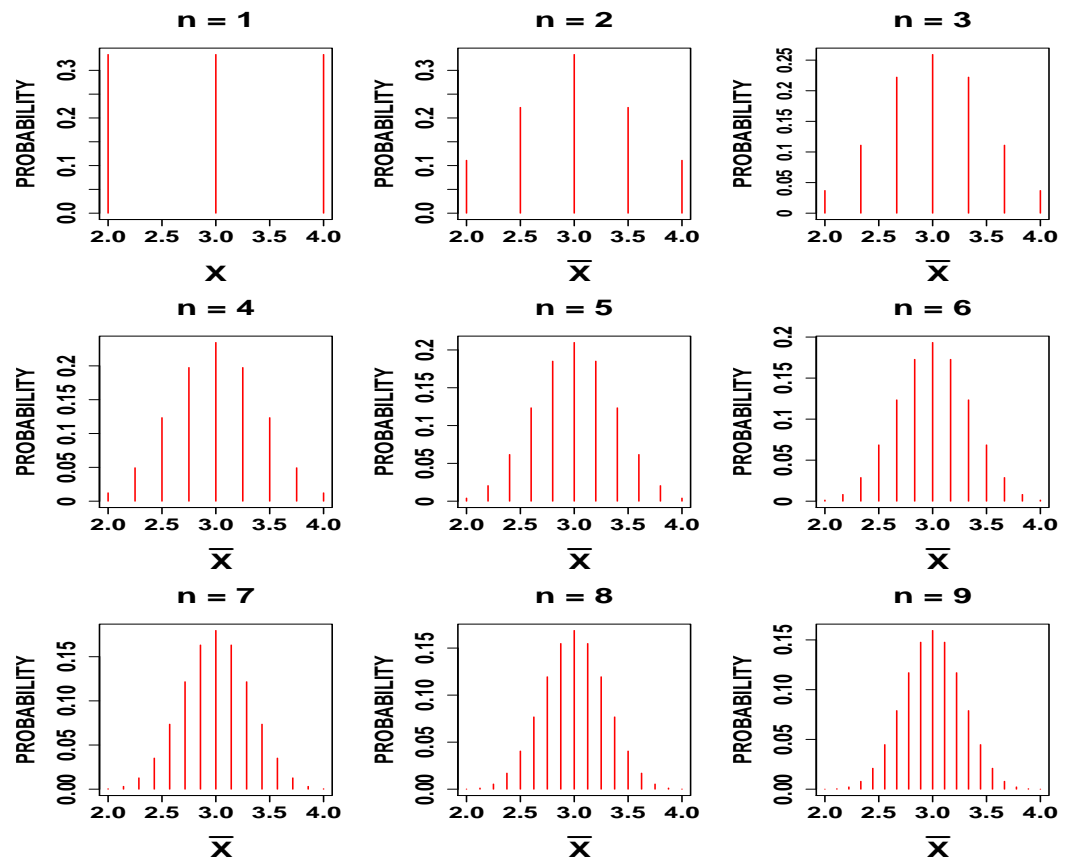
## 7.4 Sampling Distribution of the Mean

**Example:** Consider a population consisting of three cards, which are labeled as  $\boxed{2}$ ,  $\boxed{3}$ , and  $\boxed{4}$ . Let  $x$  be the value of a card drawn.

