RANDOM SAMPLING

1. PUBLIC OPINION POLLS

Gallup Poll, Harris Poll, Newspapers,

Magazines

2. MARKET RESEARCH

Consumer Preferences, Nielsen Ratings, Pew Research Center

3. GOVERNMENT STATISTICS

Census Bureau, Current Population Survey

Bureau of Labor Statistics

4. ACCEPTANCE SAMPLING

Inspections and Quality Control

5. ACCOUNTING DATA

Audits, Taxes-IRS, Airlines, ASCAP

- **Chapter 20: Chance Errors In Sampling**

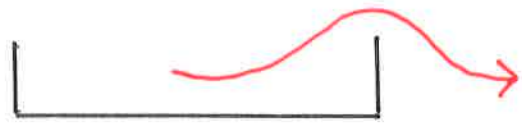
- A simple random sample is like drawing at random, without replacement, from a box of tickets.

- In practice, we use a computer to generate a simple random sample because it's quite difficult to properly randomize paper tickets.

- A simple random sample is affected by chance variability.

Expected Value and Standard Error

Suppose you randomly draw n times with replacement from a box and consider the sum of the draws:



Draw n times

EV for the sum of the draws $=$ Box AV x n

The sum of the draws $=$ EV for the sum $+$ chance error

SE for the sum of the draws $=$ Box SD x $\sqrt{n}$

= the likely size of the chance error.

EV for sum = EV for sum ± SE for sum

Suppose you randomly draw n times with replacement from a box and consider the average of the draws:

$$\text{AV of draws} = \frac{\text{sum of draws}}{n}$$

EV for the average of the draws $=$ Box AV

The average of the draws $=$ Box AV $+$ chance error

$$\text{SE for the average of the draws} = \frac{\text{Box SD x } \sqrt{n}}{n}$$

Suppose you randomly draw n times with replacement from a $0-1$ box and consider the percentage of 1s drawn.

EV for the % of 1s drawn $=$ % of 1s in the box $\qquad \dfrac{\text{EV for sum}}{n} \times 100\%$

$$= \frac{\text{Box AV} \times n}{n} \times 100\%$$

The % of 1s drawn $=$ % of 1s in the box $+$ chance error

$$\text{SE for the \% of 1s drawn} = \frac{\text{Box SD} \times \sqrt{n}}{n} \times 100\% = \text{the likely size of the chance error.}$$

EV for % 1s = EV for % 1s ± SE for % 1s

Comments:

The likely size of the chance error is the standard error or SE; it measures the spread in the probability histogram. The SE applies to chance variability; the SD applies to a list of numbers.

If you are drawing without replacement, then the standard errors must be multiplied by the correction factor. When the number of tickets in the box is much larger than the number of draws, the correction factor is nearly one and may be omitted.

For either the sum of the draws, the average of the draws, or for the percentage of 1s drawn, if the number of draws is large enough and the histogram is placed in standard units, the histogram follows the normal curve. This is the normal approximation theorem.

Examples:

Draw 25 with replacement from the box [0 , 1 , 1 , 1 , 1].

$Box\ AV = \frac{4}{5}$, $Box\ SD = \sqrt{\frac{1}{5} \cdot \frac{4}{5}} = \frac{2}{5}$

EV for sum = $Box\ AV \times n = \frac{4}{5} \times 25 = 20$

SE for sum = $Box\ SD \times \sqrt{n} = \frac{2}{5} \times \sqrt{25} = 2$

sum of draws = EV for sum $\pm$ SE for sum

$= 20 \pm 2$

Draw 25 with replacement from the box [0 , 1 , 1 , 1 , 1].

$Box\ AV = \frac{4}{5}$, $Box\ SD = \sqrt{\frac{1}{5} \cdot \frac{4}{5}} = \frac{2}{5}$

EV for average = $Box\ AV = \frac{4}{5}$

SE for average = $\dfrac{Box\ SD \times \sqrt{n}}{n}$

$= \dfrac{\frac{2}{5} \times \sqrt{25}}{25} = \dfrac{2}{25} = .08$

average of Draws = EV for AV $\pm$ SE for AV

$= \frac{4}{5} \pm .08$

Draw 25 with replacement from the box [0 , 1 , 1 , 1 , 1].

$Box\ AV = \frac{4}{5}$, $Box\ SD = \sqrt{\frac{1}{5} \cdot \frac{4}{5}} = \frac{2}{5}$

EV for % 1s drawn = Box % 1s = 80%

SE for % 1s drawn = $\frac{\text{Box SD} \times \sqrt{n}}{n} \times 100\%$

$= \frac{\frac{2}{5} \times \sqrt{25}}{25} \times 100\% = 8\%$

% 1s drawn = % 1s in box ± SE for % 1s

= 80% ± 8%

Draw 100 with replacement from the box [0 , 1 , 1 , 1 , 1].

