Chapter 3

What Do Samples Tell Us?

A *parameter* is a numerical quantity calculated from a population, whereas a *statistic* is a numerical quantity calculated from a sample. The value of a statistic is used to estimate an unknown parameter.

A statistic which estimates a parameter closely is considered better than one that does not do this, so a "good" sampling method used to estimate the parameter from a statistic has both small *bias* and small *variability*. Bias arises if, as explained earlier, sampling is *non*-random which leads to systematic error where the statistic value either overstates or understates the true parameter value and, so, to reduce bias it makes sense to sample at random, to use a statistic based on a simple random sample (SRS) to estimate the parameter. Furthermore, variability of a statistic based on a SRS should be reduced; that is, the average statistic values should be made consistently close to one another (although not necessarily the parameter), by taking large samples, by increasing n.

Both small bias and small variability are required together, not just one or the other, for a "good" statistic. As an analogy, a gun shot (statistic) is a "good" one if the gun is not only aimed close to the target (parameter) and not somewhere else, has small bias, but also each shot lands consistently close to one another, has small variability.

Correcting (reducing) bias by using a SRS only works if the statistic value is measured correctly, there is no *measurement (instrument) error*, and also if the statistic actually measures what it is supposed to measure, namely, the parameter, it does not accidently measure the wrong parameter, it is a *valid* statistic. So bias can persist even if a SRS is used. Furthermore, an estimate is *reliable* if, under repeated sampling, closely similar estimate values occur, or, in other words, there is little variability in the estimate values. Validity, reliability and measurement error are discussed in greater detail later on.

Margin of error measures variability in a sampling; it says how close the average of a statistic is to a parameter, assuming there is no measurement error and the statistic is valid. There are different types of margin of error, but, as discussed in this chapter,

the 95% confident margin of error for population proportion parameter p using sample proportion statistic $\hat{p}$ from SRS of size n is roughly $\frac{1}{\sqrt{n}}$. A *(quick method) 95% confidence statement* for p equals $\hat{p} \pm \frac{1}{\sqrt{n}}$ and means 95% of confidence statements will include p under repeated sampling. A more sophisticated 95% confidence statement, called confidence *interval*, for p is given later on.

Exercise 3.1 (What Do Samples Tell Us?)

1. *Parameter and statistic: smoking.* What proportion of smokers (ignore the nonsmokers for the moment) have heart attacks, lung cancer and other diseases (read: poor health)? Since 345 of one thousand randomly chosen smokers have poor health, it seems reasonable to infer approximately $\frac{345}{1000}$ths or 34.5% of **all** smokers have poor health.

   (a) *Population* is
      i. *all* smokers and nonsmokers.
      ii. *all* nonsmokers.
      iii. the one thousand smokers, selected at random.
      iv. *all* smokers.

   (b) *Sample* is
      i. one thousand smokers and nonsmokers, selected at random.
      ii. *all* nonsmokers.
      iii. one thousand smokers, selected at random.
      iv. *all* smokers.

   (c) Since *statistic* summarizes *sample* data, it is
      i. proportion with poor health, among *all* smokers.
      ii. proportion with poor health, among 1000 randomly chosen smokers.

   (d) Value of statistic is **34.5%** / **51%** / **74%**.

   (e) Since *parameter* summarizes *population* data, it is
      i. proportion with poor health, among *all* smokers.
      ii. proportion with poor health, among 1000 randomly chosen smokers.

   (f) **True** / **False** Statistic is *known*; it is 34.5%. On the other hand, parameter is *unknown*. Often, statistic used to *estimate* unknown parameter.

   (g) Match columns.

statistical terms	smoker example
(a) population	(A) proportion with poor health, among all smokers
(b) sample	(B) proportion with poor health, among 1000 chosen
(c) statistic	(C) all smokers
(d) parameter	(D) 1000 smokers

terms	(a)	(b)	(c)	(d)
smoker example				

2. *Parameter and statistic: average height of Americans.* Since one thousand randomly chosen Americans have an average height of 5.6 feet tall, it seems reasonable to infer average height of **all** Americans is around 5.6 feet tall. Match columns.

terms	political preference example
(a) population	(A) 1000 Americans
(b) sample	(B) average height, among all Americans
(c) statistic	(C) all Americans
(d) parameter	(D) average height, among 1000 chosen

terms	(a)	(b)	(c)	(d)
example				

3. *Parameter and statistic: average commute distances.* At PNW, 120 students are randomly surveyed from entire 11,500 and asked their commute distance to campus. Average of 9.8 miles is computed from 120 selected. We infer from data *all* students have 9.8 average commute.

terms	travel example
(a) population	**(A)** average commute distance for 120 students
(b) sample	**(B)** all students at PNW
(c) statistic	**(C)** commute distance for any PNW student
(d) parameter	**(D)** average commute distance for all students
	(E) 120 students
	(F) 8 mile commute distance for a particular student

terms	(a)	(b)	(c)	(d)
travel example				

4. *Variability and bias: darts* Describe dart hit patterns on following four dart boards.

