

Chapter 10: Introduction to Inference

The purpose of sampling is to "infer" a conclusion about the population. If we sample the OHS student body many times and find that the mean GPA of our samples is 3.0, then we want to infer that the mean GPA of the whole student body is 3.0.

Statistical Inference - methods for drawing conclusions about a population from sample data.

10.1 - Confidence Intervals and 10.2 - Tests of Significance

- the two most common types of formal statistical inference
- both based on sampling distributions of statistics
- formal inference is based on probability theory (long run behavior)
- properly randomized design is critical
 - don't use formal inference unless you are satisfied that the data merits the analysis

Estimating

Ex (10.2 p.537):

We want to estimate the mean SAT Math score for the >350,000 HS seniors in California. About 49% of CA students take the test. The ones who do are self-selected, so not representative of all CA seniors - meaning that we couldn't make inferences about the population based on any samples of this group.

Instead, we take an SRS of 500 CA seniors and get $\bar{x} = 461$. Now can we draw an inference?

We know that since the sample was random, our mean score should be close to μ , which we don't know. We guess that μ is somewhere near 461, but since we want to be more accurate, we take many samples of the same size (500) and study the collection of mean values.

Key points about the sample mean distribution:

✓ CLT says that mean \bar{X} of 500 scores will have an approx normal dist.

✓ $\mu_{\bar{x}} = \mu$

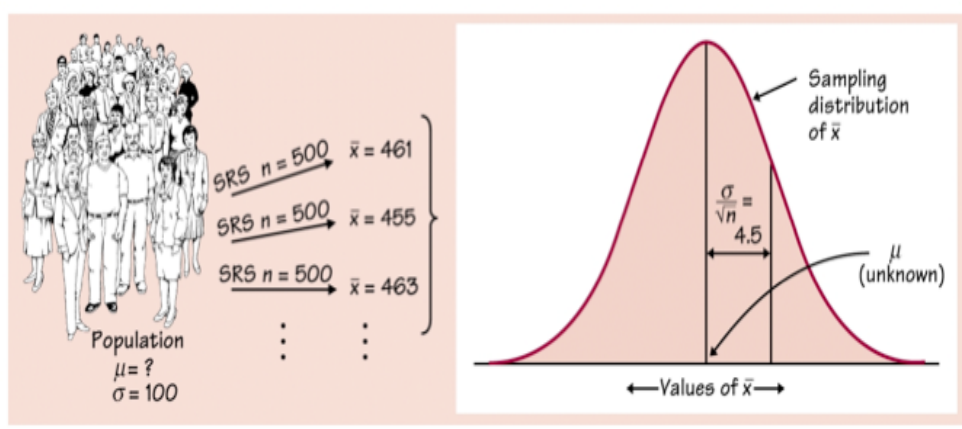
✓ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{500}}$ (σ is for all CA HS seniors)

Normally we won't know σ , so for now let's pretend it is 100.

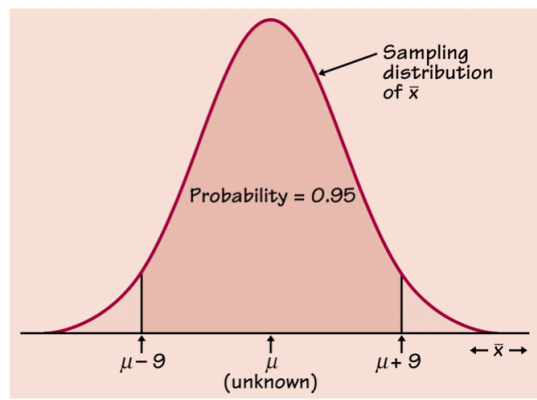
$$\sigma_{\bar{x}} = \frac{100}{\sqrt{500}} = 4.5$$

As we sample many times, we will get many different sample means. If we collect all possible, \bar{X} will be $N(\mu, 4.5)$

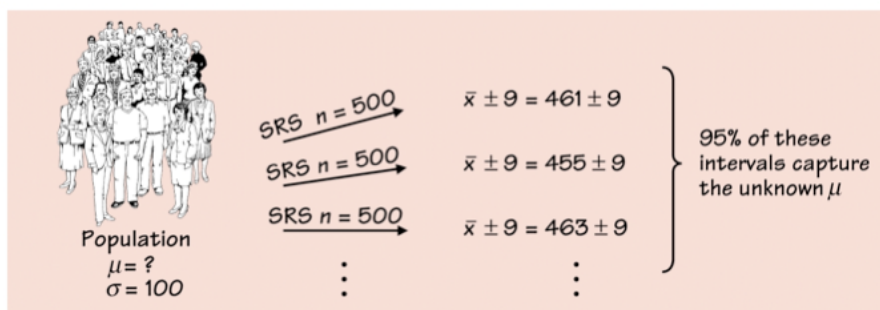
This is what our sampling distribution will look like, but we still don't know the μ .



Same sampling distribution, but with the $P=.95$ Empirical rule shaded.



The Empirical rule says that in 95% of all samples, the mean score \bar{x} will be within 2σ (9 points) of μ .



