Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

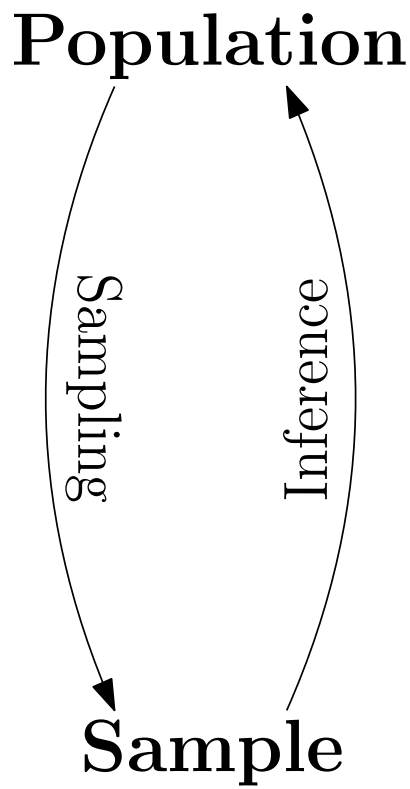
Today: Chapter 4 (beginning)

- Point estimate, sample size, variations
- Population distribution VS Sampling distribution
- Sample distribution for the sample mean
- Confidence intervals
- Stating things like: "We are XX% confident that..."

Population VS Sample

Population: Everything you want to study
Parameter: A value summarizing the population
Example:

- Population: All adults in the US
- Parameter: Mean weight in kilograms



Sample: A representative subset
Statistic: Value summarizing the sample
Example:

- Sample: 30 adults from a nearby mall
- Average of their 30 weights

A statistic from a sample is also known as a **point estimate** (of the population parameter).

Point Estimates are... Estimates

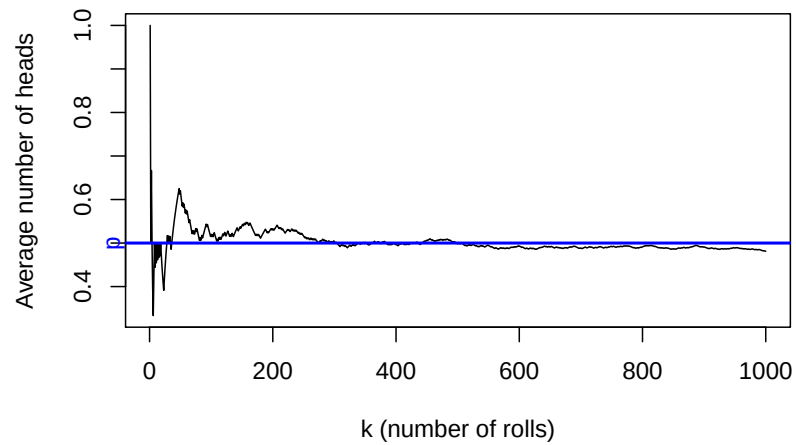
Because a sample represents a loss of information, statistics based on it are usually wrong (= they don't give the exact parameter values).

In other words, because of randomness, there are inherent variations in any statistic.

The size of the sample (denoted n) gives some indication of how good the statistic is in estimating the parameter.

$$\text{statistic} \xrightarrow[n \to \text{population size}]{\qquad\qquad} \text{parameter}$$

Though, values of the statistic fluctuate when n is fixed.



But How Bad Will My Estimate Be?

The answer to this depends on your sample size n, the particular members of your sample, and the distribution of the population you are trying to study.

Example:
You land on an alien planet and are tasked with finding the average weight of the alien species.
Which set up makes you feel better about your point estimate?