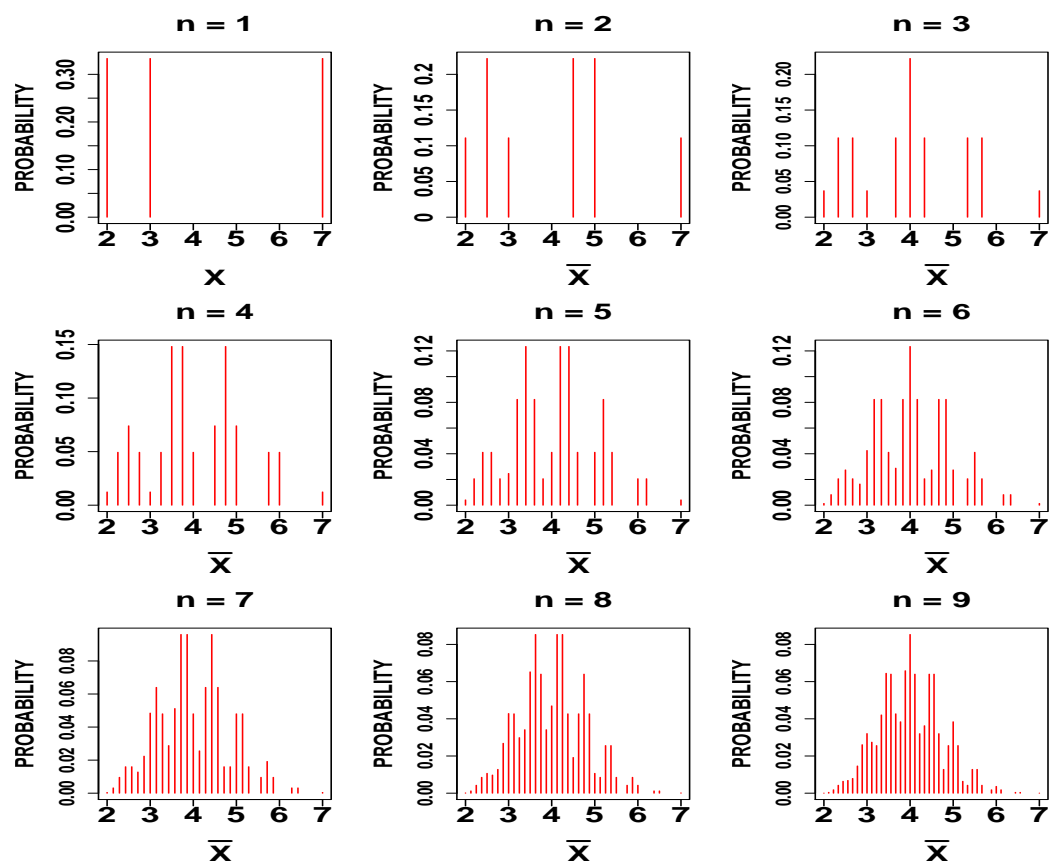
(a) Determine the **probability distribution** of  $X$ .

(b) Graph the *probability distribution* of  $X$ .

- (c) Determine the *mean* of  $X$ .
- (d) Let  $\bar{X}$  be the sample mean, based on **two** observations independently sampled (i.e., **with** replacement) from this population. Determine the **sampling distribution** of  $\bar{X}$ .
- (e) Graph the *sampling distribution* of  $\bar{X}$ .
- (f) Determine the *mean* of  $\bar{X}$ .
- (g) Additional graphs of the *sampling distribution* of  $\bar{X}$  are below, based on independent observations and sample size  $n$ .



(h) Repeat part (g), using cards labeled  $\boxed{2}$ ,  $\boxed{3}$ , and  $\boxed{7}$ .



□

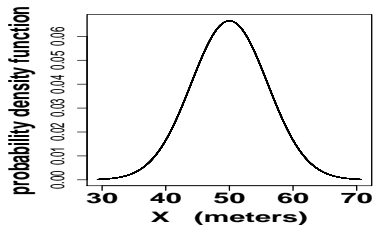
Case A: Sample **with** replacement. Hence, observations are independent.

Case B: Sample **without** replacement, but the population size is quite large compared to  $n$ . Hence, observations are nearly independent.

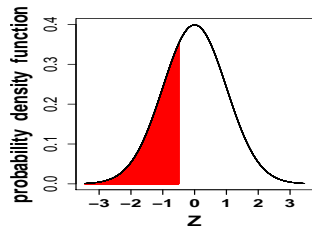
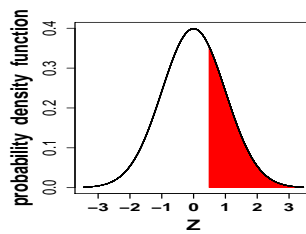
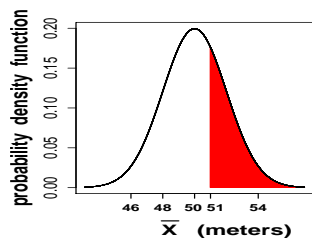
- (a)  $\mu_{\bar{X}} = \mu$  always.
- (b)  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$  (called the **standard error** of  $\bar{X}$ ), exactly for Case A and approximately for Case B.
- (c) (A version of the Central Limit Theorem) The sample mean,  $\bar{X}$ , is approximately normally distributed for Cases A and B (and positive finite  $\sigma$ ), for **large**  $n$  (usually  $n \geq 30$ , if neither tail of the distribution is too heavy).

