Confidence intervals

Let's look at the Gallup poll on the approval rating of the US President.
Suppose 60% of the 140 million likely voters approve of the way the president is handling his job.

Gallup polls 1,000 of them. The resulting approval percentage in the sample will be off the population percentage of 60% due to chance error. How much?
The SE tells us the likely size of the chance error.

Confidence intervals give a more precise statement.

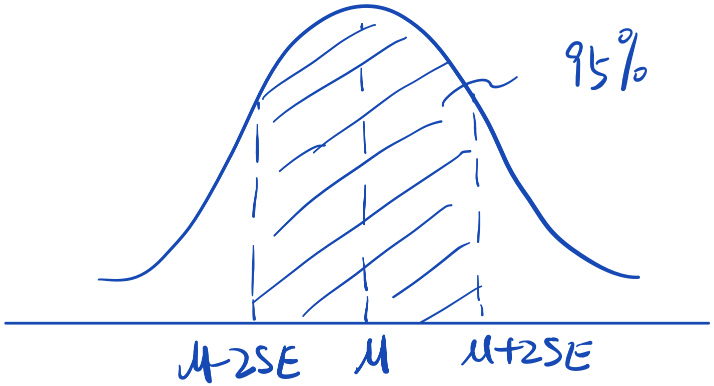
Confidence intervals

According to the central limit theorem, the sample percentage follows the normal curve with expected value $\mu = 60\%$ and SE equal to $\frac{\sigma}{\sqrt{1000}} = \frac{0.49}{31.6} = 1.6\%$.

standard error

(Because we sample from a population of 140 million labels, of which 60% are 1s and 40% are 0s.)

Counting : Introduce labels



By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage μ.

Confidence intervals

By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage μ.

But that is the same as saying that the population percentage is no more than 2 SEs away from the sample percentage.

So once we compute the sample percentage from our sample, say 58%, then we can give a range of plausible values for the unknown population percentage by going 2 SEs in each direction:

[54.8%, 61.2%] is a 95% **confidence interval** for the population percentage.

Interpretation of a confidence interval

[54.8%, 61.2%] is a 95% confidence interval for the population percentage.

Why 'confidence' instead of 'probability'?

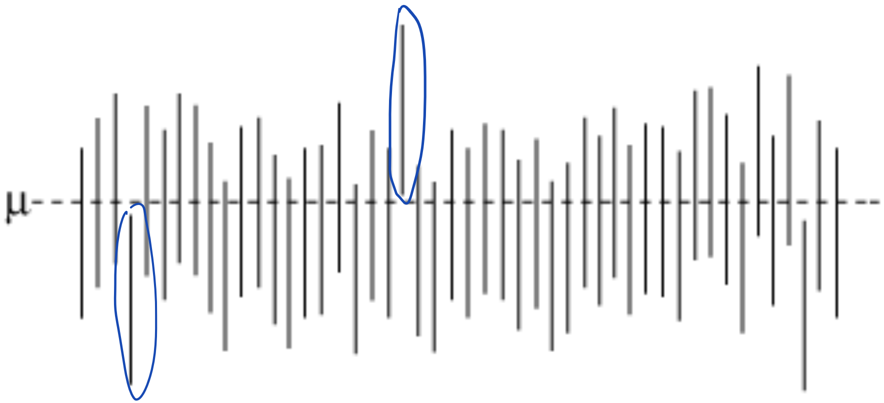
The population percentage μ is a *fixed* number, which either falls into [54.8%, 61.2%] or not. There are no chances involved.

Rather, the chances are in the sampling procedure:
a different sample of 1,000 voters will give a slightly different interval.

If one does many polls (the Gallup poll is done frequently), then 95% of these intervals trap the population percentage, and 5% will miss it.

95% is called the **confidence level**.

Interpretation of a confidence interval



"I am 95% confident that the President's approval rating is between 54.8% and 61.2%" means that 95% of the time I am correct when making such a statement based on a poll.

Keep in mind that the interval varies from sample to sample, while the population percentage is a fixed number.

Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter* μ.

Usually the confidence interval is centered at an estimate for μ which is an average.

CLT tells us that the distribution of averages of samples of salaries is close to a normal distribution even if the distribution of all salaries is quite skewed.

Examples:

1. μ = approval percentage among all 140 million likely voters.
   estimate = approval percentage among voters in the sample.
2. μ = speed of light.
   estimate = average of 30 measurements.

Since the central limit theorem applies for averages, the confidence interval has a simple form:

standard error

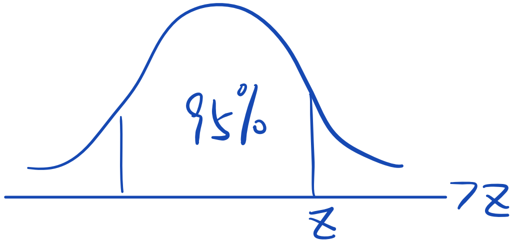
$$\text{estimate} \pm z\,\text{SE}$$

where z is the z-score corresponding to the desired confidence level:

95% confidence level → $z = 1.96$

90% confidence level → $z = 1.65$

99% confidence level → $z = 2.58$



Estimating the SE with the bootstrap principle

SE is the standard error of the estimate. If the estimate is an average (e.g. a percentage), then we know that

$$\boxed{\text{SE} = \frac{\sigma}{\sqrt{n}}}$$

where σ is the standard deviation of the population.

Now we have a problem: we don't know σ because we don't know all the data in the population (that's the reason why we sample in the first place!)

The **bootstrap principle** states that we can estimate σ by its sample version s and still get an approximately correct confidence interval.

$$\begin{cases} \sigma: \text{population standard deviation} \\ s: \text{sample standard deviation} \end{cases}$$

Estimating the SE with the bootstrap principle

The **bootstrap principle** states that we can estimate σ by its sample version s and still get an approximately correct confidence interval.

Examples:

1. We poll 1,000 likely voters and find that 58% approve of the way the president handles his job.

   binomial distribution

   SE $= \frac{\sigma}{\sqrt{n}} \times 100\%$, where $\sigma = \sqrt{p(1-p)}$, $p =$ proportion of all voters who approve.

   The bootstrap principle replaces σ by $s =$ standard deviation of the 0/1 labels in the sample $= \sqrt{0.58(1-0.58)} = 0.49$.

   So a 95% confidence interval for p is

if p=0.2

$$58\% \pm 2\frac{0.49}{\sqrt{1000}}, \qquad \text{which is } [54.9\%, 61.1\%]$$

Estimating the SE with the bootstrap principle

2. estimate = average of 30 measurements of the speed of light.

   What is the population or probability histogram from which we sample?

   The reason we get 30 different measurementss is because each is off by a chance error:

   $$\text{measurement} = \overset{\text{fixed number}}{\text{speed of light}} + \text{measurement error}$$

   The measurement error follows a probability histogram that is unknown to us. We estimate the standard deviation σ of this probability histogram by the standard deviation s of the sample of 30 measurements.

Later we will see how the bootstrap principle can be used to approximate the SE of estimates other than averages.

More about confidence intervals

- The width of the confidence interval is determined by $z\,\text{SE}$, which is called the **margin of error**.

  A larger sample size n will result in a smaller margin of error since $\text{SE} = \frac{\sigma}{\sqrt{n}}$, but we need *four times the sample size to cut the width in half*.

  We can also make the width smaller by making z smaller, e.g. use a 80% confidence level instead of 95%. Then the price for more precision (i.e. shorter interval) is less confidence that it covers the parameter μ.