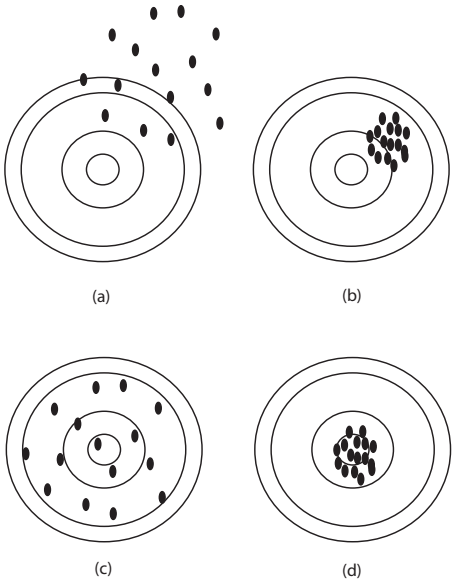


Figure 3.1 (Variability and bias)

(a) **True** / **False** Dart hits on board (a) have *large variability* because they are spread out and *large bias* because they are thrown in wrong place, up and away from bull's eye.

(b) Dart hits on board (b): Clustered together but thrown in wrong place, so **small** / **large** variability and **small** / **large** bias

(c) Dart hits on board (c): Spread out but thrown in right place, so **small** / **large** variability and **small** / **large** bias

(d) Dart hits on board (d): Clustered together and thrown in right place, so **small** / **large** variability and **small** / **large** bias

(e) If bull's eye represents parameter and dart hits represent different statistic values from repeated sampling, best dart hit pattern occurs on board (choose one) **(a)** / **(b)** / **(c)** / **(d)**

5. *More variability and bias: political preference.* It is known 68% of registered voters in Berrien County, Michigan, are registered as Democrats. We call 500 Berrien County voters at random and ask their party. We do this 5 times. Results are 59.2%, 58.9%, 60.5%, 57.4% and 61.3% Democratic.

   (a) Sampling method appears to have (choose one)

      (i) large variability and large bias

      (ii) large variability and small bias

      (iii) small variability and large bias

      (iv) small variability and small bias

Hint: 68% is the *parameter* ("bull's eye") and 59.2%, 58.9%, 60.5%, 57.4% and 61.3% are *statistic* values ("dart hits"). Recall, variability describes how clustered together the statistic values (dart hits) are and bias describes how close the statistic values (dart hits) are to the parameter (bull's eye).

(b) *Sample size, n.*
Number of voters in each sample, $n =$ (choose one) **5** / **68%** / **500**.

(c) *Number of repetitions (simulations).*
Number of samples of size $n = 500$ taken: (choose one) **5** / **10%** / **100**.

(d) It is usually the case the parameter value (actual proportion of voters who are Democrats, p, in this case) is (choose one) **known** / **unknown**.
If we knew p, what would be the point of sampling?

(e) What is wrong? Choose one.

(i) Nothing, sampling (calling) method (instrument) is fine, no bias.

(ii) Calling method probably incorrect, bias due to measurement (instrument) error.

(iii) Measuring wrong statistic, should be measuring proportion of Republicans instead, so bias due to a validity problem.

6. *Bias-Variability trade-off: reading ability versus brightness.* Curves (a), (b) and (c) are three possible sample-based models of the scatterplot of data. Which of these sample-based models best fits the (one) population-based model?

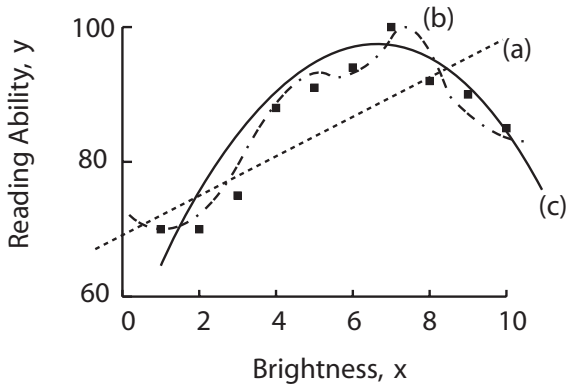


Figure 3.2 (Bias-Variability Trade-Off)

(a) All curves (a), (b) and (c) are (choose one)

(i) "exact" models, based on population of all data.

