Section 4.1 Part a: Sampling and Surveys

What are the appropriate methods for collecting data? Why collect data in the first place? Sometimes as researchers we may be required to find answers to specific questions with an acceptable level of uncertainty. For instance we may wish to know "What percent of Americans believe we should have been involved in a war in Afghanistan?" Or perhaps "What proportion of Americans save at least 10% of their income?" To answer these questions we obviously can't question the entire American population since it is not practical or even feasible so we will pose these questions with a **survey** to a **sample** (part of) of the **population** (the whole group we want information about). A survey is one example of what is referred to in statistics as an observational study. When we observe subjects and measure variables of interest while making no attempt to influence the response we are conducting an observational study. But how can we be sure that the sample in an observational study truly represents the entire population accurately?

Sample design is the method for choosing a sample from the population. Sampling refers to studying data from part of a population in order to gain information about the entire population. When choosing a sample from the population we must be sure to eliminate any **bias** in the selections process. This is critical in the design process. We must be sure that the sample is a random representation of the population. The idea of a random sample is fundamental in sample design since it reduces (or hopefully eliminates) any bias in the selection process. If we collect data from the entire population we are conducting a census. The United States conducts a census every ten years. The last one was conducted in 2010.

1. You may hear results from polls or other statistical studies reported in the media with the emphasis that the samples were randomly selected. Discuss the following two questions in your group.

a. Why do you think there is an emphasis on "random selection"?

b. What do you think could be the problem with allowing a good statistician to choose respondents for a survey as opposed to randomly selecting the respondents?

2. In this activity you will compare subjective sampling with **random sampling** and determine which method better represents the entire population. The goal is to learn why randomization is an important part of sample design.

a. Near the end of this investigation you will find a sheet of 100 rectangles. **DO NOT LOOK** at the sheet of rectangles until your teacher tells you to. When given the signal from your teacher you will have 8 seconds to look at the sheet of rectangles and estimate the average area of these rectangles.

Estimated Average Area:

b. Now select five rectangles, that in your judgment, are representative of the rectangles on the page. Write down the number of the rectangle and the corresponding area for each of the five rectangles in the table below. Calculate the average area of these five rectangles and compare this average with your estimate from part *a*.

Rectangle #					
Area					

Calculated Average Area:

c. Next we will randomly select 5 rectangles. But what is really meant by random selection? How can we be sure that the random selection is truly random? Truly random values are surprisingly hard to get. How good are humans at picking random numbers? To illustrate this idea quickly pick a number at random from the numbers shown below.

1 2 3 4

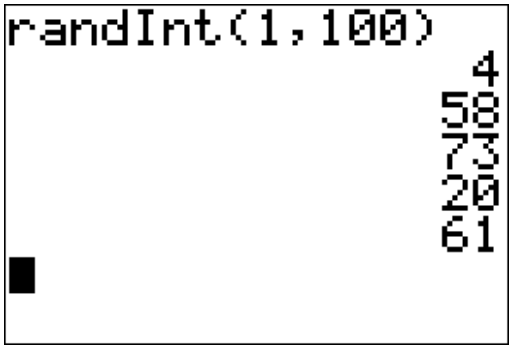
What number did you pick?

Complete the table below showing the percentages of those in your class you picked each corresponding number.

Number	**1**	**2**	**3**	**4**
Class Count				
Percent				

Based on these table results do you think humans can randomly select a number?

(Note that about 75% of the population will "randomly" pick the number three.)



d. Humans' can't effectively select values randomly. So how should we generate random values? Computers and calculators have become popular methods and even though they do a much better job at it than humans, computers and calculators can't generate truly random numbers either. They must use some type of algorithm to produce these numbers. This is why the lottery uses numbered balls in mixing machines. But we will use our graphing calculator or a table of random digits. Use your graphing calculator (or the attached table of random digits) to generate a random integer from 1 to 100 inclusive as shown in the screen shot at the right. Enter the five random numbers in the table below and find the corresponding area for the rectangle with that number. Compute the average area of these five random rectangles.

