CHAPTER 7 –DATA FOR DECISIONS

The ***population*** in a statistical study is the entire group of individuals about which we want information.

A ***sample*** is a part of the population from which we actually collect information used to draw conclusions about the whole.

Sampling refers to the process of choosing a sample from the population.

A ***convenience sample*** is a sample of individuals who are selected because they are members of a population who are the most convenient to reach.

A ***voluntary response sample*** consists of people who choose themselves by responding to a general appeal. usually gets extreme opinions

The design of a statistical study is ***biased*** if it systematically favors certain outcomes.

A ***simple random sample (SRS)*** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be in the sample actually selected.

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate.

A polling company surveys 200 people outside a county courthouse concerning tighter restrictions on smoking in public buildings in the county.

(a) What is the population? people who use public buildings in that county

(b) What is the sample? 200 people outside courthouse

(c) Is this a SRS? NO (Convenience sample)

(d) What type of bias may be present?
May be on a smoke break
sort of Small sample
Employee vs visitor

To determine the proportion of voters who favor a certain candidate for governor, the campaign staff phones 2500 residents of the state chosen from the state property tax rolls.

(a) What is the population? voters in that state

(b) What is the sample? 2500 property owners

(c) Is this a SRS? NO b/c we only asked property owners and there are voters who do not own property.

(d) What type of bias may be present?
Financial - property owners only
How did they pick the 2500? Could cause bias
Only contacted people via phone - some people don't have phone.

In order to determine the proportion of voters in a small town who favor a candidate for mayor, the campaign staff takes out an ad in the paper asking voters to call in their preference for mayor.

(a) What is the population? Voters in the small town

(b) What is the sample? people who call in

(c) Is this a SRS? NO Voluntary response

(d) What type of bias may be present? only one response avenue (phone)
Only people w/ strong opinions will take time to call in
could have non-resident callers, callers who vote multiple times.
Only people who take paper would see the call.

A polling company conducted a survey of voters to obtain data for a political campaign. They selected 3500 voters randomly from the 16,800 names on the voter registration lists of the county. Each voter contacted can reply by mail, phone or internet.

(a) What is the population? 16,800 Voters in that county

(b) What is the sample? 3500 voters randomly chosen

(c) Is this a SRS? simple random sample Yes b/c every voter had equal chance of being chosen

(d) What type of bias may be present?
No obvious biases, but it depends on how selected voters were contacted

How can we choose a SRS? Use a table of random digits.

A table of random digits is a list of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with the following two properties:

- Each entry in the table is equally likely to be any of the 10 digits 0 through 9
- The entries in the table are independent of each other.

To use the table of random digits to generate a SRS, do the following:

1. Give each member of the population a numerical label of the same length.
2. Read from the table strings of digits of the same length as the labels.
3. Skip values that are not in the range.
4. Ignore spaces and don't carry numbers over from the previous line

↑ pretrence for this class

A group of people want to order random items off a menu with 40 different items. How will the items be labeled? Which items will be chosen if the group wants 4 items? Start at line 102. Will the items chosen change if the items are labeled differently? ← Yes

01, 02, 03, ... , 39, 40 ← other possibilities as well

line # only

Pick 24, 40, 36, 35

LINE	RANDOM DIGITS					
101	08705	42934	79257	89138	21506	26797
→ 102	00755	39242	50772	44036	54518	56865
103	35486	59500	20060	89769	54870	75586
104	87788	73717	19287	69954	45917	80026
105	51052	25648	02523	84300	83093	39852
106	88988	12439	73741	30492	19280	41255

If a business has 4,490 employees, how could they be assigned labels?