$Box\ AV = \frac{4}{5}$, $Box\ SD = \sqrt{\frac{1}{5} \cdot \frac{4}{5}} \approx \frac{2}{5}$

EV for sum = $Box\ AV \times n = \frac{4}{5} \times 100 = 80$

SE for sum = $Box\ SD \times \sqrt{n} = \frac{2}{5} \times \sqrt{100} = 4$

EV for % 1s = % 1s in Box = 80%

SE for % 1s = $\frac{Box\ SD \times \sqrt{n}}{n} \times 100\%$

$= \frac{\frac{2}{5} \times \sqrt{100}}{100} \times 100\% = 4\%$

% 1s = 80% ± 4%

Draw 400 with replacement from the box [0 , 1 , 1 , 1 , 1].

$Box\ AV = \frac{4}{5}$, $Box\ SD = \sqrt{\frac{1}{5} \cdot \frac{4}{5}} = \frac{2}{5}$

EV for sum = $Box\ AV \times n = \frac{4}{5} \times 400 = 320$

SE for sum = $Box\ SD \times \sqrt{400} = \frac{2}{5} \times 20 = 8$

SE for AV = $\frac{Box\ SD \times \sqrt{n}}{n} = \frac{\frac{2}{5} \times 20}{400} = \frac{2}{100}$

EV for % 1s = % 1s in box = 80%

SE for % 1s = $\frac{Box\ SD \times \sqrt{n}}{n} \times 100\% = \frac{2}{100} \times 100\%$

$= 2\%$

% 1s drawn = 80% ± 2%

As the number of draws increases, what happens to the SEs?

Standard Error and Sample Size

$$SE\ \ for\ \%\ 1s\, drawn = \frac{SE\ \ for\ \ sum}{size\, of\ \ sample} \times 100\%$$

Multiplying the size of the sample by some factor divides the SE for the percentage not by the whole factor but by its square root.

(see previous examples)

- **The SE for the sum of the draws goes up like the square root of the sample size.**

- **The SE for the percentage of 1s drawn goes down like the square root of the sample size.**

- **Example:** If I take a simple random sample of 100 people and get SE% = 5%, how many people do I need to sample to make SE% = 1%?

Multiply the sample size by 4
& you halve the SE.

Multiply the sample size by 9
& you divide the S.E. by 3.

Multiply the sample size by 16
& you divide the SE by 4

Multiply the sample size by 25
& you divide the SE by 5.

$\boxed{1, 0, 0, 0, 0}$ → Draw 400 & consider the ______.

- **Example:** In a population of 100,000 people, we know that 20,000 are minorities. I take a simple random sample of 400 people from this population.

a) *How many* minorities do I expect to get in my sample? Give-or-take how many?

Box AV × 400 $= \frac{1}{5} \times 400 = 80$

b) *What percentage* of minorities do I expect to get in my sample? Give-or-take what percentage?

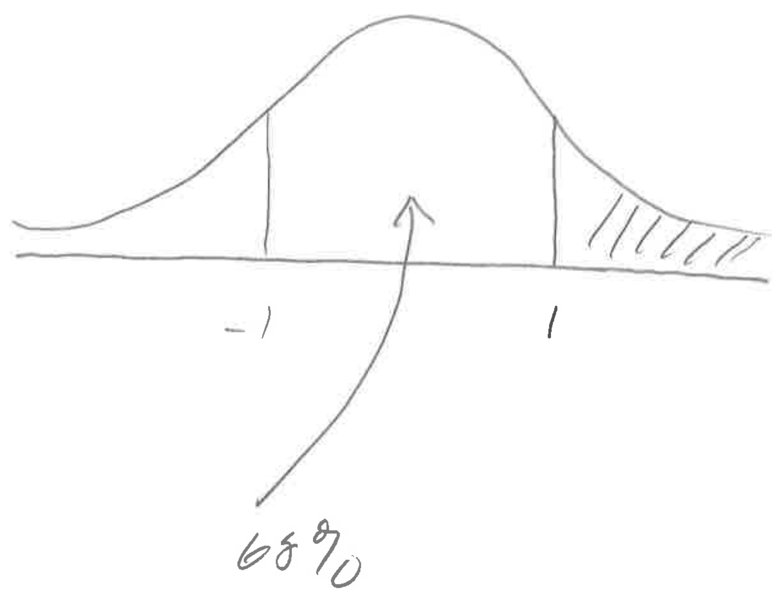
$20\% \pm$ SE for % is $= 20\% \pm \frac{\text{Box SD} \times \sqrt{400}}{400} \times 100\%$

c) Find the chance of getting more than 22% minorities in the sample.

$= \underline{\underline{20\% \pm 2\%}}$

Normal approximation!

$\frac{22-20}{2} = 1\%$



$A(1) = 68\%$

$\boxed{16\%}$

- **Example:** In a certain town, 25% of college students own an automobile. If we take a simple random sample of 400 of these students, what is the chance that less than 20% of the sampled students own an automobile?

$\boxed{1, 0, 0, 0} \rightarrow$ Draw 400 & consider the % 1s drawn.