(ii) "approximate" models, based on sample of data.

(b) If only *one* scatterplot of data, the best-fitting sample-base model is

(i) dashed linear line (a).

(ii) dashed and dotted curved line (b).

(iii) solid curved line (c).

(c) However, the pattern of the scatterplot of a second *random* sample of data from the *same* population would probably be (choose one)

(i) exactly the same as the first scatterplot pattern.

(ii) similar but not exactly the same as the first scatterplot pattern.

(iii) radically different to the first scatterplot pattern.

(d) Sample-based dashed linear line (a): Not curvy enough, so
small / **large** variability because (a) will tilt a little for each random set of points
small / **large** bias because (a) is far from population curve

(e) Sample-based dashed and dotted (sample) curved line (b): Too curvy so
small / **large** variability because (b) will change shape a lot for each random set of data
small / **large** bias because (b) is (fairly) close to population curve

(f) Sample-based solid (sample) curved line (c): Curved just right, so
small / **large** variability because (c) change shape a little for each random set of data
small / **large** bias because (c) is (fairly) close to the correct population curve

(g) Consequently, with repeated sampling, best-fitting sample-based model is

(i) dashed linear line (a).

(ii) dashed and dotted curved line (b).

(iii) solid curved line (c).

7. *Measuring variability: confidence statement (quick method)* $\hat{p} \pm \frac{1}{\sqrt{n}}$ *for* p. Proportion, p, of voters in Berrien County registered as Democrats is *un*known.

(a) *Repeatedly calculating* $\hat{p}$. We call 100 Berrien County voters in a SRS (simple random sample) and ask their party. We repeat this 5 times. In first sample of 100, we find 56, or $\hat{p} = \frac{56}{100} = 0.56$ or 56% are Democrats. In remaining samples, we find 49, 45, 51 and 51, respectively, of 100 voters are Democrats or, in other words, 49%, 45%, 51% and 51%. Summary of proportions given in following table.

class	number	proportion	%
45% to 49%	2	$\frac{2}{5} = 0.4$	40%
50% to 54%	2	$\frac{2}{5} = 0.4$	40%
55% to 59%	1	$\frac{1}{5} = 0.2$	20%
total	5	1.0	

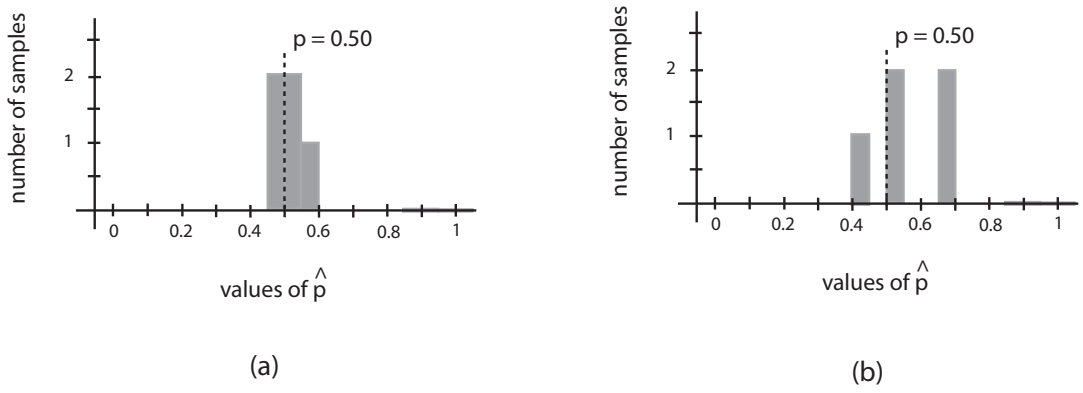


Figure 3.3 (Histograms for sample proportions, $\hat{p}$)

Which histogram best represents this information? Choose one: **(a)** / **(b)**

(b) *Variability of $\hat{p}$.* Variability of $\hat{p}$ about proportion p measured by *margin of error* $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{100}} =$ (choose one) **10%** / **11%** / **13%**.

Margin of error measures how close a *sample* proportion of Democrats, $\hat{p}$, is to the *unknown* actual proportion of Democrats, p; in this case, $\hat{p}$ is within 10% of p. But p is *unknown*, so margin of error 10% cannot be known exactly.