0001, 0002, 0003, ... 4490

or 0000, 0001, 0002, - - - - - 4489

or 1000, 1001, 1002, - - - - - - - - 5489 or lots of other ways

Choose four employees from the table of random digits below starting at line 122. Labels 0001, 0002, 0003, . . . 4490

Pick: 3227, 0298, 1585, 3152

120	2	7	0	3	1	0	3	8	9	7	1	6	7	3	8	3	1	4	5	3	0	7	5	4	5
121	3	5	1	8	6	0	3	9	5	1	6	8	2	0	8	7	3	4	6	0	7	5	3	1	4
→ 122	3	2	2	7	4	6	7	4	9	2	2	1	6	2	5	3	0	2	9	8	1	5	8	5	5
123	9	7	8	8	6	3	1	4	8	0	9	6	6	1	1	3	9	0	3	1	3	1	5	2	5
124	4	0	1	3	5	2	2	6	0	9	7	1	8	7	5	7	3	4	3	3	1	2	8	3	8
125	8	7	5	3	8	7	4	6	3	3	4	0	0	0	2	7	4	4	7	9	8	8	1	1	3
126	5	1	3	4	9	3	9	8	8	5	2	9	9	9	5	3	7	8	5	8	1	8	3	1	3
127	7	0	7	1	8	4	0	9	4	1	2	8	7	0	6	7	5	5	1	0	0	5	8	3	2
128	9	0	2	3	4	7	4	9	8	3	3	7	7	3	2	3	7	0	2	4	4	1	7	1	8
129	0	0	9	6	2	9	3	9	5	8	4	6	9	8	5	9	4	9	8	9	3	0	2	2	1
130	2	7	2	1	9	6	7	2	6	0	8	2	7	4	0	1	8	9	4	6	2	9	1	7	0

← ignore b/c not long enough for label

An ***experiment*** deliberately imposes a treatment on individuals in order to observe their responses. The purpose of an experiment is to study whether the treatment *causes* a change in the response.

Variables are said to be ***confounded*** when their effects on the outcome cannot be distinguished from each other.

How can we deal with confounded variables? Use a ***control group*** that does not receive the treatment.

A ***controlled experiment*** is an experiment that has a control group. An ***uncontrolled experiment*** is an experiment that lacks a control group.

The ***placebo effect*** is the effect of a dummy treatment on the response of the subjects.

In a ***double-blind*** experiment neither the experimental subjects nor the observers know which treatment the subjects are given.

An observed effect so large that it would rarely (less than 5% of the time) occur by chance is called ***statistically significant***.

An ***observational study*** is a passive study of a variable of interest. The study *does not attempt to influence* the responses and is meant to describe a group or situation.

A ***prospective study*** is an observational study that records slowly developing effects of a group of subjects over a long period of time.
A ***retrospective study*** is an observational study that uses interviews or records to collect information about past behaviors in two or more groups.

A group of 200 students is identified. Half took Latin in high school and half did not. The students are compared to see if the students who took Latin received higher SAT verbal scores. Was this a study or experiment?

b/c we did NOT impose a treatment

A group of people with high blood pressure were given a magnesium supplement or a placebo for 4 weeks. The two groups are compared to see if the magnesium supplement lowered blood pressure. Was this a study or an experiment? b/c we did impose a treatment

Controlled experiment b/c we had a control group

A group of adults were asked if they drank fluoridated water when they were young and asked how many dental fillings or crowns they have. Was this a study or an experiment?

retrospective study

If you wanted to know how smoking affects a baby's birth weight, would you do an experiment or a study? b/c unethical to impose treatment of forcing moms to smoke

Statistical inference refers to methods used for drawing conclusions about an entire population on the basis of data from a sample. A ***confidence interval*** is one type of inference method.

Statistical inference will only be valid if the data is from a random sample or a randomized comparative experiment.

A ***parameter*** is a fixed (and usually unknown) number that describes a population.

A ***statistic*** is a number that describes a sample.

If the parameter for the proportion of successes is called p, then the corresponding statistic for the proportion of successes is called $\hat{p}$.

An opinion poll uses random digit dialing to dial 300 (sample) phone numbers in an area code. There were 72 cell numbers dialed which is not surprising as 25% of the numbers in that area code are cell phone numbers.

$\hat{p}$ The number 72/300 is statistic b/c comes from sample

p The number 25% is parameter b/c comes from population

An online poll asks if you had been to a movie at a theater in the last week. This poll had 450 responses, and 200 of them were positive. However, the information from the local theaters indicates that 20% of the residents go to a theater each week. What are p and $\hat{p}$?