- There is an easy to remember formula for a 95% confidence interval for a percentage:

  rule of thumb

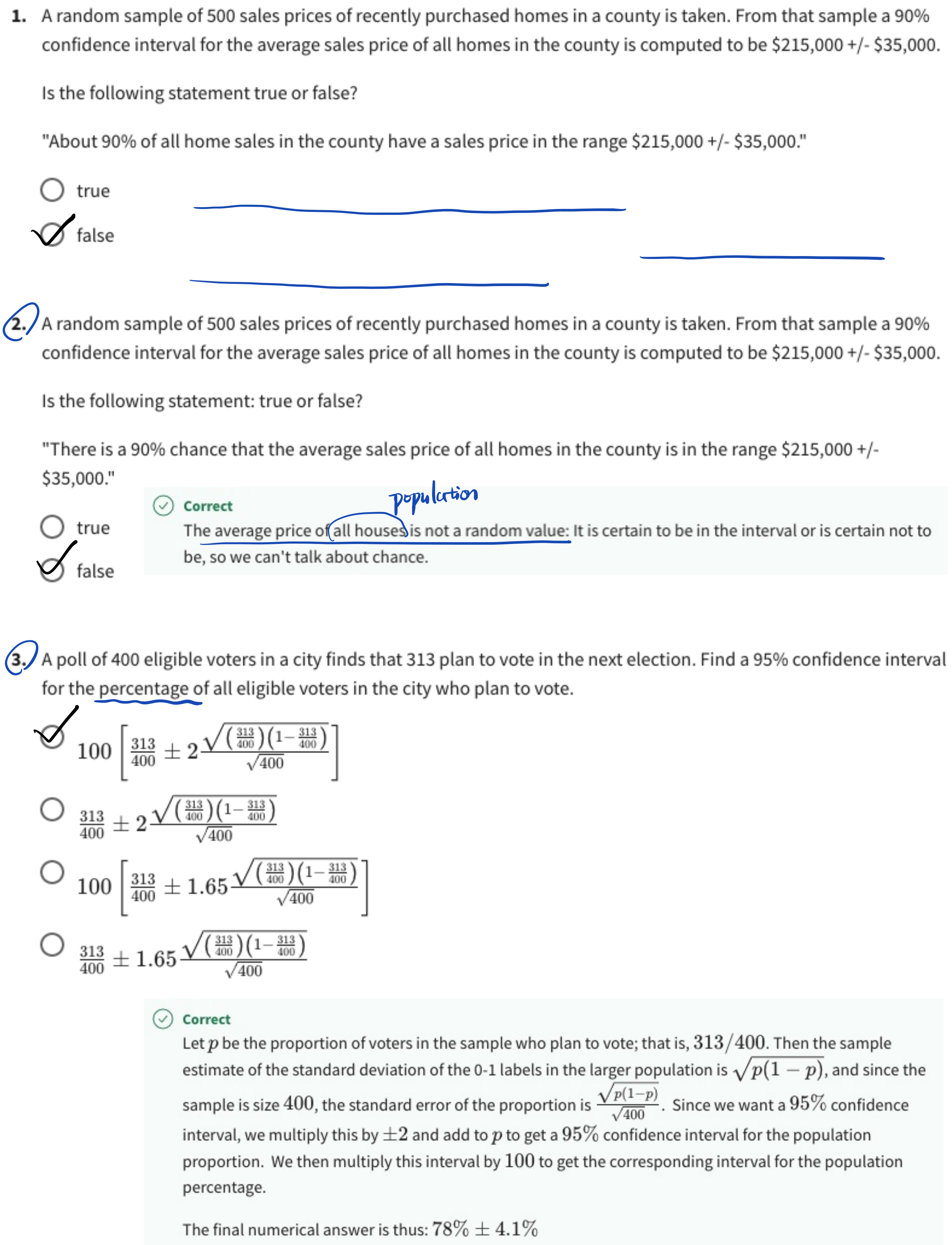
  $$\text{estimated percentage} \pm \frac{1}{\sqrt{n}}$$

  $zSE = 2 \cdot \frac{1}{2} = 1.$

  That's because $\sigma = \sqrt{p(1-p)} \leq \frac{1}{2}$ no matter what p is.

- It is journalistic convention to use a 95% confidence level unless stated otherwise.

