Goal of sample surveys: an accurate snapshot of the population at large. I.e., we want the sample to be *representative*.

How is survey data been collected?

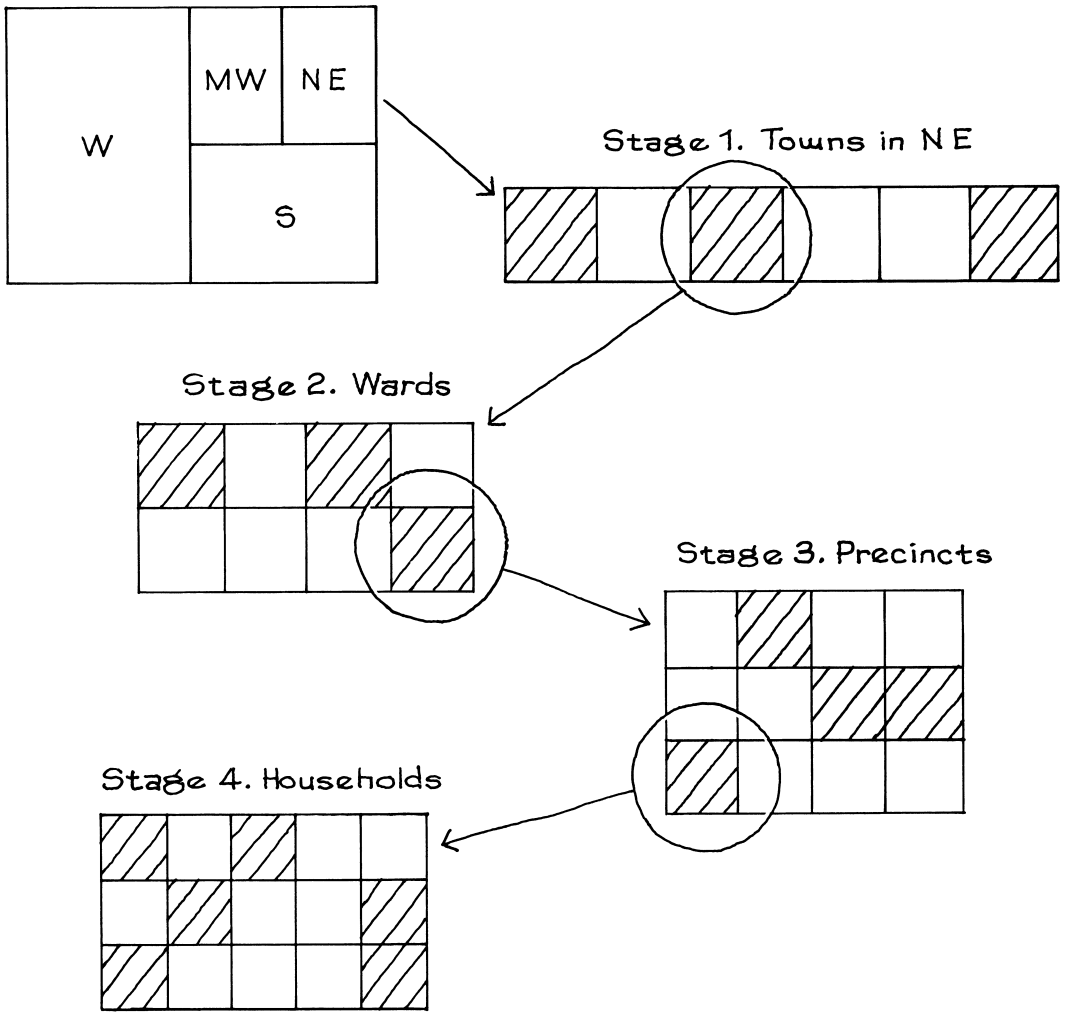
(*) ***Convenience samples*** — think Literary Digest poll of 1936.

(*) ***Quota sampling*** — Pollsters are sent to different regions/cities etc. with specific *quotas* of different types of people. E.g., "*two women age 45-55, three white men, four hispanic women, etc.*" This was Gallup's original method.

(*) ***Simple random samples*** — Like drawing tickets from a box *without* replacement. This is best, theoretically speaking, but impractical, so instead people use...

(*) ***Probabilistic methods*** — Multistage cluster samples, and other more complicated schemes.

Figure 1. Multistage cluster sampling.