20% $\frac{200}{450}$

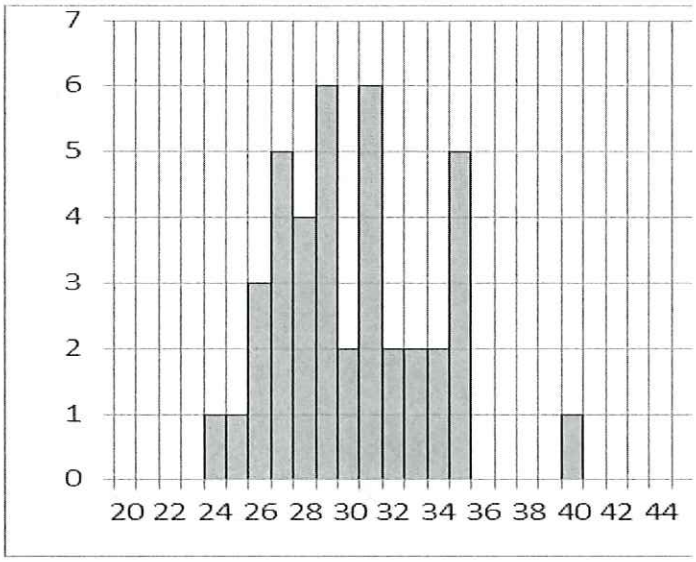
The ***sampling distribution*** of a statistic is the distribution of values taken on by the statistic in all possible samples of the same size from the same population.

At a certain school 60% of the students are football fans. A SRS of 50 students are chosen and asked if they like football. The result was 31 students liked football (31/50 = 0.62 = 62%). A different SRS of 50 students was chosen and that survey found 27 students liked football (27/50 = 54%).

The next few results were 33 students (66%), 29 students (58%), and 34 students (64%).

The results of the 40 surveys done are shown.

The mean result of the 40 surveys was 30.6 and a standard deviation of 3.5.



When a SRS of size n is chosen from a large population that has a proportion of success p.

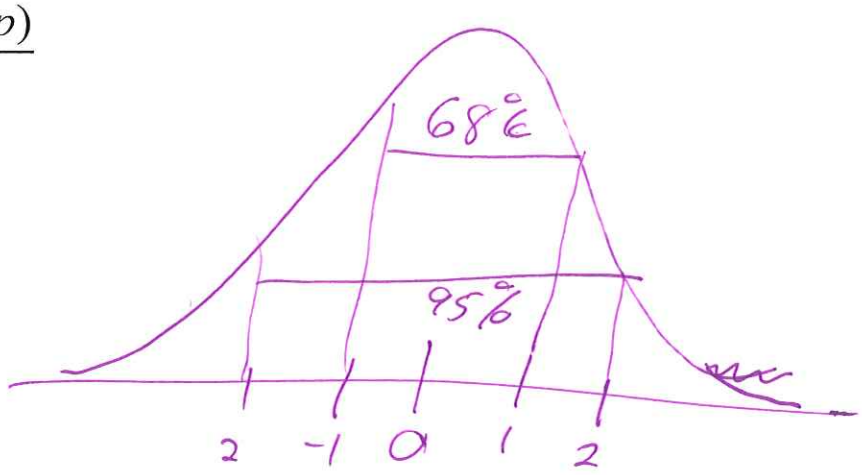
- The shape of the sampling distribution of $\hat{p}$ will be approximately normal if n is 30 or more. # people in sample
- The mean of the sampling distribution of $\hat{p}$ is p.

use mean of sample to approximate mean of population.