Rectangle #					
Area					

Calculated Average Area:

f. Using Fathom enter your three averages so we can create three dots plot for the class data of calculated average areas. After everyone has entered all three of their averages estimate the centers of the three dot plots and compare the overall distributions. How do the plots compare to each other?

e. After looking at the two subjective samplings and the random sample, which do you think is doing a better job of determining the average area of the rectangles? Why?

3. The actual mean area for all 100 rectangles is 7.42.

a. Do any of the three plots have a center that is very close to the true average?

b. Do any of the plots have a center that is larger or smaller than the true average?

c. Discuss the concept of bias in sampling and how it relates to the two sampling methods, subjective and random, we just used.

Section 4.1 Part a: Sampling and Surveys

As you have seen in the previous activities, bias was present when you selected the rectangles and calculated the average areas even though you may not have been conscious of it. Notice also that we collected several samples of size five from the class and calculated the average of those five rectangles in the sample. We then found the mean of those averages. This average of the averages (or the mean of the averages), is an important calculation in statistics.

Bias in the selection process is problematic since it favors certain outcomes. The type of bias present in selecting the rectangles is typically referred to as **selection bias**. Other types of bias exist as illustrated in the following examples.

4. You may have noticed that television news stations like to conduct call-in polls of public opinion. The station will announce a question and then asks viewers to call a particular telephone number to respond "Yes" or another number to respond "No". The ABC show *Nightline* once asked viewers whether the United Nations should continue to have its headquarters in the United States. More than 186,000 callers responded and 67% said "No"

a. Do you think that 186,000 is a large enough sample?

b. How much confidence would you have that the 67% who responded "No" represents the greater population of the United States? Describe the type of bias you believe may be present if any?

5. Often you will find volunteers in shopping malls gathering information by conducting surveys. This is typically a fast and inexpensive way to reach a sample of the population.

a. Do you think that people who shop at malls represent a random sample of the population as a whole? Why or why not?

b. Do you think that the person conducting the survey may tend not to choose certain individuals to include in the survey? How might that impact the survey results?

Section 4.1 Part a: Sampling and Surveys

6. Consider the survey question shown below.

a. *Given that the threat of nuclear war is higher now than it has ever been in human history, and the fact that a nuclear war poses a threat to the very existence of the human race, would you favor an all-out nuclear test ban?*

Describe the bias in the wording of this survey question?

What would be a more appropriate way to word this survey question?

b. Describe the type of bias present in the cartoon below?

Activity 4 is an example of **voluntary response bias** or **voluntary response sampling**. Only people who feel strongly will respond to the call in poll. Additionally, only people with phones can call in, so in some sense this is also an example of **selection bias**. Activity 5 illustrates another type of bias called **convenience bias** or **convenience sampling**. It is fast and cheap to sample mall shoppers but people at malls tend to be upper or middle class, or they are teenagers or retirees that may not be representative of the entire population. Moreover, the interviewers may tend to select "safe-looking" individuals from the customers. Activity 6 and the *Wizard of Id* cartoon above are examples of misleading or **poorly worded questions**.

One way to avoid these types of bias is to use chance when selecting a sample from the population as you did when selecting five random rectangles based on a random number generator. This is the idea of **simple random sampling**.

Definition: Simple Random Sample (SRS)

A simple random sample of size n consists of n cases from the population that are chosen in such a way that every set of size n cases has an equal chance of being selected.

A SRS must meet these two criteria.

1. All subjects have an equal chance of being selected ($n = 1$).
2. All groups of subjects have equal chances of being selected. ($n > 1$).

