THE CHANCE ERROR IN A SAMPLE PERCENTAGE

• A certain town has a population of 100,000 people age 18 and over. 20% of these people are well-educated, that is, have college degrees. A simple random sample of 1,600 people will be drawn from this population.

- • Simple random sampling means drawing at random *without* replacement.
  - • At each draw, all the people not already selected have an equal chance of being chosen.
- • The percentage of well-educated people in the sample will be around _____%, give or take _____% or so.
- • The chance that between 19% and 21% of the people in the sample are well-educated is about _____ .

How do you go about filling in the blanks?

• Step 1. Realize that

$$\text{Percentage of well-educated people in the sample} = \frac{\text{Number of well-educated people in the sample}}{\text{size of the sample}} \times 100\%$$

- • Any statement about the chance variability in the *number* of well-educated people in the sample can be converted to a similar statement about the chance variability in the *percentage* of such people, by dividing by the _____ and multiplying by _____ .

population size 100,000, 20% well-educated; sample size 1,600

• Step 2. Set up a box model. The number of well-educated people in the sample is like the sum of _____ draws made at random _____ replacement from the box

- • We want the process of selecting people from the population to be like drawing tickets from the box. So there should be:
  - • _____ ticket in the box for each person in the population, and
  - • _____ draw for each person in the sample.
- • So there should be _____ tickets in the box, and _____ draws.
- • The tickets for the well-educated people should be marked ☐, and the other tickets as ☐.
  - • We want the number of well-educated people in the sample to be like the sum of the draws from the box. This can be achieved by classifying and counting. To illustrate:

The sample people	Tom ☺	Dick ☺	Harry ☹	Number who are well-educated =
Their tickets	☐	☐	☐	Sum of the draws =

• Since 20% of the population is well-educated, there should be _____ ☐'s in the box, and _____ ☐'s.

- Step 3. Realize that for drawing from the box

20,000 $\boxed{1}$'s 80,000 $\boxed{0}$'s

there isn't much difference between drawing 1600 times *without* replacement and drawing 1600 times *with* replacement.

  - When drawing with replacement, the chance of getting a $\boxed{1}$ is _____ in _____, or 1 in 5, for each draw.
  - When drawing without replacement 1600 times, the chance of getting a $\boxed{1}$ changes from draw to draw, but stays close to 1 in 5, because the percentage of $\boxed{1}$'s in the box can't change much during the course of the draws.
    - The number of draws is _____ compared to the numbers of $\boxed{1}$'s and $\boxed{0}$'s in the box.
  - So, the number of well-educated people in the sample is almost like the sum of 1600 draws made at random *with* replacement from the box.

- Step 4. Find the expected value and SE for the sum of 1600 draws (made with replacement) from the box.
  - The expected value for the sum of the draws from the box equals (number of draws) $\times$ (average of box); that's

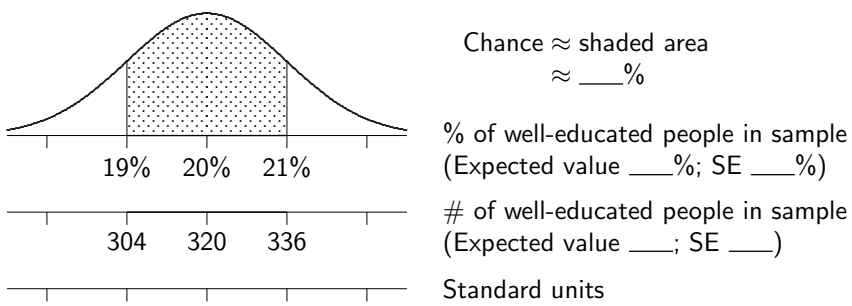
    _____ $\times$ _____ = _____.

  - The SE for the sum of the draws from the box equals $\sqrt{\text{number of draws}} \times$ (SD of box); that's

    _____ $\times$ _____ = _____.

  - The sum of the draws from the box will be around _____, give or take _____ or so.

- Step 5. Interpret the results of Step 4 in terms of the sample.
  - The number of well-educated people in the sample will be about _____, give or take _____ or so.
- Step 6. Convert to percents, relative to the size of the sample.
  - 16 is _____% of 1600.
  - 320 is _____% of 1600.
  - So the percentage of well-educated people in the sample will be around _____% (i.e., the percentage of well-educated people in the population), give or take _____% or so.
- Step 7. Use the normal approximation to estimate chances about the percentage of well-educated people in the sample.
  - The chance that between 19% and 21% of the people in the sample are well-educated is the chance that the number of well-educated people in the sample lies between $320 - 16 = 304$ and $320 + 16 = 336$.



  - The normal approximation can be used, because the sample size (and therefore the number of draws from the box) is reasonably _____.