Sample size: Weighing 30 aliens VS **300 aliens**

Weight in your sample: **all around 90kg** VS wildly different weights

Population distribution: **Normal on [80,90]** VS Uniform on [80,90]

Sample Mean

Assume you have data $X_1, X_2, \ldots, X_n$, that you sampled in a population of interest. Recall that given the **sample mean** is:

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

This is a natural candidate to estimate the **population mean**:

$$\mu = E(X)$$

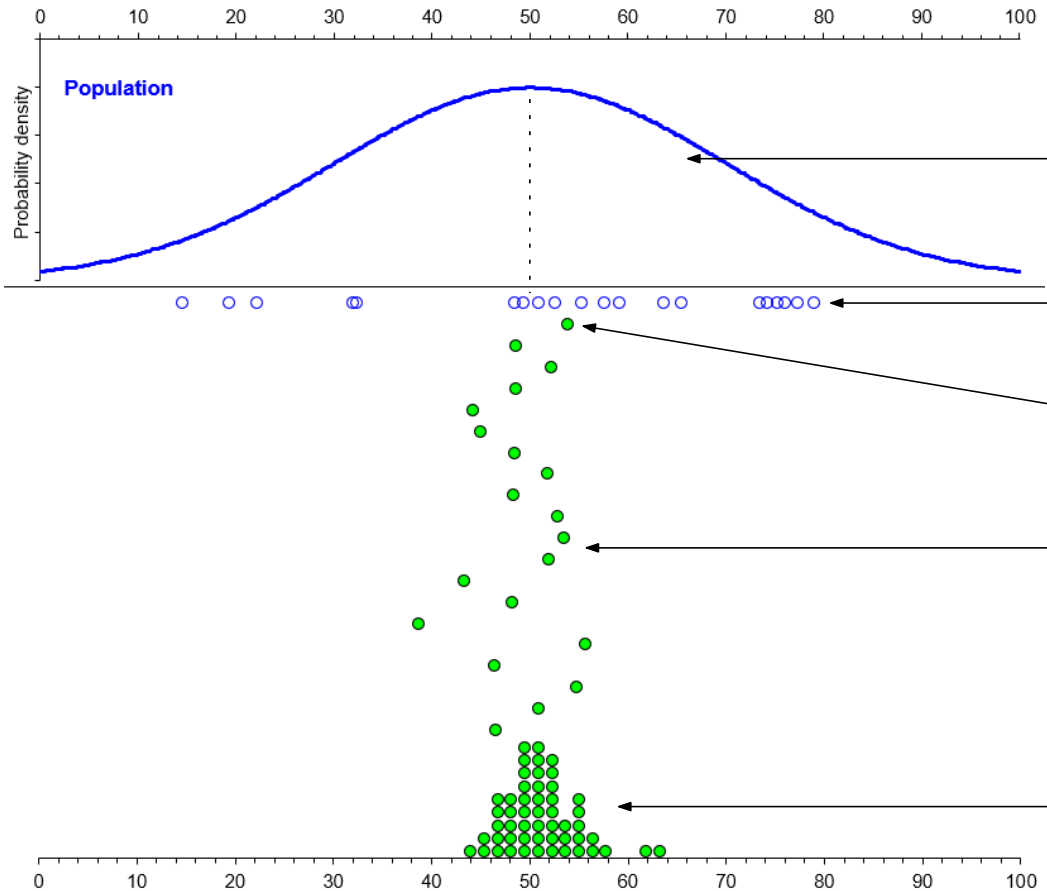
Key Idea: Variability in sampling yields variability in $\bar{X}_n$.

Goal: Understand the variability of $\bar{X}_n$.

In order to: Understand the weakness/assets of $\bar{X}_n$.

Exploratory Software for Confidence Intervals

ESCI: Excel spreadsheets (by G. Cummings) to explore sampling and confidence intervals. (free download, enable Excel macros to use)



Distribution of the idea we're studying (e.g. weight of alien)
This is invisible to us!