(c) *Confidence statement for p.*
If $\hat{p} = 0.56$ or 56%, 95% confidence interval for p is given by
$\hat{p} \pm \frac{1}{\sqrt{n}} = 0.56 \pm \frac{1}{\sqrt{100}} = 0.56 \pm 0.10 = (0.56 - 0.10, 0.56 + 0.10) =$
(39%, 59%) / **(41%, 61%)** / **(46%, 66%)**.
So, 95% *confident* parameter p in interval $(46\%, 66\%)$.

(d) *Another confidence statement for p.*
If $\hat{p} = 0.49$ or 49%, 95% confidence interval for p is given by
$\hat{p} \pm \frac{1}{\sqrt{n}} = 0.49 \pm \frac{1}{\sqrt{100}} = 0.49 \pm 0.10 = (0.49 - 0.10, 0.49 + 0.10) =$
(39%, 59%) / **(41%, 61%)** / **(46%, 66%)**.
So, 95% *confident* parameter p in interval $(39\%, 59\%)$.

(e) *And another confidence statement for p.*
If $\hat{p} = 0.51$ or 51%, 95% confidence interval for p is given by
$\hat{p} \pm \frac{1}{\sqrt{n}} = 0.51 \pm \frac{1}{\sqrt{100}} = 0.51 \pm 0.10 = (0.51 - 0.10, 0.51 + 0.10) =$
(39%, 59%) / **(41%, 61%)** / **(46%, 66%)**.
So, 95% *confident* parameter p in interval $(41\%, 61\%)$.

Which of (46%,66%), (39%,59%) or (41%,61%) is "correct"; that is, which one includes unknown actual proportion of Democrats, p? Although we do not know with 100% certainty exactly where the unknown p is, it is possible, under some assumptions (most importantly, samples are chosen at random), to say 95% (in this case) of *all* possible intervals (including the three here) based on all possible sample proportions, $\hat{p}$, do include p. Which *particular* interval includes p is unknown, though.

(f) *Number of repetitions (simulations).*
Number of samples of size $n = 100$ taken: (choose one) **5** / **10%** / **100**.

8. *Sample size and variability: political preference again.* We call $n = 100$ Berrien County voters at random and ask if their party is Democrat. We repeat this 1000 times. One thousand sample proportions are calculated and combined in histogram (a) of figure. Same, but $n = 2500$ for histogram (b) in figure.

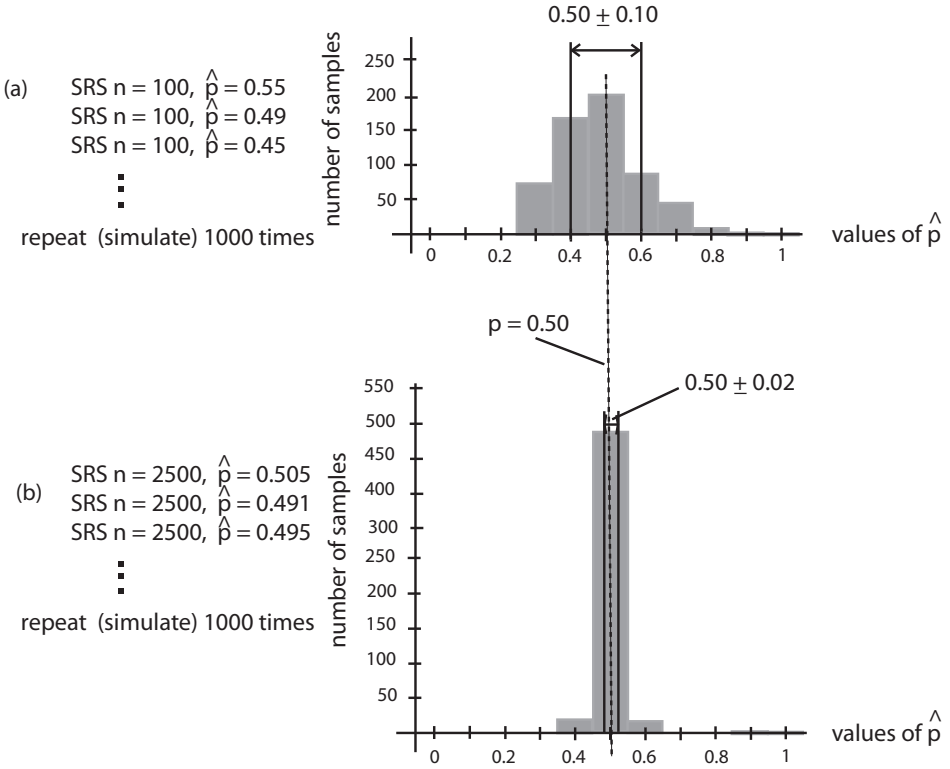


Figure 3.4 (Simulation, variability and bias)

(a) *Simulate SRS* $n = 100$ *(so* $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.10$*) for* $\hat{p}$ *thousand times.*
If $\hat{p} = 0.50$ or 50%, 95% confidence interval for p is given by
$\hat{p} \pm \frac{1}{\sqrt{n}} = 0.50 \pm \frac{1}{\sqrt{100}} = 0.50 \pm 0.10 = (0.50 - 0.10, 0.50 + 0.10) =$
(39%, 59%) / (40%, 60%) / (48%, 52%).