- The spread in the sampling distribution of $\hat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$

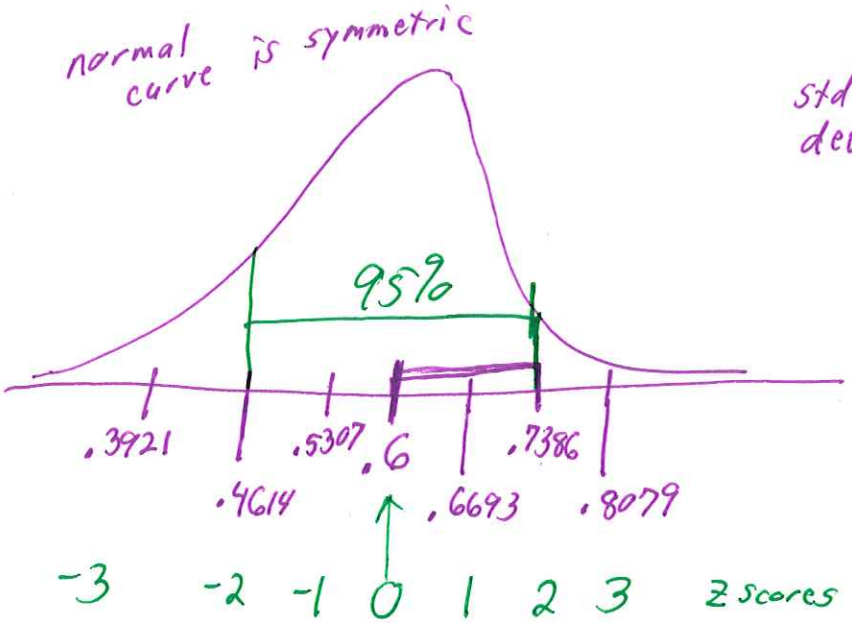
Standard deviation

The margin of error in a sampling distribution at 95% confidence is $2\sqrt{\frac{p(1-p)}{n}}$



normal curve w/ mean of .6 & stddev of $\sqrt{\frac{p(1-p)}{n}}$

With our population p = 0.6 and SRS of size $n = 50$, what do we expect the sampling distribution to look like? What is the margin of error (at 95% confidence level)? Between what two values should 95% of our results fall? Between what two values should 99.7% of our results fall?