Question:

> *A **simple random sample** of 3500 likely voters from California is surveyed. Of these, 2175 say that they support a ballot initiative raising the tax rate on capital gains. What is the likely percentage of all California voters that support this initiative?*

Answer:

The *sample percentage* of voters that support the initiative is ***probably close to*** the *population percentage* that support the initiative.

Conclusion: Approximately

$$\frac{2175}{3500} \approx 62.14\%$$

of California voters support the initiative.

Follow up questions: (i) How ***close***? (ii) How ***probable***?

(0 1)-Boxes

In a (0 1)-box all the tickets are labeled with a 0 or a 1.

- The ***sum*** of all the tickets in a (0 1)-box is equal to the ***number*** of 1s in the box.

- The ***average*** of a (0 1)-box equals the ***proportion***, p, of 1s in the box. Equivalently, the ***percentage*** of 1s in the box is $p \cdot 100\%$.

- The ***SD*** of a (0 1)-box is computed using the shortcut

$$SD_{box} = \sqrt{p \cdot (1-p)},$$

where p is the fraction of 1s in box and $(1-p)$ is the fraction of 0s in box.

Sampling at random with replacement from a (0 1)-box.

If n tickets are drawn at random with replacement from a (0 1)-box...

- The ***expected number*** of 1s in the sample is $p \cdot n$, where p is the *proportion* of 1s in the box (i.e., the average of the box). So...
- The ***expected percentage*** of 1s in the sample is

$$\frac{\text{expected number of 1s}}{n} \cdot 100\% = \frac{p \cdot \cancel{n}}{\cancel{n}} \cdot 100\% = p \cdot 100\%.$$

$\Rightarrow$ The expected percentage of 1s is equal to the percentage of 1s in the box.

- The ***standard error*** for the ***percentage*** of 1s in the sample is

$$SE_{\%} = \frac{SE(\#1s)}{n} \times 100\% = \frac{\sqrt{n} \cdot SD_{box}}{n} \times 100\% = \frac{SD_{box}}{\sqrt{n}} \times 100\%.$$

What changes when we sample at random ***without*** *replacement, i.e., when the sample is a* ***simple random sample****?*

(*) The number of tickets in the box needs to be considered.

- The ***expected percentage*** of $\boxed{1}$s in the sample is still equal to the percentage of $\boxed{1}$s in the box.
- The ***standard error*** for the ***percentage*** of $\boxed{1}$s in a simple random sample is ***smaller*** than when the tickets are drawn with replacement. Specifically

$$SE_{\%} = CF \times \overbrace{\frac{SD_{box}}{\sqrt{n}} \times 100\%}^{SE_{\%} \text{ with replacement}} \quad ,$$

where the ***correction factor*** is $CF = \sqrt{\frac{N-n}{N-1}}$.

When should we include the correction factor?

(*) For simple random samples it is always more accurate to include the correction factor when calculating the SE.

(*) If the sample size n is very small compared to the population size N, then the correction factor has a negligible effect (and can be usually ignored).

Example: If $N = 4000$ and $n = 400$, then

$$CF = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{3600}{3999}} \approx 0.949,$$

so the correction factor will have a small but noticeable effect on the $SE_{\%}$, and should be included in the calculation.

On the other hand, if $N = 400000$ and $n = 400$, then

$$CF = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{399600}{399999}} \approx 0.9995,$$

so the correction factor will have a negligible effect on the $SE_{\%}$, and we don't need to include it in the calculation.

Normal approximation

- When a simple random sample is drawn from a ($\boxed{0}\,\boxed{1}$)-box, the observed percentage of $\boxed{1}$s in the sample differs from the expected percentage of $\boxed{1}$s by some ***chance error***. This chance error is generally no larger than one or two $SE_{\%}$s.
- If the sample size is ***large enough***, then the probability histogram for the *sample percentages of* $\boxed{1}$s is well approximated by the *normal curve* (after converting to *standard units*).
- This means that if the sample size is ***large enough***, then

$$P(|(\text{observed }\%) - (\text{expected }\%)| < Z \cdot SE_{\%}) \approx Table(Z),$$

  where $Table(Z)$ is the area under the normal curve from $-Z$ to Z (in the table at the back of the book).
- How large is ***large enough***? If p is the fraction of $\boxed{1}$s in the population (box) and n is the sample size, then the normal approximation starts to become reasonably accurate once both $np > 5$ and $n(1-p) > 5$ (though bigger is better).

