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.

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 inside Kyle Field on a football Saturday concerning seeking opinions about college athletics from people residing in America.

(a) What is the population? people residing in America

(b) What is the sample? 200 people inside Kyle Field

(c) Is this a SRS? NO Convenience sample

(d) What type of bias may be present?
people inside Kyle Field have already dedicated money and time to be there so have expressed in college Football.
People in sample are from a specific region of US. Not entire US
Biased toward TAMU Fans

To determine the proportion of voters who favor a certain candidate for school board, the school staff phones 120 residents of the school district chosen randomly from the list of parents of students in the district.

(a) What is the population? Voters in school district

(b) What is the sample? 120 residents of school district who were called

(c) Is this a SRS? NO because a SRS would give an equal chance to everyone in the population to be chosen. This sample was taken only from parents of students in the district.

(d) What type of bias may be present?
Chose only parents to ask. other people can vote too.
Used only phones to contact.

In order to determine the proportion of voters on campus who favor a certain candidate for yell leader, the campaign staff takes out an ad in the paper asking voters to text in their preference for yell leader.

(a) What is the population? eligible voters on campus

(b) What is the sample? people who texted in response

(c) Is this a SRS? NO Voluntary response

(d) What type of bias may be present?
people who don't get the paper wouldn't see the question.
people who can't see to read the paper wouldn't get the question
Some people (professors/staff) could respond but can't vote
People could text multiple times to have their vote counted multiple times.

A polling company conducted a survey of voters to obtain data for a political campaign. They selected 2500 voters randomly from the 76,800 names on the voter registration lists of the district. Each voter contacted can reply by mail, phone or internet.

(a) What is the population? 76,800 voters of the district

(b) What is the sample? 2500 voters who were contacted

(c) Is this a SRS? Yes b/c every person in the population had an equal chance of being chosen

(d) What type of bias may be present?

May get a non response bias if not enough of the contacted people respond.

A table of random digits is one way to choose a SRS.

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
5. For this class, don't carry numbers over from the previous line

A teacher wants to randomly choose students from her class of 45 students. How will the students be labeled to use the table?

01, 02, 03, 04, . . ., 45 11, 12, 13, 14, ... 55

00, 01, 02, 03, . . ., 44 10, 11, 12, 13, . . . 54

Which student numbers will be chosen if the teacher wants 4 students and starts at line 103? Pick 35, 20, 06, 08

Using

Will the students chosen change if the students are labeled differently?

Line #

LINE	RANDOM DIGITS					
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

If a business has 5,490 employees, how could they be assigned labels?

0001, 0002, 0003, . . . 5490
0000, 0001, 0002, . . . 5489
1000, 1001, 1002, . . . 6489

Choose five employees from the table of random digits below starting at line 124.

using

Pick: 4013, 5226, 3433, 1283, 3340

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

Leave off remaining digits

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 adults were asked how many sodas they drank daily when they were young and asked how many dental fillings or crowns they have. Was this a study or an experiment? b/c Not imposing a treatment

retrospective

A group of 200 students is identified. Half took 4 years of math in high school and half did not. The students are compared to see if the students who took 4 years of math received higher SAT math scores. Was this a study or experiment? b/c Not impose a treatment

A group of people with osteoporosis were given a vitamin supplement or a placebo for 4 weeks. The two groups are compared to see if the bone density had changed. Was this a study or an experiment? b/c imposed a treatment

Controlled

If you wanted to know how drug use affects a baby's birth weight, would you do an experiment or a study?

Unethical to force people to do drugs

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 400 phone numbers in an area code. There were 312 cell numbers dialed which is not surprising as 75% of the numbers in that area code are cell phone numbers.

The number 312/400 is a Statistic b/c represents the sample so $\hat{p} = \frac{312}{400}$

The number 75% is a parameter b/c represents the population so $p = 75\% = 0.75$

An online poll asks if you had been to a restaurant in the last week. This poll had 750 responses, and 600 of them were positive. However, the information from the local restaurants indicates that 62% of the residents go to a restaurant each week. What are p and $\hat{p}$?