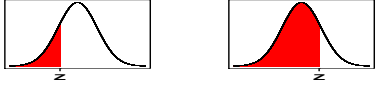
- (d) (A special case) The sample mean,  $\bar{X}$ , is approximately normally distributed for Cases A and B (and positive finite  $\sigma$ ), if the **original population** is approximately **normally distributed** (for **any** sample size  $n$ ).

**Example:** Suppose  $X \sim N(\mu = 50 \text{ meters}, \sigma = 6 \text{ meters})$ . Sample nine independent observations of  $X$ .



- (a) Determine the *mean* of  $\bar{X}$ .
- (b) Determine the *standard deviation* of  $\bar{X}$ ; i.e., the *standard error* of  $\bar{X}$ .
- (c) Determine the probability that  $\bar{X}$  exceeds 51 meters.



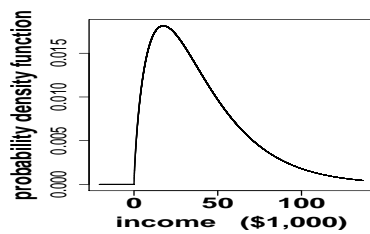


The Cumulative Standard Normal Distribution, pp. 914–915, Table E.2

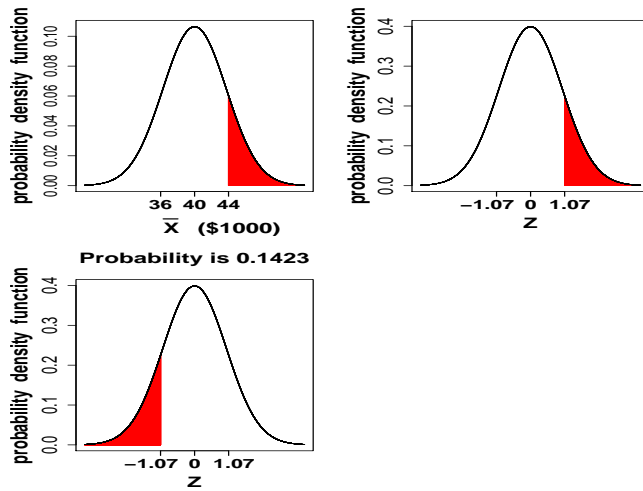
Cumulative Probabilities										
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

**Example:** Suppose personal income,  $X$ , in the U.S. has mean  $\mu = \$40,000$  and standard deviation  $\sigma = \$30,000$ . Sample **without** replacement.



(a) Determine  $P(\bar{X} > \$44,000)$ , for  $n = 64$ .

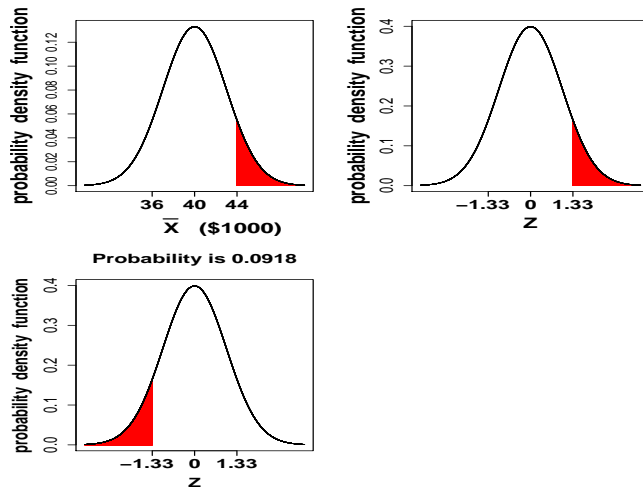


The Cumulative Standard Normal Distribution, pp. 914–915, Table E.2

Cumulative Probabilities										
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(b) Determine  $P(\bar{X} > \$44,000)$ , for  $n = 100$ .