(b) *Simulate SRS* $n = 2500$ *(*$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{2500}} = 0.02$*) for* $\hat{p}$ *thousand times.*
If $\hat{p} = 0.50$ or 50%, 95% confidence interval for p is given by
$\hat{p} \pm \frac{1}{\sqrt{n}} = 0.50 \pm \frac{1}{\sqrt{2500}} = 0.50 \pm 0.02 = (0.50 - 0.02, 0.50 + 0.02) =$
(39%, 59%) / (40%, 60%) / (48%, 52%).

(c) As sample size, n, increases 100 to 2500, variability **increases / decreases**.
margin of error for $n = 2500$ $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{2500}} = 0.02$ smaller than for $n = 100$, $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.10$
in other words, 95% confidence interval for $n = 2500$ shorter (more "precise") than for $n = 100$

9. *Sample size and variability; random sampling and bias.*

(a) To reduce bias of estimates from a sample (choose one):

(i) use a small sample.
(ii) use voluntary response sampling.
(iii) use a convenience sampling.
(iv) use a SRS.
(v) use a large sample.

bias occurs in sampling when one individual is favored over another; *random* sampling prevents this
assuming the statistic is measured correctly and measures what it is supposed to measure

(b) To reduce variability of estimates from a SRS (choose one):

(i) increase number of simulations.
(ii) increase bias.
(iii) use a convenience sampling.
(iv) use a SRS.
(v) use a large sample.

10. *95% confidence statements of proportion: decayed teeth.*

(a) $n = 500$. If 345 of a SRS of $n = 500$ children have decayed teeth,

i. sample proportion is $\hat{p} = \frac{345}{500} =$ (choose one) **69%** / **71%** / **73%**
ii. variability in sample proportion, measured by margin of error, is $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{500}} \approx$ (choose one) **3.4%** / **4.5%** / **5.4%**
iii. 95% confidence statement of parameter proportion $69\% \pm 4.5\% \approx$ **(64.5%, 73.5%)** / **(63.5%, 74.5%)** / **(62.5%, 75.5%)**

(b) $n = 3000$. If 2070 of a SRS of $n = 3000$ children have decayed teeth,

i. sample proportion is $\hat{p} = \frac{2070}{3000} =$ (choose one) **69%** / **71%** / **73%**
where, notice, sample proportion same as before
ii. variability in sample proportion, measured by margin of error, is $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{3000}} \approx$ (choose one) **0.4%** / **1.8%** / **2.4%**
where, notice, margin of error smaller than before because sample size $n = 3000 > n = 500$
iii. 95% confidence statement of parameter proportion $69\% \pm 1.8\% \approx$ **(67.2%, 70.8%)** / **(63.5%, 74.5%)** / **(62.5%, 75.5%)**
where confidence statement shorter than before again because $n = 3000 > n = 500$
iv. We are **more** / **less** certain of $\hat{p} = 69\%$ when $n = 3000$ than $n = 500$.

11. *More 95% confidence statements and margin of errors of proportion.*

(a) *Defective widgets.* If 551 of a SRS of $n = 1500$ widgets defective,

i. sample proportion is $\frac{551}{1500} =$ (choose one) **19.0%** / **27.1%** / **36.7%**
ii. variability in sample proportion, measured by margin of error, is $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1500}} \approx$ (choose one) **2.6%** / **3.5%** / **4.4%**

iii. (quick method) 95% confidence statement of parameter proportion $36.7\% \pm 2.6\% =$ **(24.5%, 53.5%)** / **(34.1%, 39.3%)**

(b) *Comparing margin of errors for two samples.* A poll of 190 statisticians and 421 mathematicians revealed 13% of statisticians and 21% of mathematicians said they had difficulty spelling. Which statement is true?