$p = 62\% = 0.62$ $\hat{p} = \frac{600}{750} = 0.80 = 80\%$

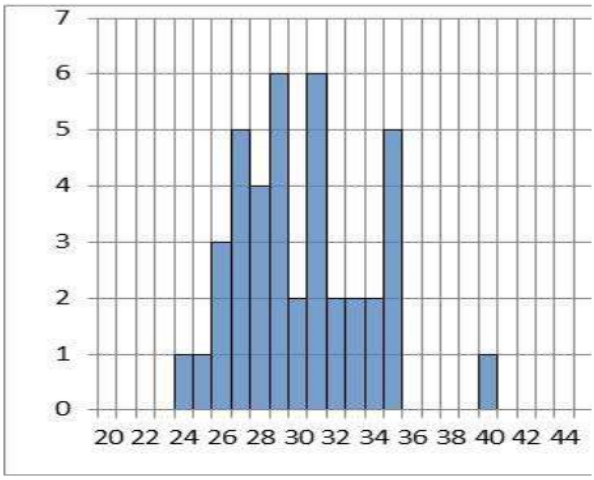
The ***sampling distribution*** of a statistic is the distribution of values taken on by the statistic in all possible samples of the same size from the same population.

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.



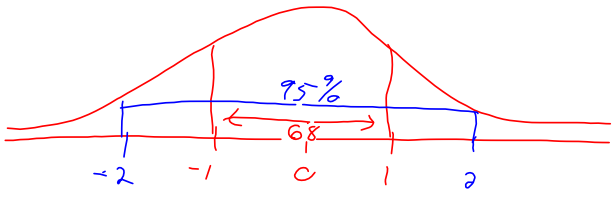
When a SRS of size n (if $n \geq 30$) is chosen from a large population that has a proportion of success p.

- **Shape**: The shape of the sampling distribution of $\hat{p}$ will be approximately normal.
- **Center:** The mean of the sampling distribution of $\hat{p}$ is p.
- **Variability:** The spread in the sampling distribution of $\hat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$.

The margin of error in a sampling distribution at 95% confidence is

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

b/c within 2 std dev comprise 95% of data



A population has a 25% chance that an adult reads a newspaper. A SRS of 400 adults were asked if they read the newspaper. If this experiment was repeated many times, what would the sampling distribution look like?

Normal curve
Centered at $p = 0.25$ with a std. dev of $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.25(1-.25)}{400}} \approx 0.02165$

What is the margin of error (at 95% confidence level)?

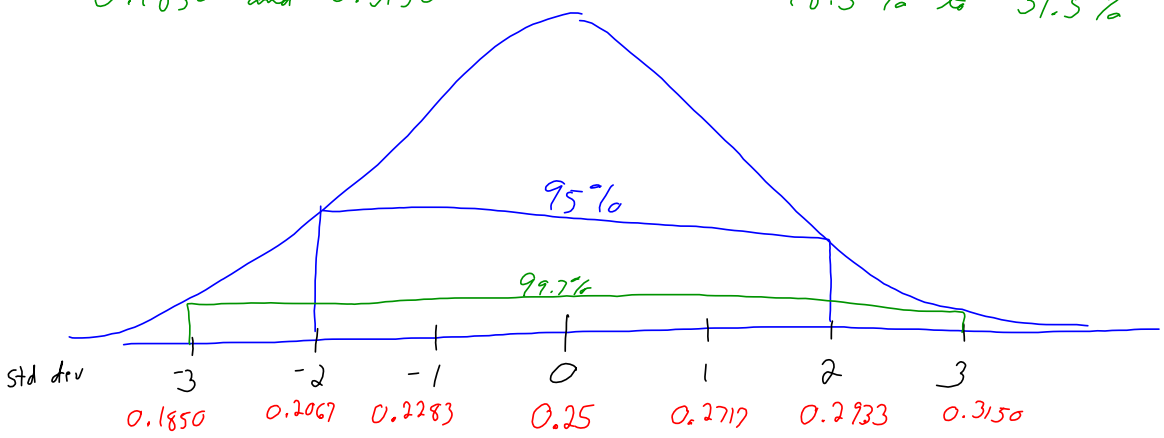
2 stddev $\approx 2(0.02165) \approx 0.0433$