We use probability (long run) theory to talk about our confidence in any one sample.

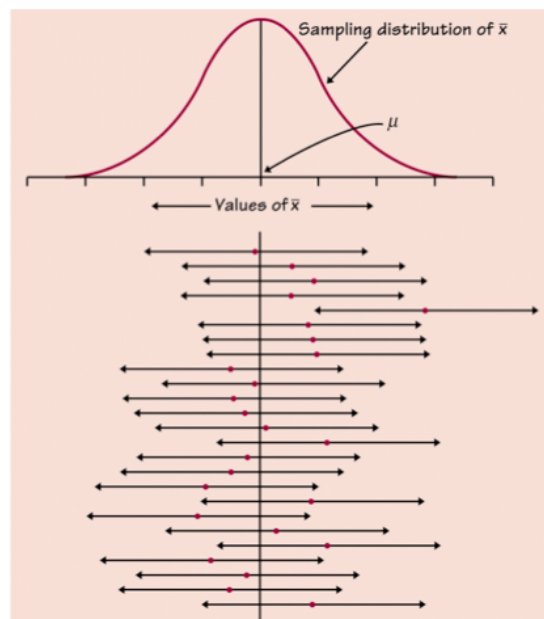
Our sample had $\bar{x} = 461$. We are 95% confident that the mean SAT Math score for all CA HS seniors likes between 452 and 470, (461 ± 9) .

One of two things must be true:

1. The interval between 452 and 470 contains the true μ .
2. Our sample was one of the 5% for which \bar{x} is not within 9 points of the true μ .

Confidence Interval for a parameter has two parts:

1. **estimate \pm margin of error** $\bar{x} \pm m/e$
2. **confidence level C** (C=.95 in our example) gives the ratio for capturing the true parameter in repeated samples.



Common graphical display of confidence interval:

C=.95 for 25 SRSs
confidence interval is the arrow
 \bar{x} is the dot.

We say that "95% of our samples will capture the mu of the population."

Constructing a Confidence Interval

What we need (the conditions):

1. data should come from an SRS
2. sampling distribution of \bar{x} should be approx. normal

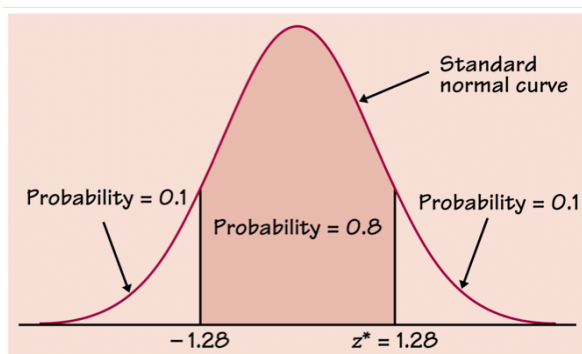
The big idea: to "catch" the right probability C under a normal curve. Measure C in terms of standard deviations - or in this case a new term: z^*

Think - if we want 95% confidence, then we will measure that with two standard deviations from the mean, so z^* would be about 2 (using the empirical rule).

Find z^* :

To find an 80% confidence interval, we need to shade the central 80% of a normal distribution. That means 10% in each tail. To find the z^* , we are really just finding the z score for .9 (and .1) of the standard normal curve.

The probability that any value is between -1.28 and 1.28 is .8 or 80%.



Note: you can use `invNorm(.9)` to generate the same value for z^* or use Table A.

Most common confidence levels:

Confidence Level	Tail Area	z^*
90%	0.05	1.645
95%	0.025	1.960
99%	0.005	2.576

***Note that we use 1.96 for 95% since it is more accurate than the 2% that the Empirical rule states

Table C gives more z^* values for a wider range of confidence levels

4 Steps to calculate a confidence interval:

1. Identify the parameter and population of interest
2. Choose the appropriate inference procedure (for now it will be a confidence interval (CI) and verify the conditions (SRS & approx normal sampling dist)
3. Calculate the interval:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

remember yesterday:
2(4.5) gave the ± 9

4. Interpret your results in the context of the problem

P.548 # 10.5

Pharmaceuticals problem...specimen selected from each batch and analyzed 3 times, reporting the mean result. Results follow normal distribution and analysis procedure has no bias. $\sigma = .0068$. Three analyses of one specimen give concentrations: 0.8403, 0.8363, 0.8447. Construct a 99% confidence interval for true concentration μ . Follow 4 step procedure.

1. I want to estimate μ , the true concentration of active ingredient in the drug.
2. Since σ is known, I will construct a Confidence Interval
 - an SRS guarantees no bias
 - it is given that the distribution is approximately normal.

3. $\bar{x} = \frac{.8403 + .8363 + .8447}{3} = .8404$

$$\bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$.8404 \pm (2.576)(.003926)$$

$$.8404 \pm .0101 \quad (.8303, .8505)$$

$$z^* = 2.576$$

$$\frac{\sigma}{\sqrt{n}} = \frac{.0068}{\sqrt{3}} = .003926$$