Example.

Suppose that a simple random sample of 400 tickets is drawn from a (0 1)-box of 5000 tickets containing 3000 1s and 2000 0s.

What percentage of 1s are we likely to see in the sample?

- The expected percentage of 1s in the sample is 60% (same as the box percentage).
- The standard error is $SE_{\%} = \sqrt{\frac{4600}{4999}} \times \frac{\sqrt{0.6 \cdot 0.4}}{20} \times 100\% \approx 2.35\%$.
- The sample percentage of 1s is likely to be in the range $60\% \pm 2.35\%$, or between 57.65% and 62.35%. The margin of error here is 1 $SE_{\%}$, and the probability that the sample percentage falls in this range is about 68%.
- If we want a higher probability that the sample percentage falls into the predicted range, we can increase the range. The probability that the sample percentage of 1s falls in the range $60\% \pm 4.7\%$ (55.3% to 64.7%) is about 95%, since the margin of error is now $2SE_{\%}$.

Back to the question:

A simple random sample of 3500 likely voters from California is surveyed. Of these, 2175 say that they support a ballot initiative raising the tax rate on capital gains. What is the likely percentage of all California voters that support this initiative?

(*) **Intuition:** Approximately

$$\frac{2175}{3500} \approx 62.14\%$$

of California voters support the initiative.

Follow up question: How *accurate* is this estimate *likely* to be?

(*) **Know:** (normal approximation)

$$P\left(\text{Pop.}\% - 2SE_{\%} < 62.14\% < \text{Pop.}\% + 2SE_{\%}\right) \approx 95\%,$$

(*) ***Don't know:*** Pop.% or $SE_{\%}$.

From the sample to the box...

The estimate

$$\boxed{P\big(\text{population \%} - 2SE_{\%} < \text{sample\%} < \text{population \%} + 2SE_{\%}\big) \approx 95\%}$$

remains accurate even when we don't know the composition of the population!

The boxed estimate above can also be written as

$$P\left(|\text{ population \%} - \text{sample\%}| < 2SE_{\%}\right) \approx 95\%$$

and this can be rewritten as

$$\boxed{P\big(\text{sample \%} - 2SE_{\%} < \text{population \%} < \text{sample \%} + 2SE_{\%}\big) \approx 95\%}$$

I.e., we can use the sample percentage to find a *likely* range of values for the population percentage!

The interval $((\text{sample \%}) - 2 \cdot SE_{\%}, (\text{sample \%}) + 2 \cdot SE_{\%})$ is called a ***95% confidence interval*** for the population percentage.

Problem:

If we don't know the composition of the box, then we don't know the SD of the box, so we can't find the $SE_{\%}$!

Solution:

Use the ***sample*** proportions of $\boxed{1}$s and $\boxed{0}$s to estimate the proportions in the box and use these estimates to approximate the SD of the box. If the sample size is big enough, this approximation will (almost always) be very good.

Back to the question (again)...

(*) The sample percentage of $\boxed{1}$s(supporters of ballot initiative) is $p_S = 0.6214$, so the standard deviation for the CA box, SD_{CA} can be approximated by...

$$SD_{CA} = \sqrt{p_{CA}(1-p_{CA})} \approx \sqrt{p_S(1-p_S)} = \sqrt{0.6214 \cdot 0.3786} \approx 0.485.$$

Hence

$$SE_{\%} \approx \frac{0.485}{\sqrt{3500}} \times 100\% \approx 0.82\%.$$

Conclusion: There is an approximately 95% chance that the percentage of California voters who support the ballot initiative is within

$$2 \cdot 0.82\% = 1.64\%$$

of the sample percentage 62.14%.

(*) A 95%-confidence interval for the percentage of California voters who support the initiative is