Between what two values should 95% of our results fall?

0.2067 and 0.2933 or 20.67% to 29.33%

Between what two values should 99.7% of our results fall?

0.1850 and 0.3150 or 18.5% to 31.5%



For the previous problem, how large would the sample need to be to have a margin of error of 3% at a 95% confidence level?

MOE at 95% CI $= 2 \cdot \text{Std dev} = 2\sqrt{\frac{p(1-p)}{n}}$

$0.03 = 2\sqrt{\frac{0.25(1-0.25)}{n}}$

Divide both sides by 2

$\frac{0.03}{2} = \sqrt{\frac{0.25(1-0.25)}{n}}$

Square both sides

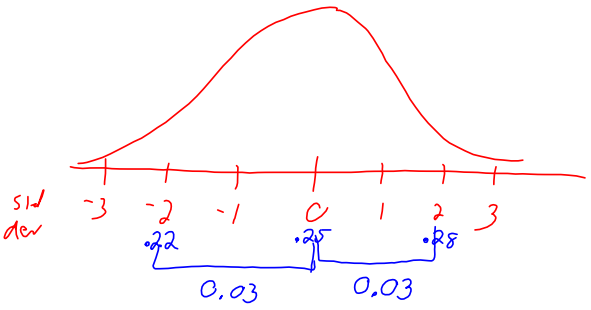
$\left(\frac{0.03}{2}\right)^2 = \frac{0.25(1-0.25)}{n}$

Multiply both sides by n

$\left(\frac{0.03}{2}\right)^2 n = 0.25(1-0.25)$

Divide both sides by the coefficient of n

$n = \frac{0.25(1-0.25)}{\left(\frac{0.03}{2}\right)^2} = \frac{0.1875}{0.000225} = 833.\bar{3}$ so we need at least 834 people



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 statistics representing the proportion of freshmen?

(A) no values are statistics (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) Post an announcement on the campus web page station asking all readers to text in their opinions.

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

(C) Ask the students who sit near you in class about their opinions.

(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 popcorn filled by its packing machines, a company inspects 80 bags per day and weighs them. In this example, the population is:

(A) the 80 bags inspected each day.

(B) the weight of the 80 bags inspected.

(C) all popcorn produced by the company.

(D) all bags of popcorn produced by the company.

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 walk through central campus. 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 200 students is randomly selected at a local high school and required to fill out yearly questionnaires on time spent reading. Students' performances on standardized tests are then followed throughout their high school years to determine if time spent reading affects test scores. This describes

(A) a comparative experiment. Didn't impose a treatment
(B) a controlled experiment.
(C) a prospective study. Follow as they progress
(D) a retrospective study.

6. An inert pill (a pill with no medication inside) 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 3000 flashlights to fill a special order. Eighty of these flashlights are selected at random from the production line to be tested, and 25 are found to be defective. The population is:

(A) the 25 defective flashlights.
(B) the 80 flashlights tested.
(C) the 3000 flashlights produced for this order. Only sampling from this special order
(D) all flashlights produced by the manufacturer.
(E) None of these