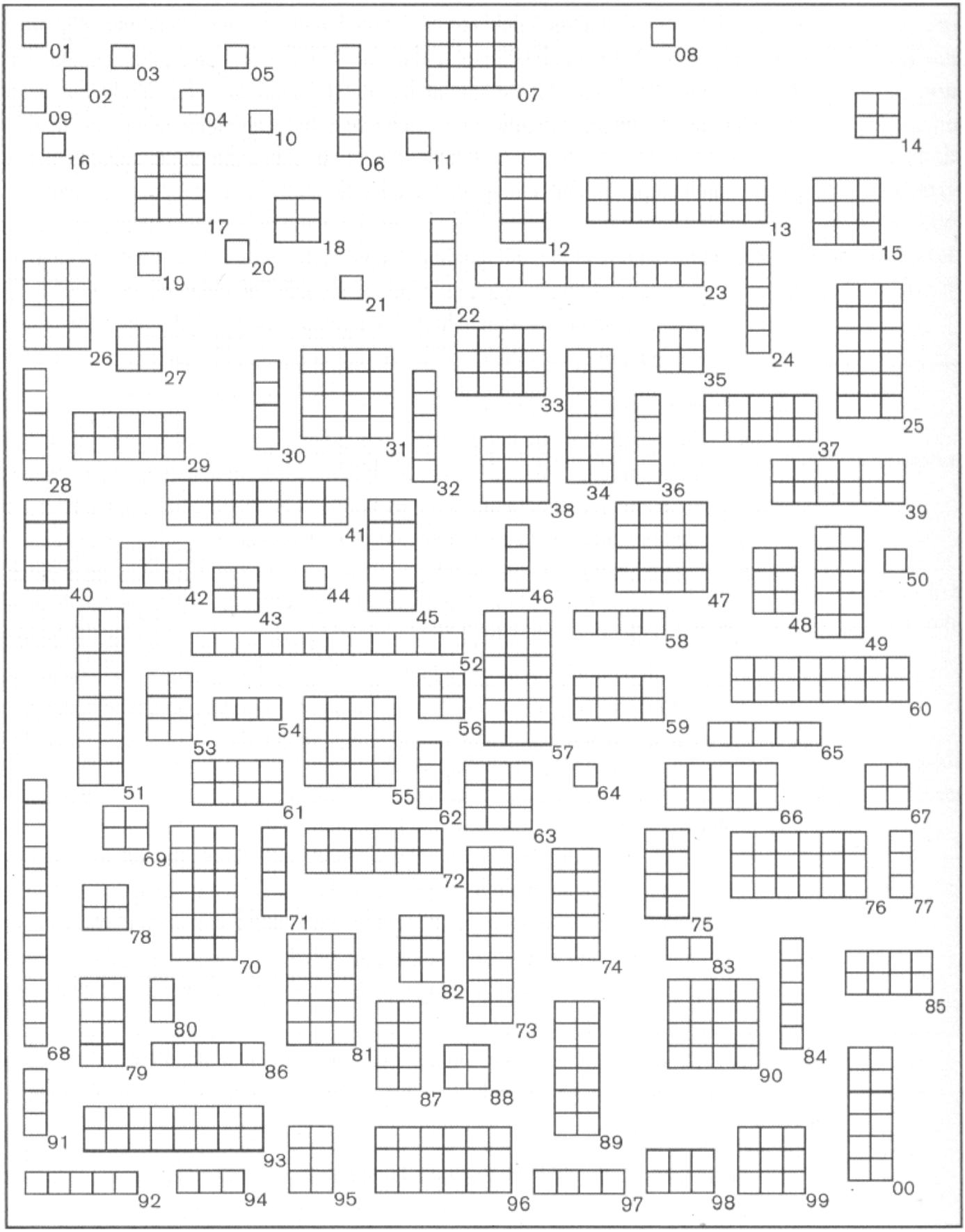

01
03
05
07
08
02
09
04
10
06
11
14
16
17
18
12
13
15
20
19
21
22
23
24
26
27
25
33
35
28
29
30
31
32
34
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
58
53
54
56
57
59
60
61
55
62
63
64
65
66
67
69
72
71
76
77
78
70
74
75
83
82
73
85
68
80
81
84
79
86
87
88
90
91
89
93
92
94
95
96
97
98
99
00