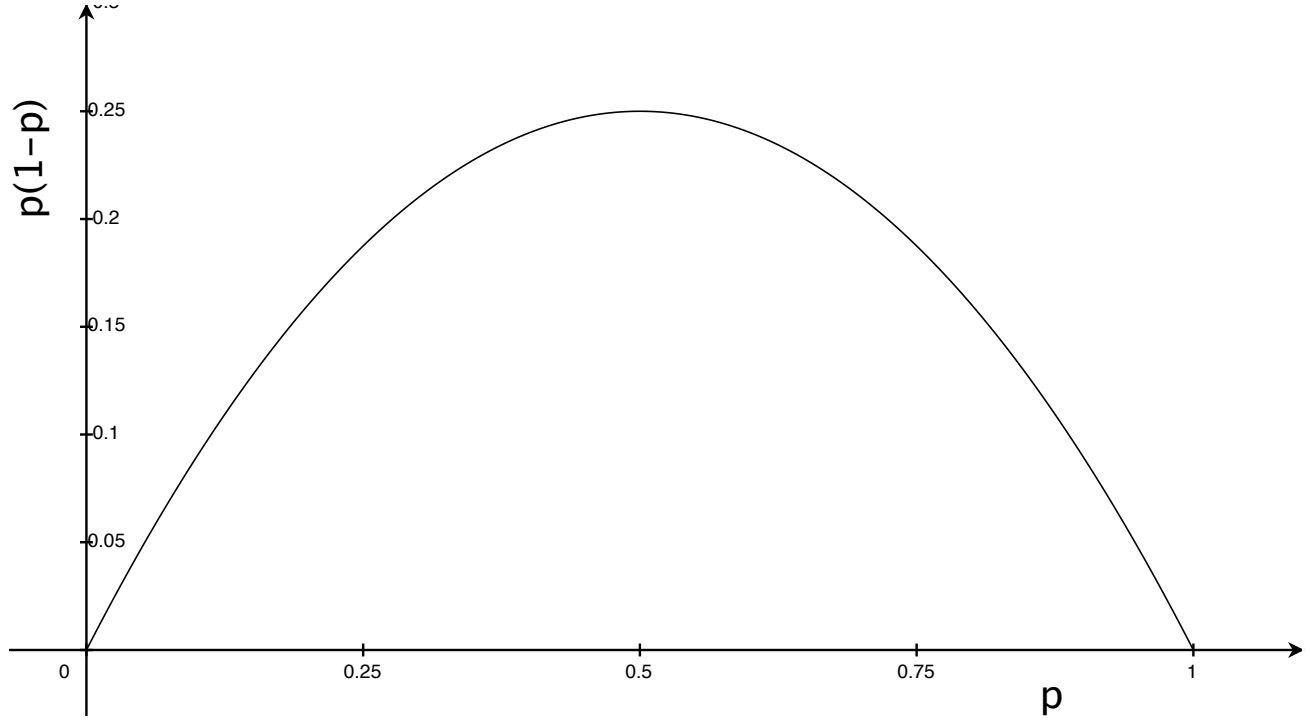
$$\text{Sample-}\% \pm 2SE_{\%} = 62.14\% \pm 1.64\% = (60.5\%, 63.78\%).$$

Comments:

(*) The *margin of error* in this (and other) examples is $2SE_{\%}$.

(*) We obtain a more *conservative* estimate if we use a bigger margin of error than $2SE_{\%}$.

Fact: The SD of a $\boxed{0}$-$\boxed{1}$ box is ***never greater than 1/2***.



Consequence: The $SE_{\%}$ is never greater than $\dfrac{0.5}{\sqrt{n}} \times 100\% = \dfrac{50\%}{\sqrt{n}}$.

What does "95%-confidence" *mean?*

(*) A confidence interval depends on the sample data. Different samples generally produce different sample data — in this case, different sample percentages.

(*) This means that 100 different samples will produce 100 different 95%-confidence intervals — though most of them will be very similar to each other, some perhaps identical.

(*) The percentage of $\boxed{1}$s in the population (box) is unknown but ***fixed***. The intervals we construct vary with the samples.

(*) The term "95%-confidence" means that about 95% of all the intervals we construct using this method will contain the true (but unknown) population percentage.

Observation

When surveying large populations the accuracy of the prediction depends primarily on the sample size, not the relative size of the sample.

What does this mean?

(*) The accuracy of the prediction is given by the margin of error, which is the $SE_{\%}$.

(*) $SE_{\%} \approx CF \times \dfrac{SD_{\text{sample}}}{\sqrt{\text{sample size}}} \times 100\%$

(*) $CF = \sqrt{\dfrac{\text{pop size} - \text{sample size}}{\text{pop size} - 1}}$

(*) If the population size is much bigger than the sample size (which is the usual case), then $CF \approx 1$ and

$$SE_{\%} \approx \frac{SD_{\text{sample}}}{\sqrt{\text{sample size}}} \times 100\%$$

which depends only on the sample size.