Quiz



1. A random sample of 500 sales prices of recently purchased homes in a county is taken. From that sample a 90% confidence interval for the average sales price of all homes in the county is computed to be $215,000 +/- $35,000.

Is the following statement true or false?

"About 90% of all home sales in the county have a sales price in the range $215,000 +/- $35,000."

○ true

✓ false

2. A random sample of 500 sales prices of recently purchased homes in a county is taken. From that sample a 90% confidence interval for the average sales price of all homes in the county is computed to be $215,000 +/- $35,000.

Is the following statement: true or false?

"There is a 90% chance that the average sales price of all homes in the county is in the range $215,000 +/- $35,000."

○ true

✓ false

Correct

population

The average price of all houses is not a random value: It is certain to be in the interval or is certain not to be, so we can't talk about chance.

3. A poll of 400 eligible voters in a city finds that 313 plan to vote in the next election. Find a 95% confidence interval for the percentage of all eligible voters in the city who plan to vote.

✓ $100\left[\frac{313}{400} \pm 2\frac{\sqrt{\left(\frac{313}{400}\right)\left(1-\frac{313}{400}\right)}}{\sqrt{400}}\right]$

○ $\frac{313}{400} \pm 2\frac{\sqrt{\left(\frac{313}{400}\right)\left(1-\frac{313}{400}\right)}}{\sqrt{400}}$

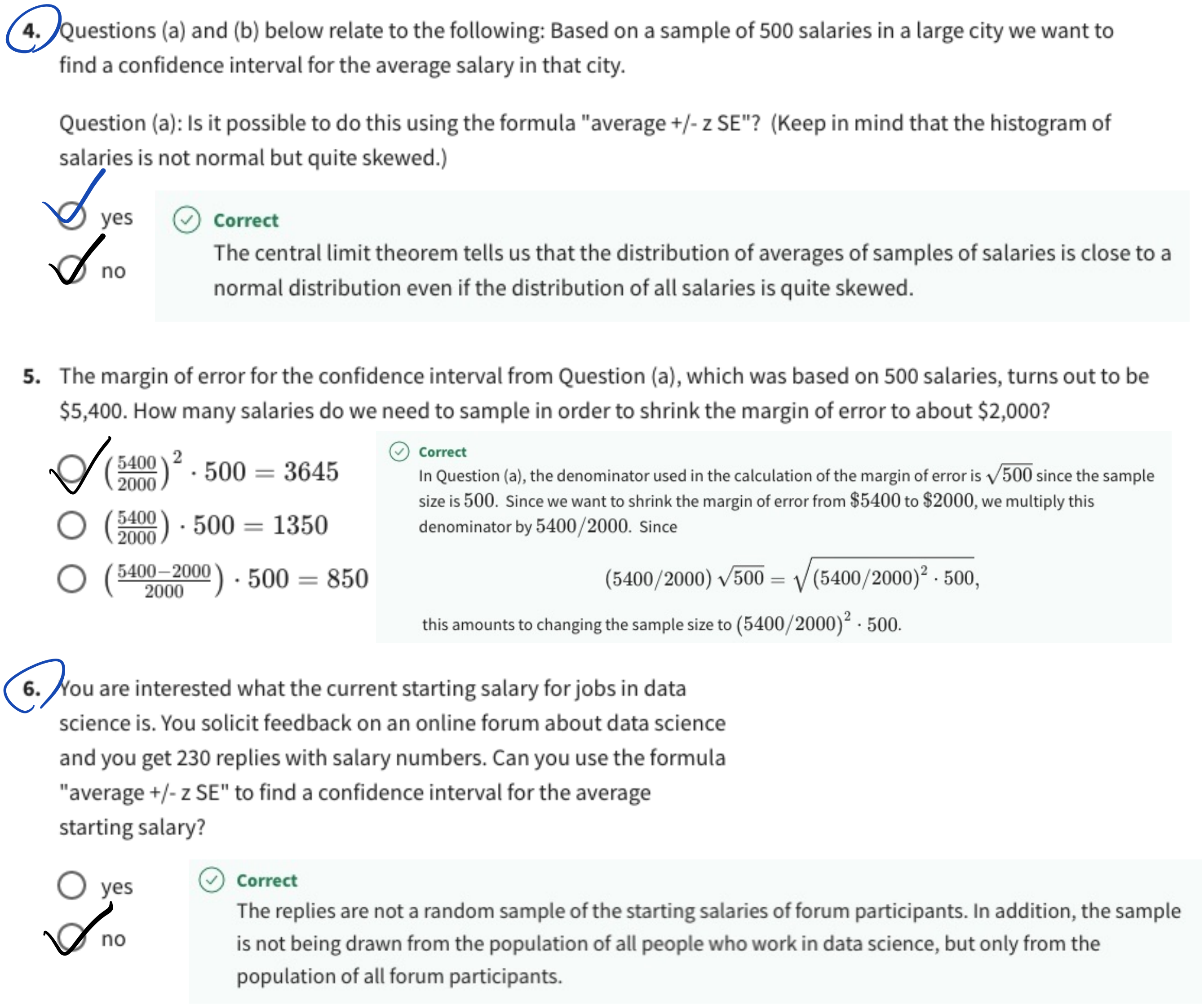
○ $100\left[\frac{313}{400} \pm 1.65\frac{\sqrt{\left(\frac{313}{400}\right)\left(1-\frac{313}{400}\right)}}{\sqrt{400}}\right]$