$$\text{std dev} = \sqrt{\frac{p(1-p)}{n}} \qquad p = .6,\ n = 50$$

$$= \sqrt{\frac{.6(1-.6)}{50}} = \sqrt{\frac{(.6)(.4)}{50}}$$

$$\approx 0.0693$$

Margin of error at 95% confidence level is

$$2 \text{ std dev} \approx 2(0.0693) \approx .1386 = 13.86\%$$

95% of results fall between

.4614 and .7386 $\pm$ 2 stddev

46.14% and 73.86%

99.7% of results fall between $\pm$ 3 std dev

39.21% and 80.79%

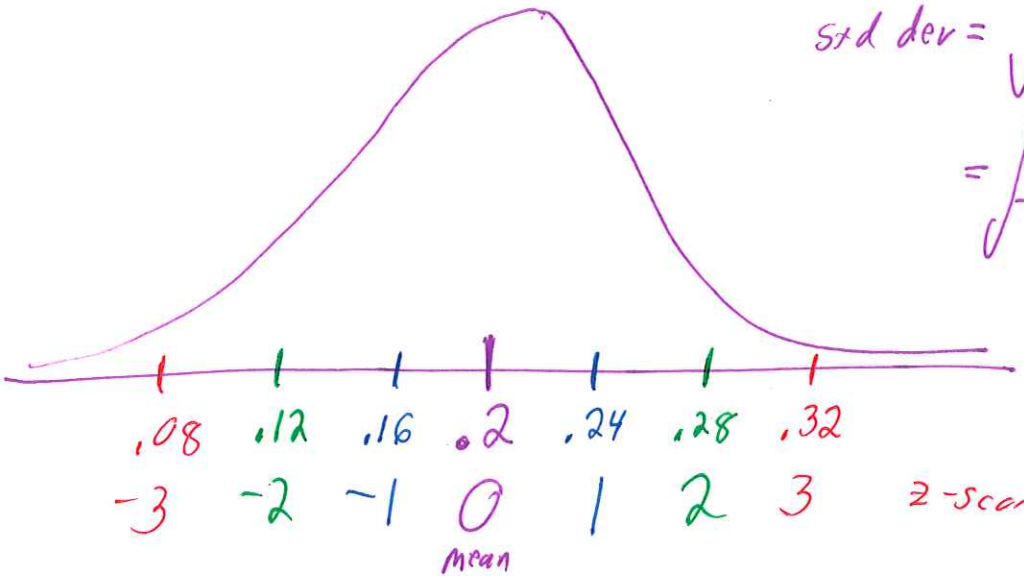
$p = .2$

A population has a 20% chance that a adult reads a newspaper. A SRS of 100 adults were asked if they read the newspaper. If this experiment was repeated many times, what would the sampling distribution look like? normal curve

How large would the sample need to be to have a margin of error of 3% at a 95% confidence level?

$$\text{std dev} = \sqrt{\frac{p(1-p)}{n}}$$

$$= \sqrt{\frac{.2(.8)}{100}} = 0.04$$



Margin of error at 95% confidence interval?

2 std dev = 2(0.04) = .08 or 8%

want MoE at 95%

$2 \text{ std dev} = 0.03$

$$2\sqrt{\frac{p(1-p)}{n}} = 0.03$$

$$2\sqrt{\frac{.2(.8)}{n}} = 0.03$$

Divide by 2

$$\sqrt{\frac{.2(.8)}{n}} = 0.015$$

Square both sides

$$\frac{.2(.8)}{n} = .000225$$

multiply by n

$$.2(.8) = .000225n$$

divide by .000225

$$\frac{.2(.8)}{.000225} = n$$

$$711.11 \approx n$$

$$712 = n$$

SAMPLE EXAM QUESTIONS FROM CHAPTER 7

1. An opinion poll selected 500 email addresses at random from a list of 2500 student emails. Of these, 111 are freshmen. This is not surprising because 25% of the students are freshmen. Which number or numbers below are parameters?

(A) no values are parameters (B) 111/500 (C) 500/2500 (D) 25%

2. You wish to survey the students at your college to determine their feelings about the quality of services in the student center. Which of the following sampling designs is best for avoiding bias?

(A) Air an announcement on the campus radio station asking all listeners to phone in their opinions. Voluntary response bias

(B) Survey every tenth student who enters the student center. Convenience sample

(C) Place an ad in the student newspaper asking all readers to mail in their opinions. Voluntary response bias

(D) Obtain a list of student names from the registrar and randomly select 250 names to contact.

3. In order to determine the mean weight of bags of chips filled by its packing machines, a company inspects 50 bags per day and weighs them. In this example, the population is:

(A) the 50 bags inspected each day. Sample

(B) all potato chips produced by the company.

(C) all bags of chips produced by the company.

(D) the weight of the 50 bags inspected.

4. To determine the proportion of students at a university who favor the construction of a parking garage, a student senate member surveys students as they leave the student union. This type of sample is a:

(A) convenience sample.
(B) simple random sample.
(C) voluntary response sample.

5. Consider the following situation: A group of 300 students is randomly selected at a local high school and required to fill out yearly questionnaires on family income. Students' performances on standardized tests are then followed throughout their high school years to determine if socio-economic status affects test scores. This describes

(A) a comparative experiment.
(B) a controlled experiment.
(C) a prospective study.
(D) a retrospective study.

6. A dummy medication (such as a salt tablet) will often help a patient who trusts the doctor who administers the medicine. This is called:

(A) confidentiality.
(B) double-blindness.
(C) confounding variables.
(D) the placebo effect.