(i) Less variability in sample proportion of poor-spelling statisticians than poor-spelling mathematicians,

(ii) More variability in sample proportion of poor-spelling statisticians than poor-spelling mathematicians,

(iii) Same variability in sample proportion of poor-spelling statisticians than poor-spelling mathematicians,

more variability in statisticians because $\frac{1}{\sqrt{190}} \approx 0.073 > \frac{1}{\sqrt{421}} \approx 0.049$

(c) *Comparing margin of errors for two different populations.* If $n = 750$ are polled in New York city, with population of about 8.3 million, and $n = 750$ are polled in Valparaiso, with population of about 29,000, the margin of error of a 95% confidence statement of population proportion interested in basketball is (choose one)

(i) biggest in New York city,

(ii) biggest in Valparaiso,

(iii) same for New York and Valparaiso,

same in both cases because $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{750}} \approx 0.037$

12. *Some final comments.*

(a) Confidence statement conclusion always applies to **sample** / **population**

Confidence statement is calculated from a sample and a sample is collected to tell us something about the population.

(b) 95% confidence statement conclusion always **certain** / **uncertain**

95% confidence statement confident only 95%, not 100% of the time.

(c) 99% confidence statement conclusion **more certain, but longer** / **more certain and shorter** than 95% confidence statement

Assuming the sample size is the *same* for both statements, the 99% confidence statement is more confident of capturing the parameter *because* it is longer.

Chapter 4

Sample Surveys in the Real World

Sampling errors and nonsampling errors occur when sampling in the real world. *Sampling errors* arise from act of sampling and include, for example, *random* sampling error (measured, previously, by the margin of error) and *undercoverage*. Undercoverage, which might lead to bias, occurs when individuals are left out of a sampling *frame*, a list from which the sample is chosen. *Nonsampling errors* have nothing to do with choosing a sample, often lead to bias which cannot be reduced by SRS, occur even in a census and include, for example, nonresponse, processing, response and survey wording errors. *Nonresponse error* occurs when subjects either refuse are cannot be contacted for a survey. *Processing errors* occurs when data is mishandled. *Response errors* occur when subjects give incorrect answers. *Wording errors* occur when survey questions are improperly or incorrectly worded.

In addition to simple random sampling (SRS), other *probability sampling* techniques include stratified sampling, cluster sampling and systematic sampling. *Stratified sampling* involves dividing population into strata and choosing a simple random sample (SRS) from each strata. *Cluster sampling* involves dividing population into clusters, choosing a subset of these clusters at random, and then using either SRS or all items of selected clusters. *Systematic sampling* involves selecting every kth item from population, where first item is chosen at random. In general, these probability sampling techniques reduce variability (improve reliability) but at the expense of increasing bias somewhat.

Exercise 4.1 (Sample Surveys in the Real World)

1. *Sampling error or nonsampling error?* Match embryonic stem cell research survey examples with types of errors.

   (a) **Sampling / Nonsampling**
   Survey, a SRS of tax payers in Michigan City, has 3% margin of error.

   (b) **Sampling / Nonsampling**
   Many, about 75%, of individuals contacted refuse to answer survey.

(c) **Sampling / Nonsampling**
Computer glitch causes one third of survey forms to disappear.

(d) **Sampling / Nonsampling**
Individuals uncomfortable truthfully answering sensitive moral question.

(e) **Sampling / Nonsampling**
Survey question is worded: "Which is more important: research that might result in new medical cures or not destroying live human embryos?"

(f) **Sampling / Nonsampling**
Survey frame includes households, but excludes students in college residences, prison inmates and individuals in shelters.

(g) **Sampling / Nonsampling**
Survey frame is city telephone list.

2. *Random or undercoverage sampling error?* Match embryonic stem cell research survey examples with *sampling* types of errors.

(a) **Random / Undercoverage**
Survey, based on carefully done SRS, has 3% margin of error.

(b) **Random / Undercoverage**
Survey frame includes households, but excludes students in college residences, prison inmates and individuals in shelters.

(c) **Random / Undercoverage**
Survey frame is city telephone list.

(d) **Random / Undercoverage**
Only people on nearby street interviewed.

(e) **Random / Undercoverage**
Survey frame is list of email addresses.

3. *Nonresponse, processing, response or wording nonsampling errors?* Match embryonic stem cell research survey examples with types of errors.

(a) **Nonresponse / Processing / Response / Wording**
Computer glitch causes one third of survey forms to disappear.