Table D Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244
126	96927	19931	36809	74192	77567	88741	48409	41903
127	43909	99477	25330	64359	40085	16925	85117	36071
128	15689	14227	06565	14374	13352	49367	81982	87209
129	36759	58984	68288	22913	18638	54303	00795	08727
130	69051	64817	87174	09517	84534	06489	87201	97245
131	05007	16632	81194	14873	04197	85576	45195	96565
132	68732	55259	84292	08796	43165	93739	31685	97150
133	45740	41807	65561	33302	07051	93623	18132	09547
134	27816	78416	18329	21337	35213	37741	04312	68508
135	66925	55658	39100	78458	11206	19876	87151	31260
136	08421	44753	77377	28744	75592	08563	79140	92454
137	53645	66812	61421	47836	12609	15373	98481	14592
138	66831	68908	40772	21558	47781	33586	79177	06928
139	55588	99404	70708	41098	43563	56934	48394	51719
140	12975	13258	13048	45144	72321	81940	00360	02428
141	96767	35964	23822	96012	94591	65194	50842	53372
142	72829	50232	97892	63408	77919	44575	24870	04178
143	88565	42628	17797	49376	61762	16953	88604	12724
144	62964	88145	83083	69453	46109	59505	69680	00900
145	19687	12633	57857	95806	09931	02150	43163	58636
146	37609	59057	66967	83401	60705	02384	90597	93600
147	54973	86278	88737	74351	47500	84552	19909	67181
148	00694	05977	19664	65441	20903	62371	22725	53340
149	71546	05233	53946	68743	72460	27601	45403	88692
150	07511	88915	41267	16853	84569	79367	32337	03316

Section 4.1 Part a: Sampling and Surveys

What are the appropriate methods for collecting data? Why collect data in the first place? Sometimes as researchers we may be required to find answers to specific questions with an acceptable level of uncertainty. For instance we may wish to know "What percent of Americans believe we should have been involved in a war in Afghanistan?" Or perhaps "What proportion of Americans save at least 10% of their income?" To answer these questions we obviously can't question the entire American population since it is not practical or even feasible so we will pose these questions with a **survey** to a **sample** (part of) of the **population** (the whole group we want information about). A survey is one example of what is referred to in statistics as an observational study. When we observe subjects and measure variables of interest while making no attempt to influence the response we are conducting an observational study. But how can we be sure that the sample in an observational study truly represents the entire population accurately?

Sample design is the method for choosing a sample from the population. Sampling refers to studying data from part of a population in order to gain information about the entire population. When choosing a sample from the population we must be sure to eliminate any **bias** in the selections process. This is critical in the design process. We must be sure that the sample is a random representation of the population. The idea of a random sample is fundamental in sample design since it reduces (or hopefully eliminates) any bias in the selection process. If we collect data from the entire population we are conducting a census. The United States conducts a census every ten years. The last one was conducted in 2010.

1. You may hear results from polls or other statistical studies reported in the media with the emphasis that the samples were randomly selected. Discuss the following two questions in your group.

a. Why do you think there is an emphasis on "random selection"? To control or eliminate bias

b. What do you think could be the problem with allowing a good statistician to choose respondents for a survey as opposed to randomly selecting the respondents?

They may be biased.

2. In this activity you will compare subjective sampling with **random sampling** and determine which method better represents the entire population. The goal is to learn why randomization is an important part of sample design.

a. Near the end of this investigation you will find a sheet of 100 rectangles. **DO NOT LOOK** at the sheet of rectangles until your teacher tells you to. When given the signal from your teacher you will have 8 seconds to look at the sheet of rectangles and estimate the average area of these rectangles.

Answers vary

Estimated Average Area:

b. Now select five rectangles, that in your judgment, are representative of the rectangles on the page. Write down the number of the rectangle and the corresponding area for each of the five rectangles in the table below. Calculate the average area of these five rectangles and compare this average with your estimate from part *a*.

Rectangle #					
Area	ANSWERS	VARY			

Calculated Average Area: Answers VARY

c. Next we will randomly select 5 rectangles. But what is really meant by random selection? How can we be sure that the random selection is truly random? Truly random values are surprisingly hard to get. How good are humans at picking random numbers? To illustrate this idea quickly pick a number at random from the numbers shown below.