7. A flashlight manufacturer sets aside a production line for the assembly of 2000 flashlights to fill a special order. Ninety of these flashlights are selected at random from the production line to be tested, and 15 are found to be defective. The population is:

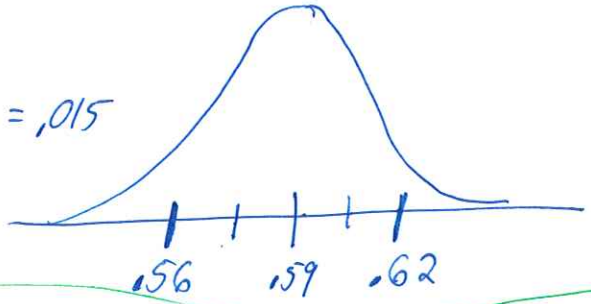
Sample

(A) the 15 defective flashlights.
(B) the 90 flashlights tested.
(C) the 2000 flashlights produced for this order.
(D) all flashlights produced by the manufacturer.
(E) None of these

8. If the 95% confidence interval is determined to be from 56% to 62%, then the standard deviation is

(A) 6%
(B) 3%
(C) 1.5%
(D) 12%
(E) None of these

$\frac{.62-.56}{4} = .015$



extra info: mean $= \frac{.56+.62}{2} = .59$

Margin of Error $= 2$ std dev $= 2(.015) = .03 = 3\%$

9. You must choose a simple random sample of 25 of the 314 members of your fly fishing club. How would you label the population in order to use a table of random digits to make your selection?

(A) 1, 2, 3, …, 24, 25
(B) 1, 2, 3, …, 313, 314 ← need same length for labels
(C) 000, 001, 002, …, 313, 314 ← has 315 people in it
(D) 001, 002, 003, …, 313, 314
(E) None of these

$\hat{p} = .35$ use $\hat{p}$ to approximate p

10. In opinion poll of 1000 adults, 35% said they thought the weather was good. What is the margin of error in this poll at a 95% confidence level?

(A) 0.0151
(B) 0.0302
(C) 0.00023
(D) 0.00046
(E) none of these

$\text{Std dev} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.35(1-.35)}{1000}} \approx 0.015083103\ldots$

Margin of Error at 95% $= 2$ std dev
$\approx 2(0.015083\ldots)$
≈ 0.030166

11. Raffle tickets are issued with the numbers 1001 to 3003 on them. Three prizes will be given out using the table of random digits below. Starting on line 112, which raffle tickets are winners?

4 digit #

110	8	2	0	2	8	4	3	1	1	7	2	6	5	6	8	9	4	1	4	3	8	8	4	9	9	8	0	6	8	3
111	4	6	2	7	9	9	4	5	5	1	0	2	2	3	8	2	0	7	7	0	6	6	2	4	0	0	7	7	0	9
→ 112	4	0	8	8	6	0	9	4	8	1	8	2	1	2	6	7	1	8	8	0	6	8	2	2	9	0	3	0	4	9
113	1	5	0	1	4	4	1	3	8	2	5	6	0	9	4	3	8	3	9	7	5	0	4	2	1	6	1	2	0	2
114	0	9	3	3	3	3	8	4	9	5	5	2	5	6	1	5	1	6	1	8	2	6	3	8	7	7	7	2	1	4
115	3	4	7	7	6	1	4	7	6	2	0	8	6	4	3	2	2	8	4	9	0	5	8	8	9	4	6	8	6	9
116	1	5	4	6	9	3	5	3	2	1	1	1	1	7	9	8	1	8	9	5	3	4	0	7	4	4	2	6	5	0
117	3	1	6	6	7	1	2	8	3	2	7	5	7	1	9	5	5	3	6	6	4	3	8	2	9	3	2	4	5	5
118	8	9	3	3	5	3	1	5	8	2	3	9	5	3	1	3	5	8	8	1	3	7	7	3	6	2	2	5	4	2

1st ticket is 1267 2nd ticket is 1880 3rd ticket is 1501

12. A pollster wants to have a margin of error of 2% in his poll at a 95% confidence level. If $p = 0.4$, how large does the sample need to be?

Looking for n

2 std dev = 0.02

$2\sqrt{\frac{.4(1-.4)}{n}} = 0.02$

divide by 2

$\sqrt{\frac{.4(.6)}{n}} = 0.01$

square both sides

$\frac{.4(.6)}{n} = (0.01)^2$

multiply by n

$.4(.6) = (0.01)^2 n$

divide by $(0.01)^2$

$\frac{.4(.6)}{(.01)^2} = n$

$2400 = n$