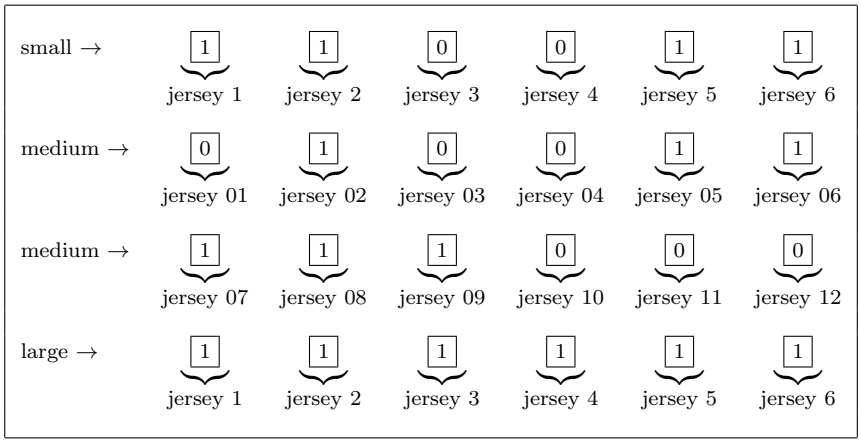
(b) **Nonresponse / Processing / Response / Wording**
Individuals uncomfortable truthfully answering sensitive moral question.

(c) **Nonresponse / Processing / Response / Wording**
Many, about 75%, of individuals contacted refuse to answer survey.

(d) **Nonresponse / Processing / Response / Wording**
Survey question is worded: "Which is more important: research that might result in new medical cures or not destroying live human embryos?"

(e) **Nonresponse** / **Processing** / **Response** / **Wording**
Individual does not answer door after five attempts.

4. *Stratified Random Sample: Jerseys.* Twenty–four football jerseys are arranged according to size in three strata, large, medium and small, and identified as being made of polyester (indicated by a "1") or of a polyester–cotton (indicated as a "0") blend, as given in following (tiny) population box model.



Estimate population proportion of polyester-made jerseys with observed sample proportion of jerseys using stratified sampling method.

(a) *Terminology.*
If "jersey size" is a *stratum*, there are **one** / **two** / **three** / **four** strata.
Second stratum has **twelve** / **thirteen** / **sixteen** / **eighteen** jerseys.
Each stratum is **homogenous** / **heterogeneous**.

(b) *Stratified Sample.*
Possible stratified sample of *eight* jerseys is (choose one *or more*)

i. one (1) small, two (2) medium and three (3) large jerseys,
ii. two (2) small, three (4) medium and three (2) large jerseys,
iii. two (2) small, four (3) medium and two (3) large jerseys,

where jerseys are chosen using SRSs from each strata.

(c) *Population Proportion.*
Proportion of *all* tickets in box with 1s represents (circle one)

i. sample proportion of polyester-made jerseys,
ii. population proportion of polyester-made jerseys,

and equal to: (expected) proportion $= \frac{1+1+0+\cdots+1}{24} \approx$ **0.67** / **0.77** / **0.89**.
Proportion (0.67, in this case) is example of **statistic** / **parameter**.

5. *Cluster Random Sample: Jerseys Again.*

small →	1	1	0	0	1	1
	jersey 01	jersey 02	jersey 03	jersey 1	jersey 1	jersey 1
medium →	0	1	0	0	1	1
	jersey 04	jersey 05	jersey 06	jersey 2	jersey 3	jersey 4
medium →	1	1	1	0	0	0
	jersey 07	jersey 08	jersey 09	jersey 3	jersey 5	jersey 6
large →	1	1	1	1	1	1
	jersey 10	jersey 11	jersey 12	jersey 4	jersey 7	jersey 8
	cluster 1			cluster 2	cluster 3	

Once again, estimate population proportion of polyester-made jerseys with observed sample proportion of jerseys but, this time, use cluster sampling method.

(a) *Terminology.*
There are **one** / **two** / **three** / **four** clusters.
Second cluster has **four** / **nine** / **twelve** / **eighteen** jerseys.
Each cluster is **homogenous** / **heterogeneous**.