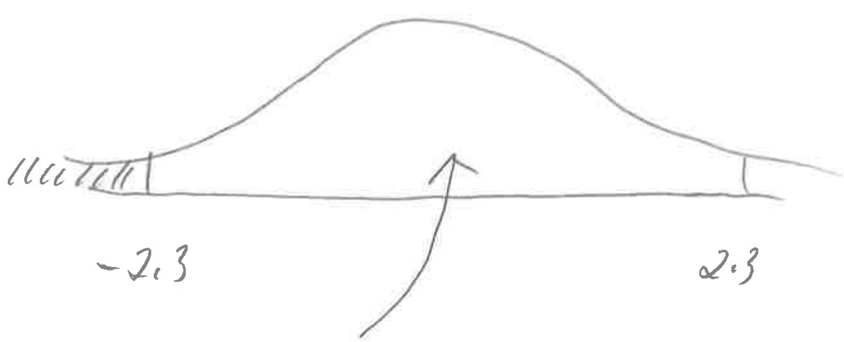
EV for % 1s drawn = 25%

SE for % 1s drawn $= \dfrac{\text{Box SD} \times \sqrt{400}}{400} \times 100\%$

$= \dfrac{\sqrt{\frac{1}{4} \cdot \frac{3}{4}} \times \sqrt{400}}{400} \times 100\% \approx 2.165\%$

Normal approximation:

$\dfrac{20 - 25}{2.165} = -2.3$



$A(2.3) \approx 97.86\%$

About 1.07%

$\approx 1\%$

Correction Factor

$$\begin{matrix} SE \text{ when drawing} \\ WITHOUT \text{ replacement} \end{matrix} = \begin{matrix} \text{correction} \\ \text{factor} \end{matrix} \times \begin{matrix} SE \text{ when drawing} \\ WITH \text{ replacement} \end{matrix}$$

$$\text{correction factor} = \sqrt{\frac{\text{number of tickets in box} - \text{number of draws}}{\text{number of tickets in box} - \text{one}}}$$

Example:

Suppose the number of draws is 2500.

Number of tickets In the box	Correction factor (to five decimals)
5,000	0.70718
10,000	0.86607
100,000	0.98743
500,000	0.99750
1,000,000	0.99875
12,500,000	0.99990

With 15,000 college students
and a sample size of 400,
the correction factor is $\sqrt{\frac{15000 - 400}{15000 - 1}} = .9866$

- **Sample Size and Accuracy**

- For simple random samples, the accuracy is determined by the *sample size itself*, not the sample size relative to the population size.

- E.g. If we want the same accuracy in Salt Lake City and Logan, we should sample the same number of people in Salt Lake City and Logan.

- The rule is true if the sample size is *small* compared to the population size – sampling 20,000 people in Logan would be a bit more accurate than sampling 20,000 in Salt Lake City.

- The rule assumes the thing being estimated is not too different in the two places.

$$\text{SE for \% 1s drawn} = \frac{\text{Box SD} \times \sqrt{n}}{n} \times 100\%$$

does not depend on the population size.

(Assuming the correction factor is ≈ 1)

- **Example :** I'm interested in estimating the percentage of people who watch Channel 5 News. In each case, will the Logan or Salt Lake City sample be more accurate?

a) A simple random sample of 500 Logan residents and a simple random sample of 500 Salt Lake City residents.

same

b) A simple random sample of 500 Logan residents and a simple random sample of 750 Salt Lake City residents.

SLC

c) A simple random sample of 5% of all Logan residents and a simple random sample of 5% of all Salt Lake City residents.

SLC

Chance Errors in Sampling

1. According to the Census, a certain town has a population of 100,000 people age 18 and over. Of them, 60% are married, 10% have incomes over $90,000 a year and 45% have college degrees. As part of a pre-election survey, a simple random sample of 1600 people will be drawn from this population. Find the chance that between 40% and 50% of the people in the sample have a college degree.

$\boxed{45\% \text{ 1s}, \ 55\% \text{ 0s}}$ → Draw 1600 & consider the % 1s drawn.

EV for % 1s = 45%

$$\text{SE for \% 1s} = \frac{\text{Box SD} \times \sqrt{1600}}{1600} \times 100\%$$

$$= \frac{\sqrt{(.45)(.55)} \times 40}{1600} \times 100\% = 1.24\%$$

$\boxed{\text{Normal } z}$

$$\frac{40-45}{1.24} \approx -4, \quad \frac{50-45}{1.24} \approx +4$$

More than 99%

2. The presidential campaign is in full swing and the focus is on the Southwest. Pollsters are trying to predict the results. There are about 1.2 million eligible voters in New Mexico, and about 12.5 million in the state of Texas. Suppose one polling agency takes a simple random sample of 2500 voters in New Mexico in order to estimate the percentage of voters in that state who are Democratic. Another polling organization does the same in Texas. Which poll is likely to be more accurate?

Their accuracy is the same.
Accuracy depends on sample size,
not population size.