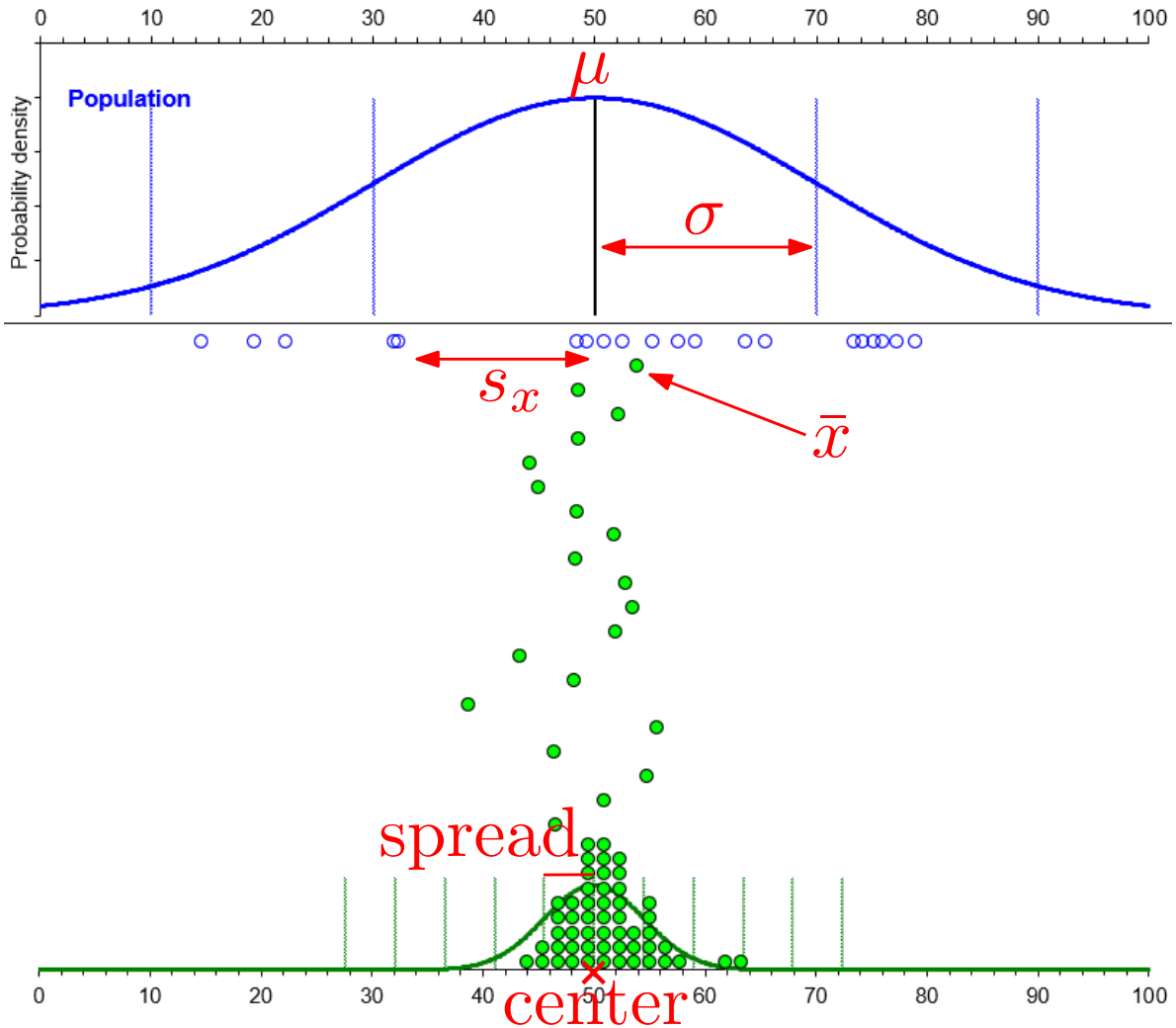
The individuals in the particular sample we drew

The mean weight from the pictured sample

Means from other samples

Making a histogram of all the means from all the samples

Some Notation



Population:
Center: μ
Spread: σ

One sample:
Center: $\bar{x}$
Spread: s_x

Sampling distribution:
Center: ?
Spread: ?