(b) *Cluster Sample.*
Possible cluster sample of *two* is (choose *one or more*!)

i. small jersey and large jersey clusters,
ii. small jersey cluster,
iii. small jersey and medium jersey and large jersey clusters,

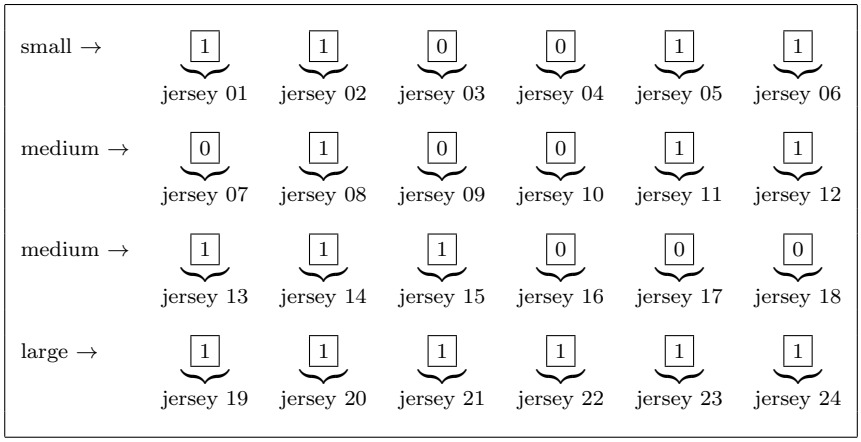
where all jerseys in chosen clusters are used in sample.

A cluster sample does not require *all* jerseys are chosen from selected clusters. Often a SRS is chosen from each selected cluster. Only one or two clusters can be chosen from three clusters in cluster sampling; if all three clusters are chosen, sampling method becomes stratified sampling.

(c) *Population Proportion.* Population proportion of polyester-made jerseys, proportion of *all* tickets in box with 1s, **remains same** / **changes** because box is same as before and so proportion $= \frac{1+1+0+\cdots+1}{24} \approx 0.67$.

(d) *Cluster sampling versus stratified sampling.*
In cluster sampling, a *subset* of all clusters are sampled from, whereas in stratified sampling, (circle one) **some** / **all** strata are sampled from.

6. *Systematic Sampling: Jerseys Again.* Once again, estimate population proportion of polyester-made jerseys with observed sample proportion of jerseys using systematic sampling method. Jerseys are re–numbered, as given in box model.



(a) *Systematic Sample and Proportion.*
After choosing number between 1 and 6 at random, 5, say, this number and every 4th number after this (5, 9, 13, 17, 21) is chosen from jersey box model. Sample proportion of polyester-made for five jerseys is
(observed) proportion $= \frac{1+0+1+0+1}{5} =$ (circle one) **0.4** / **0.6** / **0.7** / **0.8**.
Sample proportion (0.6, in this case) is example of **statistic** / **parameter**.

(b) *Another Systematic Sample.*
If starting number 4 is used, systematic sample is (choose one)

i. 1, 5, 9, 13, 17, 21
ii. 2, 6, 10, 14, 18, 22
iii. 3, 7, 11, 15, 19, 23
iv. 4, 8, 12, 16, 20, 24
v. 6, 10, 14, 18, 22

(c) *Population Proportion.* Population proportion of polyester-made jerseys, proportion of *all* tickets in box with 1s, **remains same** / **changes** because box is same as before and so proportion $= \frac{1+1+0+\cdots+1}{24} \approx 0.67$.

(d) If a systematic sample is started by a number chosen at random from 1 to 6, there are **four** / **five** / **six** possible systemic samples *of size 6*.

(e) *Systematic versus SRS.* Systematic sampling is **same as** / **different from** a simple random sampling (SRS) because it is possible to choose two tickets next to one another in an SRS but not so in a systematic sample.

(f) *Systematic, Cluster and Stratified.* Systematic sampling is special case of **cluster sampling** / **stratified sampling** because both result in a choosing a subset from all possible clusters in population.

7. *Simple, Stratified, Cluster or Systematic?* Match racing horses examples with sampling technique. Recall "SRS" means "simple random sample".

(a) **Simple** / **Stratified** / **Cluster** / **Systematic**
An SRS is taken from all racing horses.

(b) **Simple** / **Stratified** / **Cluster** / **Systematic**
All racing horses are listed from lightest to heaviest. Sample consists of taking every seventh racing horse from this list.

(c) **Simple** / **Stratified** / **Cluster** / **Systematic**
All racing horses are listed alphabetically, by name. Sample consists of taking every third racing horse from this list.

(d) **Simple** / **Stratified** / **Cluster** / **Systematic**
All racing horses are classified as light, middle or heavy weight horses. Sample consists of taking SRSs from the racing horses in each weight class.

(e) **Simple** / **Stratified** / **Cluster** / **Systematic**
Horses racing occurs in many cities in the U.S.A. Sample consists of taking SRSs from the racing horses in New York, Los Angeles and Chicago.

8. *Convenience Sampling.*
This sampling procedure chooses, in a non-random way, from only easy-to-access part of population. This is a poor sampling technique. Sampling only jerseys at top of a box of jerseys **is** / **is not** a case of convenience sampling.