1 2 3 4

What number did you pick? Answers VARY

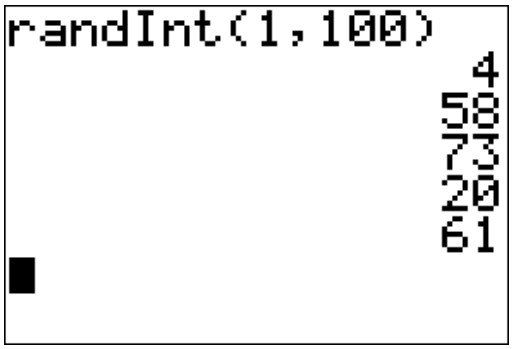
Complete the table below showing the percentages of those in your class you picked each corresponding number.

Number	**1**	**2**	**3**	**4**
Class Count	Answers	VARY		
Percent				

Based on these table results do you think humans can randomly select a number?
(Note that about 75% of the population will "randomly" pick the number three.)

NO!

d. Humans' can't effectively select values randomly. So how should we generate random values? Computers and calculators have become popular methods and even though they do a much better job at it than humans, computers and calculators can't generate truly random numbers either. They must use some type of algorithm to produce these numbers. This is why the lottery uses numbered balls in mixing machines. But we will use our graphing calculator or a table of random digits. Use your graphing calculator (or the attached table of random digits) to generate a random integer from 1 to 100 inclusive as shown in the screen shot at the right. Enter the five random numbers in the table below and find the corresponding area for the rectangle with that number. Compute the average area of these five random rectangles.



Rectangle #					
Area	ANSWERS VARY				

Calculated Average Area:

ANSWERS VARY

f. Using Fathom enter your three averages so we can create three dots plot for the class data of calculated average areas. After everyone has entered all three of their averages estimate the centers of the three dot plots and compare the overall distributions. How do the plots compare to each other?

They all have similar amounts of variability but the first two methods have centers that are a bit larger than the random method.

e. After looking at the two subjective samplings and the random sample, which do you think is doing a better job of determining the average area of the rectangles? Why?

Answers vary.

3. The actual mean area for all 100 rectangles is 7.42.

a. Do any of the three plots have a center that is very close to the true average?

yes, the randomly selected DATA

b. Do any of the plots have a center that is larger or smaller than the true average?

yes, the non-random/subjective DATA

Section 4.1 Part a: Sampling and Surveys

c. Discuss the concept of bias in sampling and how it relates to the two sampling methods, subjective and random, we just used.

We have a tendency to notice and select the larger rectangles, thus our subjective method overestimates the average area while the random method avoids this bias and thus results in a more accurate estimate of the true mean area.

Section 4.1 Part a: Sampling and Surveys

As you have seen in the previous activities, bias was present when you selected the rectangles and calculated the average areas even though you may not have been conscious of it. Notice also that we collected several samples of size five from the class and calculated the average of those five rectangles in the sample. We then found the mean of those averages. This average of the averages (or the mean of the averages), is an important calculation in statistics.

Bias in the selection process is problematic since it favors certain outcomes. The type of bias present in selecting the rectangles is typically referred to as **selection bias**. Other types of bias exist as illustrated in the following examples.

4. You may have noticed that television news stations like to conduct call-in polls of public opinion. The station will announce a question and then asks viewers to call a particular telephone number to respond "Yes" or another number to respond "No". The ABC show *Nightline* once asked viewers whether the United Nations should continue to have its headquarters in the United States. More than 186,000 callers responded and 67% said "No"

a. Do you think that 186,000 is a large enough sample?

IT IS

b. How much confidence would you have that the 67% who responded "No" represents the greater population of the United States? Describe the type of bias you believe may be present if any?

Not much, the people responding must have strong opinions to even bother responding.

5. Often you will find volunteers in shopping malls gathering information by conducting surveys. This is typically a fast and inexpensive way to reach a sample of the population.

a. Do you think that people who shop at malls represent a random sample of the population as a whole? Why or why not?