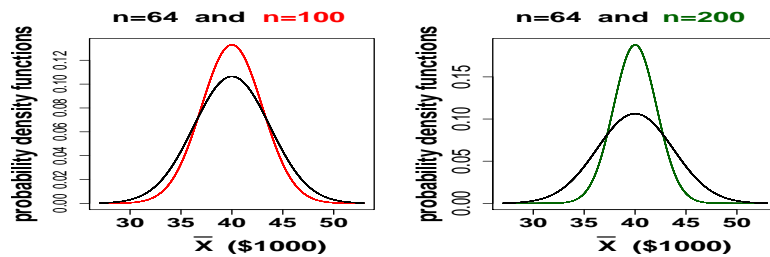




The Cumulative Standard Normal Distribution, pp. 914–915, Table E.2

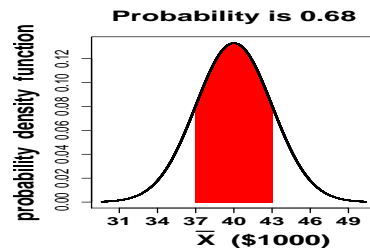
Cumulative Probabilities										
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(c) What happens to  $P(\bar{X} > \$44,000)$  as we increase  $n$  to 200?

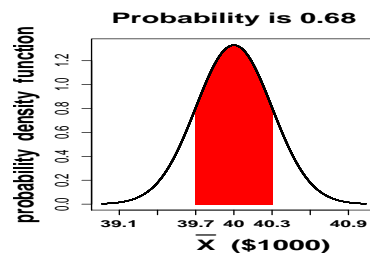


(d) Determine  $P(\bar{X} > \$44,000)$ , for  $n = 10$ .

- (e) Determine the 68% part of the empirical rule for  $n = 100$ .



- (f) Determine the 68% part of the empirical rule for  $n = 10,000$ .



□

## 7.5 Sampling Distribution of the Proportion

A proportion is a special case of a mean.

**Example:** Sample *independent* observations from a population which is 30% Democrat. Let  $p$  be the sample proportion of Democrats.

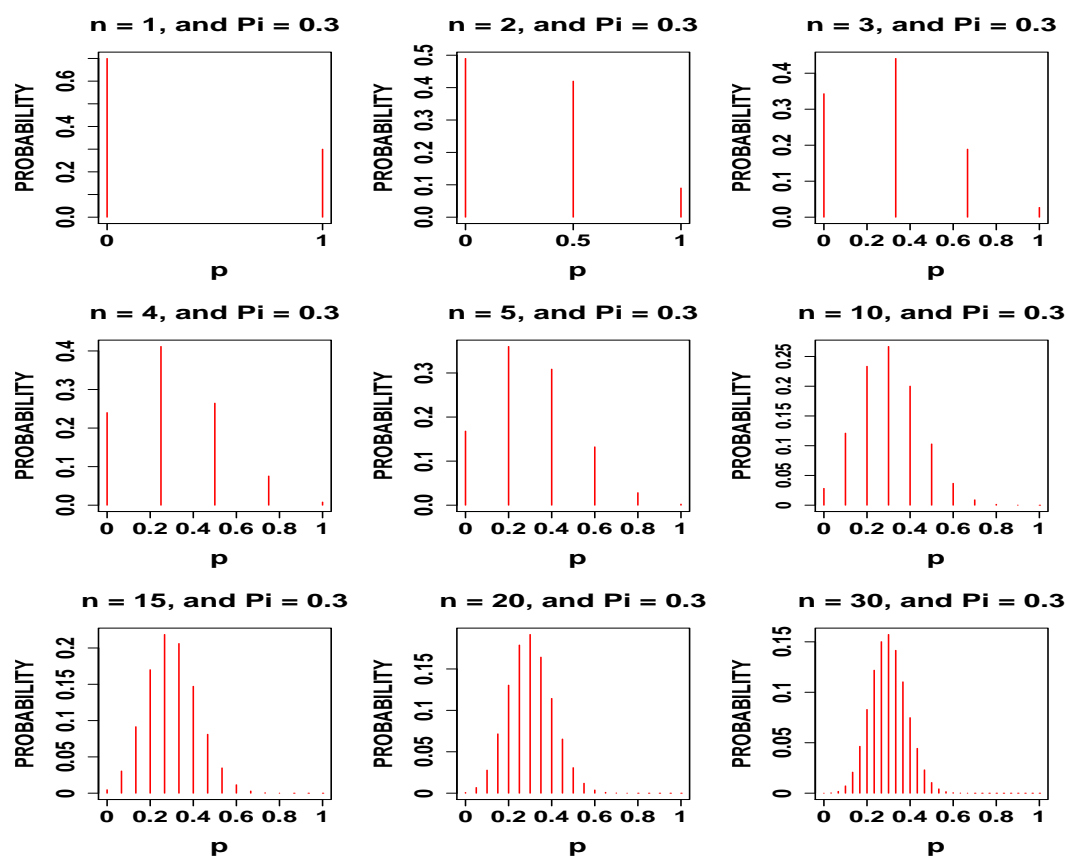
- (a) State the **population distribution** in a chart, and construct the *line graph* of the **population distribution**.