THE STANDARD ERROR FOR A SAMPLE PERCENTAGE

• Consider taking a simple random sample from a box of tickets, some of which are marked 1, the others 0:

$$\boxed{\text{some } \boxed{1}\text{'s} \qquad \text{some } \boxed{0}\text{'s}} \leftarrow \begin{array}{c}\text{The "population",}\\ \text{in abstract form}\end{array}$$

Suppose the number of draws is large in absolute terms, but small relative to the number of tickets in the box.

• The number of $\boxed{1}$'s in the sample is expected to be about

$$\begin{aligned}&(\text{number of draws}) \times (\text{average of the box})\\ &= (\text{number of draws}) \times (\text{fraction of } \boxed{1}\text{'s in the box}),\end{aligned}$$

give or take

$$\begin{aligned}&\text{SE for the number of } \boxed{1}\text{'s in the sample}\\ &= \sqrt{\text{number of draws}} \times (\text{SD of the box})\end{aligned}$$

or so.

• Now convert to percents relative to the number of draws, by dividing by the number of draws and multiplying by 100%. The percentage of $\boxed{1}$'s in the sample is expected to be about

$$\frac{\text{expected number of } \boxed{1}\text{'s in the sample}}{\text{number of draws}} \times 100\%$$

$$= \boxed{\text{percentage of } \boxed{1}\text{'s in the box,}}$$

give or take

$$\frac{\text{SE for the number of } \boxed{1}\text{'s in the sample}}{\text{number of draws}} \times 100\%$$

$$= \boxed{\frac{\text{SD of the box}}{\sqrt{\text{number of draws}}} \times 100\%}$$

or so.

• To put it another way, as an estimate of the percentage of $\boxed{1}$'s in the box, the percentage of $\boxed{1}$'s in the sample will be off by a chance amount similar in size to the quantity

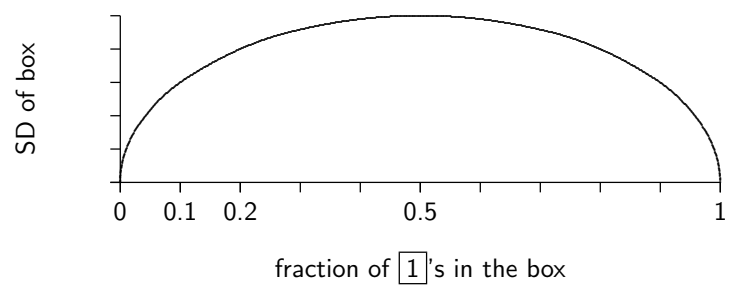
$$\frac{\text{SD of the box}}{\sqrt{\text{number of draws}}} \times 100\%;$$

this quantity is called the standard error (SE) for the sample percentage.

• How does the SE for the sample percentage depend on the box?

  • Only through the variability of the tickets in the box, as measured by their _____, which equals

SD of a box containing $\boxed{0}$'s and $\boxed{1}$'s, as a function of the fraction of $\boxed{1}$'s



• How does the SE depend on the number of draws?

  • The SE is inversely proportional to the square root of the size of the sample. In particular, large samples are more accurate than small ones. Multiplying the size of the sample by some factor divides the SE for the sample percentage by the square root of that factor.

• Reminder: for a SRS from a 0–1 box, the likely size of the chance error in the sample percentage as an estimate of the population percentage equals

• How does this SE depend on the number of tickets in the box?

- _____ ! When estimating percentages, it is the _____ size of the sample which determines accuracy, not the size _____ to the population. This is true when the sample is only a _____ part of the population, which is the usual case.

• For reasonably large samples, the normal approximation can be used, just as for the sum of the draws. In particular, the chance error in the sample percentage is

- smaller than 1 SE with probability about _____ ;
- smaller than 2 SEs with probability about _____ ;
- smaller than 3 SEs with probability about _____ .

• The largest the SD of a 0–1 box could be is _____. So for a SRS of size 2,500, the largest the SE for the sample percentage could be is _____ = 1%.

- For a SRS of size a few _____, the sample percentage is quite likely to be within a few percent of the population percentage — that is, the sample is quite likely to be quite representative of the population.

SAMPLING WITHOUT REPLACEMENT VERSUS SAMPLING WITH REPLACEMENT

• Suppose you're going to draw tickets at random from a 0–1 box:

some $\boxed{1}$'s	some $\boxed{0}$'s

and use the percentage of 1's in the sample as an estimate of the percentage of 1's in the box. Which method would be more accurate, and why?

- Sampling with replacement.
- Sampling without replacement, i.e., taking a SRS.

• Sampling _____ replacement is more accurate, because sampling _____ replacement is inefficient.

- Suppose you sample with replacement, and you put a check mark on each ticket you draw, before replacing it in the box.
- If you draw a ticket with a check mark, then you've seen that ticket before; drawing it the second time doesn't tell you anything new about the tickets in the box.