No, people shopping have money to spend. This may not be true of the entire population.

b. Do you think that the person conducting the survey may tend not to choose certain individuals to include in the survey? How might that impact the survey results?

Yes, those that don't frighten or intimidate them. Like choosing rectangles, it may exclude certain parts of the population.

Section 4.1 Part a: Sampling and Surveys

6. Consider the survey question shown below.

a. *Given that the threat of nuclear war is higher now than it has ever been in human history, and the fact that a nuclear war poses a threat to the very existence of the human race, would you favor an all-out nuclear test ban?*

Describe the bias in the wording of this survey question?

IT SETS UP THE SITUATION AS DIRE AND PREDISPOSES The respondent TO ANSWER IN THE AFFIRMATIVE.

What would be a more appropriate way to word this survey question?

Do you favor an all-out nuclear test ban?

b. Describe the type of bias present in the cartoon below?

None of the response choices are negative.

Section 4.1 Part a: Sampling and Surveys

Activity 4 is an example of **voluntary response bias** or **voluntary response sampling**. Only people who feel strongly will respond to the call in poll. Additionally, only people with phones can call in, so in some sense this is also an example of **selection bias**. Activity 5 illustrates another type of bias called **convenience bias** or **convenience sampling**. It is fast and cheap to sample mall shoppers but people at malls tend to be upper or middle class, or they are teenagers or retirees that may not be representative of the entire population. Moreover, the interviewers may tend to select "safe-looking" individuals from the customers. Activity 6 and the *Wizard of Id* cartoon above are examples of misleading or **poorly worded questions**.

One way to avoid these types of bias is to use chance when selecting a sample from the population as you did when selecting five random rectangles based on a random number generator. This is the idea of **simple random sampling**.

Definition: Simple Random Sample (SRS)

A simple random sample of size *n* consists of *n* cases from the population that are chosen in such a way that every set of size *n* cases has an equal chance of being selected.

A SRS must meet these two criteria.

3. All subjects have an equal chance of being selected ($n = 1$).
4. All groups of subjects have equal chances of being selected. ($n > 1$).