$\hat{p} = \frac{25}{80}$

8. If the 95% confidence interval is determined to be from 55% to 63%, then the standard deviation is

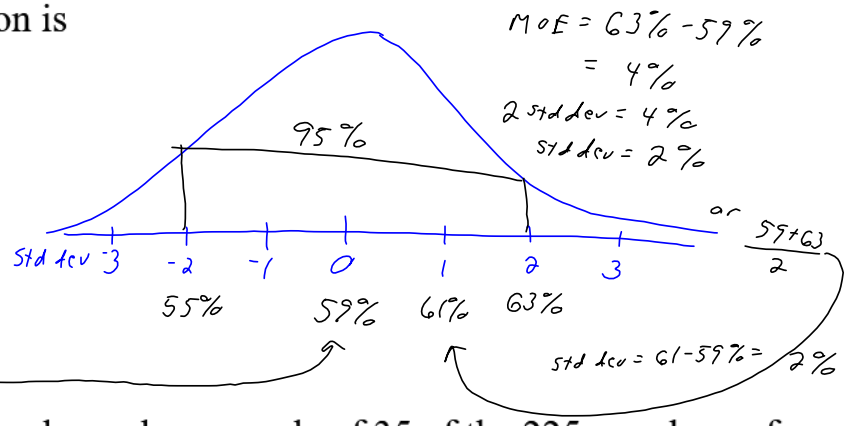
(A) 8%

(B) 4%

(C) 2%

(D) 16%

(E) None of these



9. You must choose a simple random sample of 35 of the 225 members of your bass 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, …, 34, 35 b/c label label entire population

(B) 1, 2, 3, …, 224, 225 b/c labels are not all the same length

(C) 000, 001, 002, …, 224, 225 Contains 226 people

(D) 001, 002, 003, …, 224, 225

(E) None of these

10. In an opinion poll of 600 adults, 20% said they thought the weather was good. What is the margin of error in this poll at a 95% confidence level?

(A) 0.0003

(B) 0.0163

(C) 0.0327

(D) 0.0653

(E) none of these

$\hat{n}$ $\hat{p}$

$$\text{std dev} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.2(1-.2)}{600}} \approx 0.01633$$

$$\text{MOE for 95\% CI} = 2 \text{ std dev} = 2(0.01633) \approx 0.03266$$

11. Raffle tickets are issued with the numbers 1001 to 3003 on them. Four prizes will be given out using the table of random digits below. Starting on line 114, which raffle tickets are winners?

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

← ignore b/c not long enough to be a 4digit

1st ticket is 1618 2nd ticket is 2638

3rd ticket is 2849 4th ticket is 1546

12. A pollster wants to have a margin of error of 4% in his poll at a 99.7% confidence level. If $p = 0.45$, how large does the sample need to be?