• Which is smaller, the SE for the percentage of 1's in the sample when drawing with replacement, or the corresponding SE for drawing without replacement?

- The "_____" SE is smaller.

• How much smaller?

- See the next slide!

• The relationship between the two SE's is:

$$\begin{array}{c}\text{SE when drawing}\\ \textit{without}\text{ replacement}\end{array} = \begin{array}{c}\text{correction}\\ \text{factor}\end{array} \times \begin{array}{c}\text{SE when drawing}\\ \textit{with}\text{ replacement}\end{array}$$

where the correction factor is

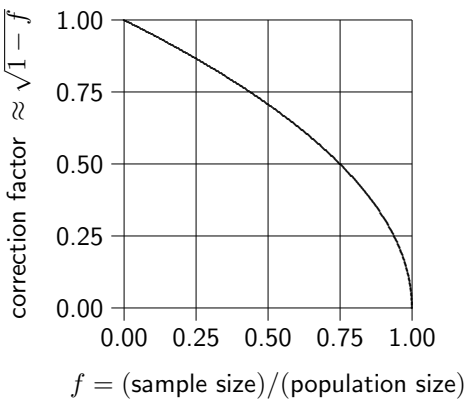
$$\sqrt{\frac{\text{number of tickets in box} - \text{number of draws}}{\text{number of tickets in box} - 1}}\ .$$

• Unless the number of tickets in the box is very small, the correction factor is nearly $\sqrt{1-f}$, where

$$f = \frac{\text{number of draws}}{\text{number of tickets}}$$

is the size of the sample, expressed as a fraction of the size of the population.

• Here's a graph of $\sqrt{1-f}$ versus f:



• Typically f is small, and the correction factor is nearly 1 and can be ignored. (We did ignore it, in our earlier examples.)

• You have hired a polling organization to take a simple random sample from a box of 100,000 tickets, and estimate the percentage of 1's in the box. Unknown to them, the box contains 50% 0's and 50% 1's. How far off should you expect them to be:

- (a) if they draw 2,500 tickets?
- (b) if they draw 25,000 tickets?
- (c) if they draw 100,000 tickets?

• (a) If they drew with replacement, the SE for the percentage of 1's in the sample would be

$$= 1\%$$

For drawing without replacement, the SE is smaller, by a factor of nearly

$$\sqrt{1 - (\text{relative size of sample})} = \sqrt{1 - \qquad} = 0.987 \approx 1.$$

So they'd be off by about 1% or so.

• (b) Increasing the sample size from 2,500 to 25,000, i.e., by by a factor of 10, reduces the SE for sampling with replacement by a factor of _____, from 1% to $1\%/\sqrt{10} = 0.32\%$. The SE for sampling without replacement is smaller still, by a factor of $\sqrt{1 - 25{,}000/100{,}000} = 0.87$. So they'd be off by about $0.87 \times 0.32\% = 0.27\% = 0.27$ of 1%.

• (c) Their estimate would be off by _____%, since they'd have all the tickets from the box in their sample.

SUMMARY

• The sample is only part of the population, so the percentage composition of the sample usually differs a bit from the percentage composition of the whole population. The difference is called the chance error in the sample percentage.

• For probability samples, the likely size of the chance error in the sample percentage is given by the standard error (SE).

• To find the SE for a percentage, first get the SE for the corresponding number, and then convert to percent, by dividing by the size of the sample and multiplying by 100%. (Alternatively, you can use the square root formula given at the top of page 20–6.)

• To figure the SE for the corresponding number, a box model is needed. When the problem involves classifying and counting, or taking percents, there should be only 0's and 1's in the box. Change the box, if necessary.

• When the sample is only a small part of the population, the number of individuals in the population has almost no influence on the accuracy of estimated percentages. It is the absolute size of the sample (that is, the number of individuals in the sample) which matters, not the size of the population.

• The square root law for the SE of the sample percentage is exact when the draws are made with replacement. When the draws are made without replacement, the formula gives a good approximation – provided the number of draws is small by comparison with the number of tickets in the box.

• When drawing without replacement, to get the exact SE it is necessary to multiply the SE for sampling with replacement by the correction factor

$$\sqrt{\frac{\text{number of tickets in box} - \text{number of draws}}{\text{number of tickets in box} - 1}}\ .$$

When the number of tickets in the box is much larger than the number of draws, the correction factor is nearly one.

• The normal approximation can be used to figure chances about a sample percentage; the conversion to standard units is done as

$$z = \frac{\text{sample percentage} - \text{population percentage}}{\text{SE for the sample percentage}}\ .$$

The normal approximation can be trusted if the number of draws from the box is reasonably large. (For sampling with replacement, the number of tickets left in the box must be reasonably large as well; that's typically the case.)