Let  $X = 0$  if non-Democrat, and  $X = 1$  if Democrat.

Note that the *sampling distribution* of  $p$  for  $n = 1$  is the same as the *population distribution* of  $X$ .

- (b) For  $n = 2$ , state the **sampling distribution** of  $p$  in a chart, and construct the *line graph* of the **sampling distribution** of  $p$ .

Recall the graphs from section 5.3.



(c) What happens to the *sampling distribution* of  $p$  as the sample size,  $n$ , gets larger?

□

**Example:** *Virginians who exercise.* According to the Centers for Disease Control and Prevention, about 48% of Virginian adults achieved the recommended level of physical activity.

*Recommended physical activity is defined as “reported moderate-intensity activities (i.e., brisk walking, bicycling, vacuuming, gardening, or anything else that causes small increases in breathing or heart rate) for at least 30 minutes per day, at least 5 days per week or vigorous-intensity activities (i.e., running, aerobics, heavy yard work, or anything else that causes large increases in breathing or heart rate) for at least 20 minutes per day, at least 3 days per week or both. This can be accomplished*

through lifestyle activities (i.e., household, transportation, or leisure-time activities).”

<http://apps.nccd.cdc.gov/PASurveillance/StateSumV.asp?Year=2001>

[www.cdc.gov/nccdphp/dnpa/physical/stats/us\\_physical\\_activity/index.htm](http://www.cdc.gov/nccdphp/dnpa/physical/stats/us_physical_activity/index.htm)

Take a sample of size  $n = 100$ , and let  $X$  be the number who achieved the recommended level of physical activity. What is the distribution of  $X$ ?

□

Case A: Sample **with** replacement. Hence, observations are independent.

Case B: Sample **without** replacement, but the population size is quite large compared to  $n$ . Hence, observations are nearly independent.

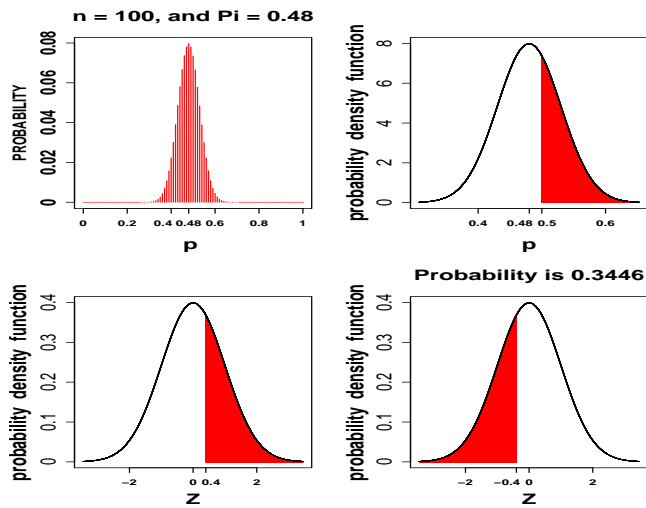
If  $n$  is a small percentage of the population size, then sampling **without** replacement is similar to sampling **with** replacement, since sampling the same person more than once would be quite unlikely.