○ $\frac{313}{400} \pm 1.65\frac{\sqrt{\left(\frac{313}{400}\right)\left(1-\frac{313}{400}\right)}}{\sqrt{400}}$

Correct

Let p be the proportion of voters in the sample who plan to vote; that is, $313/400$. Then the sample estimate of the standard deviation of the 0-1 labels in the larger population is $\sqrt{p(1-p)}$, and since the sample is size 400, the standard error of the proportion is $\frac{\sqrt{p(1-p)}}{\sqrt{400}}$. Since we want a 95% confidence interval, we multiply this by ± 2 and add to p to get a 95% confidence interval for the population proportion. We then multiply this interval by 100 to get the corresponding interval for the population percentage.

The final numerical answer is thus: $78\% \pm 4.1\%$



4. Questions (a) and (b) below relate to the following: Based on a sample of 500 salaries in a large city we want to find a confidence interval for the average salary in that city.

Question (a): Is it possible to do this using the formula "average +/- z SE"? (Keep in mind that the histogram of salaries is not normal but quite skewed.)

[✓ ○ yes]

[✓ ○ no]

✓ **Correct**

The central limit theorem tells us that the distribution of averages of samples of salaries is close to a normal distribution even if the distribution of all salaries is quite skewed.

5. The margin of error for the confidence interval from Question (a), which was based on 500 salaries, turns out to be $5,400. How many salaries do we need to sample in order to shrink the margin of error to about $2,000?

[✓ ○ $\left(\frac{5400}{2000}\right)^2 \cdot 500 = 3645$]

○ $\left(\frac{5400}{2000}\right) \cdot 500 = 1350$

○ $\left(\frac{5400-2000}{2000}\right) \cdot 500 = 850$

✓ **Correct**

In Question (a), the denominator used in the calculation of the margin of error is $\sqrt{500}$ since the sample size is 500. Since we want to shrink the margin of error from $\$5400$ to $\$2000$, we multiply this denominator by $5400/2000$. Since

$$(5400/2000)\sqrt{500} = \sqrt{(5400/2000)^2 \cdot 500},$$

this amounts to changing the sample size to $(5400/2000)^2 \cdot 500$.

6. You are interested what the current starting salary for jobs in data science is. You solicit feedback on an online forum about data science and you get 230 replies with salary numbers. Can you use the formula "average +/- z SE" to find a confidence interval for the average starting salary?

○ yes

[✓ ○ no]

✓ **Correct**

The replies are not a random sample of the starting salaries of forum participants. In addition, the sample is not being drawn from the population of all people who work in data science, but only from the population of all forum participants.