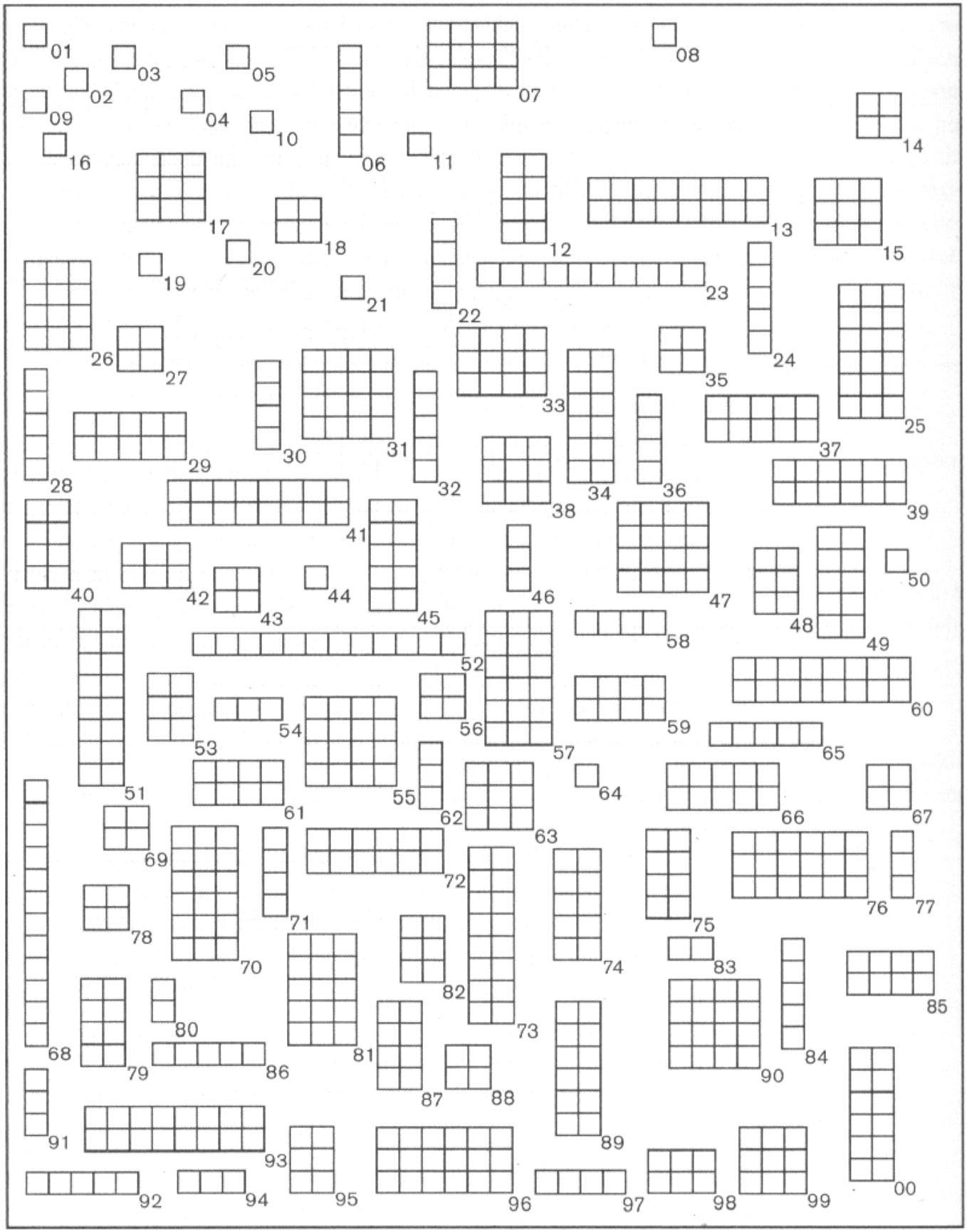

01
03
05
07
08
02
09
04
10
06
11
14
16
17
18
12
13
15
20
19
23
21
22
24
26
27
33
35
25
29
30
31
32
37
28
34
36
38
39
41
40
42
43
44
45
46
47
48
49
50
58
52
53
54
56
57
59
60
65
51
55
61
62
63
64
66
67
69
72
71
75
76
77
78
70
74
83
82
73
85
68
80
81
84
79
86
87
88
90
91
89
93
92
94
95
96
97
98
99
00

Table D Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244
126	96927	19931	36809	74192	77567	88741	48409	41903
127	43909	99477	25330	64359	40085	16925	85117	36071
128	15689	14227	06565	14374	13352	49367	81982	87209
129	36759	58984	68288	22913	18638	54303	00795	08727
130	69051	64817	87174	09517	84534	06489	87201	97245
131	05007	16632	81194	14873	04197	85576	45195	96565
132	68732	55259	84292	08796	43165	93739	31685	97150
133	45740	41807	65561	33302	07051	93623	18132	09547
134	27816	78416	18329	21337	35213	37741	04312	68508
135	66925	55658	39100	78458	11206	19876	87151	31260
136	08421	44753	77377	28744	75592	08563	79140	92454
137	53645	66812	61421	47836	12609	15373	98481	14592
138	66831	68908	40772	21558	47781	33586	79177	06928
139	55588	99404	70708	41098	43563	56934	48394	51719
140	12975	13258	13048	45144	72321	81940	00360	02428
141	96767	35964	23822	96012	94591	65194	50842	53372
142	72829	50232	97892	63408	77919	44575	24870	04178
143	88565	42628	17797	49376	61762	16953	88604	12724
144	62964	88145	83083	69453	46109	59505	69680	00900
145	19687	12633	57857	95806	09931	02150	43163	58636
146	37609	59057	66967	83401	60705	02384	90597	93600
147	54973	86278	88737	74351	47500	84552	19909	67181
148	00694	05977	19664	65441	20903	62371	22725	53340
149	71546	05233	53946	68743	72460	27601	45403	88692
150	07511	88915	41267	16853	84569	79367	32337	03316