(a)  $\mu_p = \pi$  always.

(b)  $\sigma_p = \sqrt{\pi(1 - \pi)/n}$  (called the **standard error** of  $p$ ), exactly for Case A and approximately for Case B.

(c) (A version of the Central Limit Theorem) The sample proportion  $p$  is approximately normal if {rule of thumb}  $n\pi \geq 5$  and  $n(1 - \pi) \geq 5$ , for Cases A and B.

**Example:** *Revisit Virginians who exercise.* Determine the probability that a majority of Virginians in a sample of size 100 achieve the recommended level of physical activity.



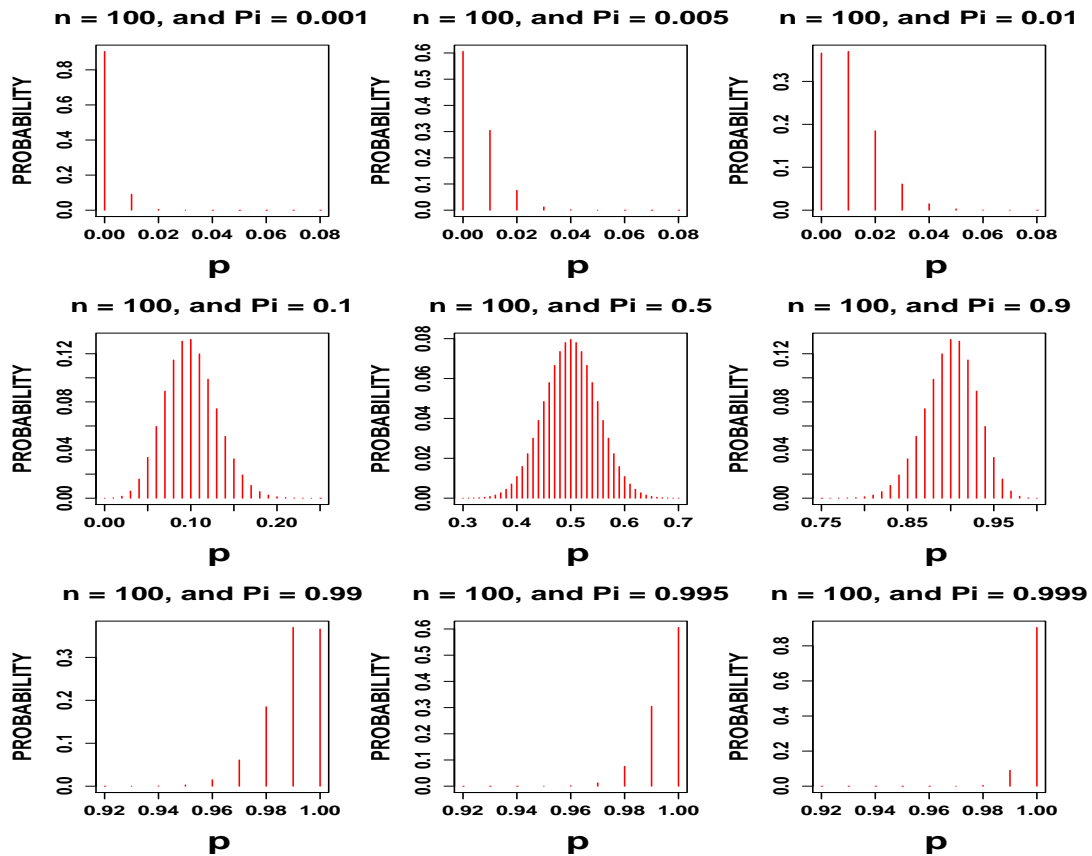
The Cumulative Standard Normal Distribution, pp. 914–915, Table E.2

Cumulative Probabilities										
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

## Why is the rule of thumb needed?

**Example:** Consider the *sampling distribution* of  $p$ , for  $n = 100$  and various  $\pi$ .



□

Read p. 287, Appendix E7, Using Microsoft Excel for Sampling and Sampling Distributions.