MOE at 99.7% CI $= 3$ Std dev $= 3\sqrt{\frac{p(1-p)}{n}}$

$$0.04 = 3\sqrt{\frac{0.45(1-0.45)}{n}}$$

Divide both sides by 3

$$\frac{0.04}{3} = \sqrt{\frac{0.45(1-0.45)}{n}}$$

Square both sides

$$\left(\frac{0.04}{3}\right)^2 = \frac{0.45(1-.45)}{n}$$

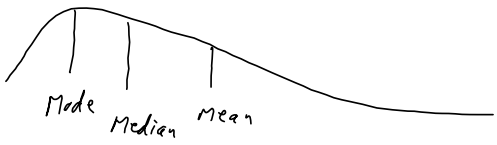
Multiply both sides by n

$$\left(\frac{0.04}{3}\right)^2 n = 0.45(0.55)$$

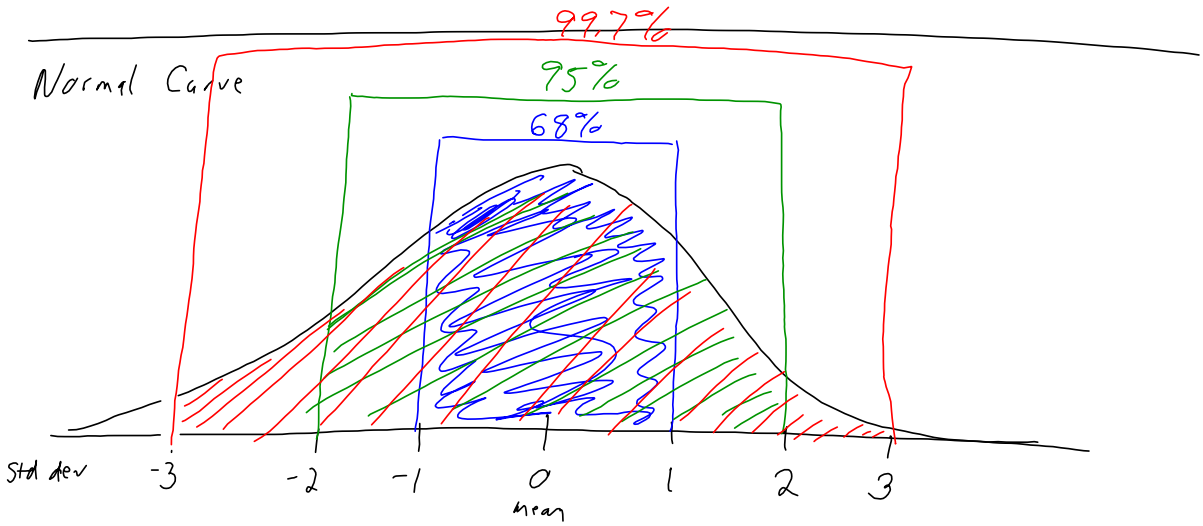
Divide both sides by the coefficient of n

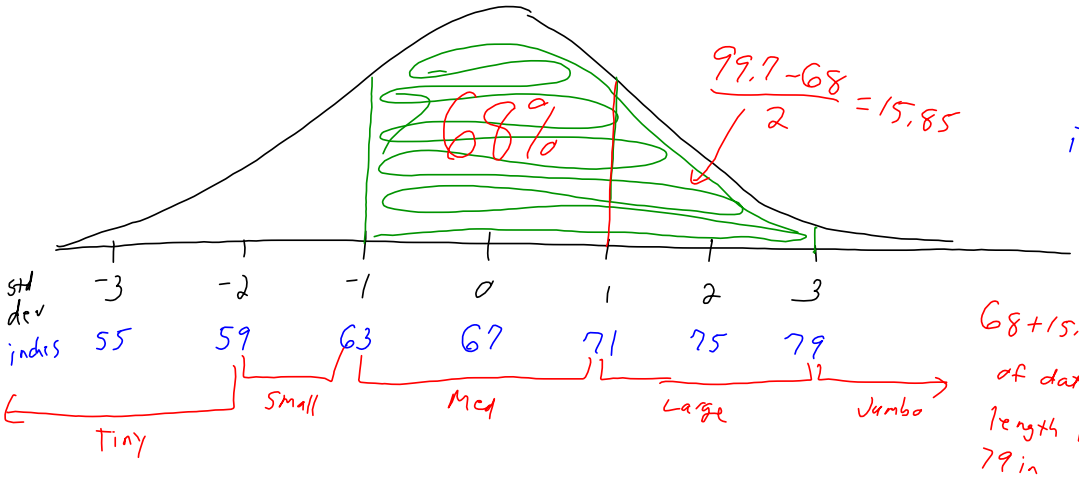
$$n = \frac{0.45(0.55)}{\left(\frac{0.04}{3}\right)^2} = \frac{0.2475}{(0.01\overline{3})^2} \approx 1392.1875$$ so we need at least 1393 people

Extra Questions:



Mean is pulled toward the tail





Length of tape in inches if mean = 67 Std dev = 4

68+15.85 = 83.85 % of data has tape length between 63 and 79 in

If you had 2000 rolls of tape, how many were Med or Large?

.8385 (2000) = 1677