Center and Spread of the Sampling Distribution

Assume that $X_1, X_2, \ldots, X_n$ are independent random variables that have a common distribution with parameters:

$$E(X) = \mu \qquad\qquad Var(X) = \sigma^2.$$

$$\begin{aligned} E(\bar{X}_n) &= E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\ &= \frac{1}{n}\left(E(X_1) + E(X_2) + \ldots + E(X_n)\right) \\ &= \frac{1}{n}\left(\mu + \mu + \ldots + \mu\right) = \mu. \end{aligned}$$

$$\begin{aligned} Var(\bar{X}_n) &= Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) \\ &= \frac{1}{n^2}\left(Var(X_1) + Var(X_2) + \ldots + Var(X_n)\right) \\ &= \frac{1}{n}\left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right) = \frac{\sigma^2}{n}. \end{aligned}$$

Center and Spread of the Sampling Distribution

If $X_1, X_2, \ldots, X_n$ are independent random variables that have a common distribution with parameters:

$$E(X) = \mu \qquad\qquad SD(X) = \sigma,$$

then,

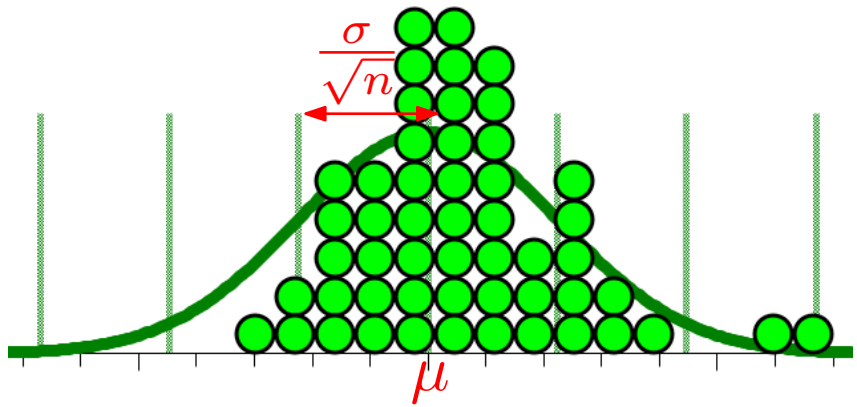
$$E(\bar{X}_n) = \mu \qquad\qquad SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}.$$

The **standard error** defined to be the spread of the sampling distribution. That is

$$SE = \frac{\sigma}{\sqrt{n}}.$$

Sample distribution in the Normal Case

Furthermore, if the population distribution is normal, then the sampling distribution of $\bar{X}_n$ is a normal curve.



As a consequence, the green curve is $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

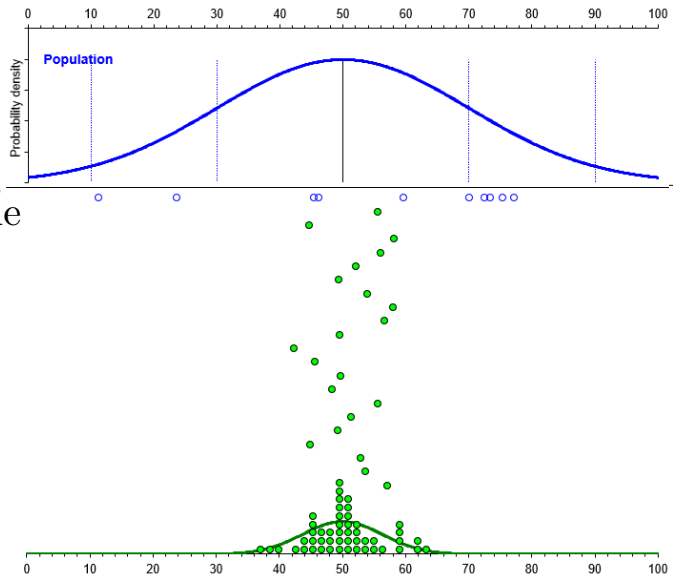
Remark: The lower the standard error, the better. Hence,

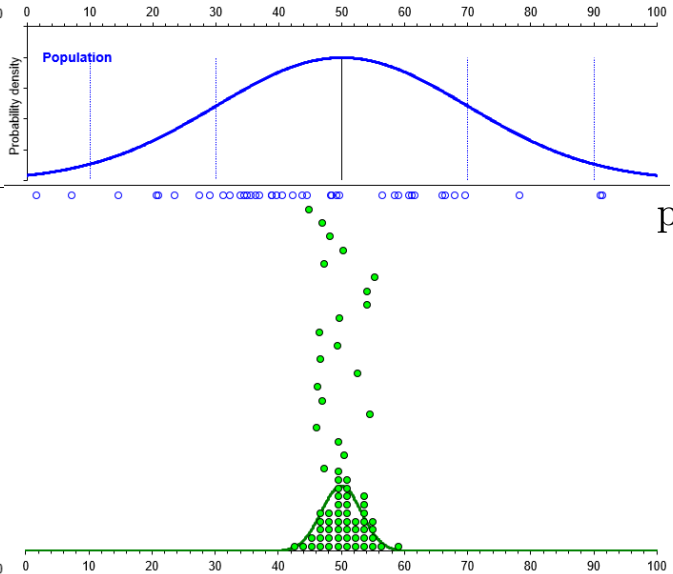
- Smaller population spread is better than larger, as σ shows up.
- Bigger sample is better than smaller sample, as $\frac{1}{\sqrt{n}}$ shows up.

Algebraically Confirming Our Intuition

If we wish to find the average weight of an alien species, why is a sample of 40 aliens better than a sample of 10 aliens?

10 aliens per sample





40 aliens per sample

$$SE_{10} = \frac{\sigma}{\sqrt{10}} \qquad SE_{40} = \frac{\sigma}{\sqrt{40}} = \frac{\sigma}{2\sqrt{10}}$$

Summary

Take a population with mean μ and standard deviation σ for some trait (e.g. weight).
For now, you need to assume that this trait is normally distributed.

Fix n and take a sample of n independent observations. Calculate the mean for the sample: $\bar{x}$. This is a point estimate for the population parameter μ.

This is just one possible estimate. If we make a histogram of all possible estimates, this is the sampling distribution, which has shape

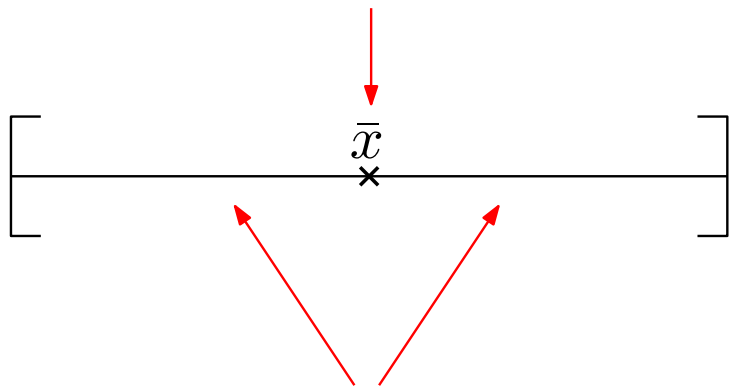
$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Remark: Since independence is hard to check for sample members, we usually check if they were randomly chosen and are $< 10\%$ of the population size.
This condition approximates the idea of independence.

Using Our Understanding of the Sampling Distribution

The sampling distribution helps us create a **confidence interval**: a range of values around a point estimate that convey our uncertainty about the population parameter (as well as a range of plausible values for it).

At the center is our point estimate based on a single sample
It is our best guess for the population parameter

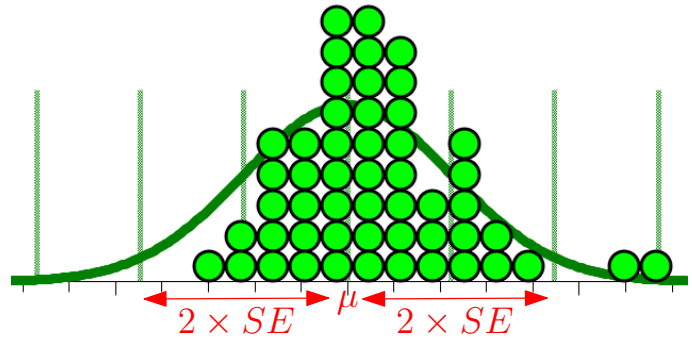


We also stretch our guess by adding the same amount on both sides to show our uncertainty

The amount you stretch your IC (= its width) is determined by how sure you want to be that the interval will contain the true population parameter μ.

Building a Confidence Interval

Recall that for any normal curve (here, the sampling distribution), about 95% of all values fall within 2 standard deviations from the mean.



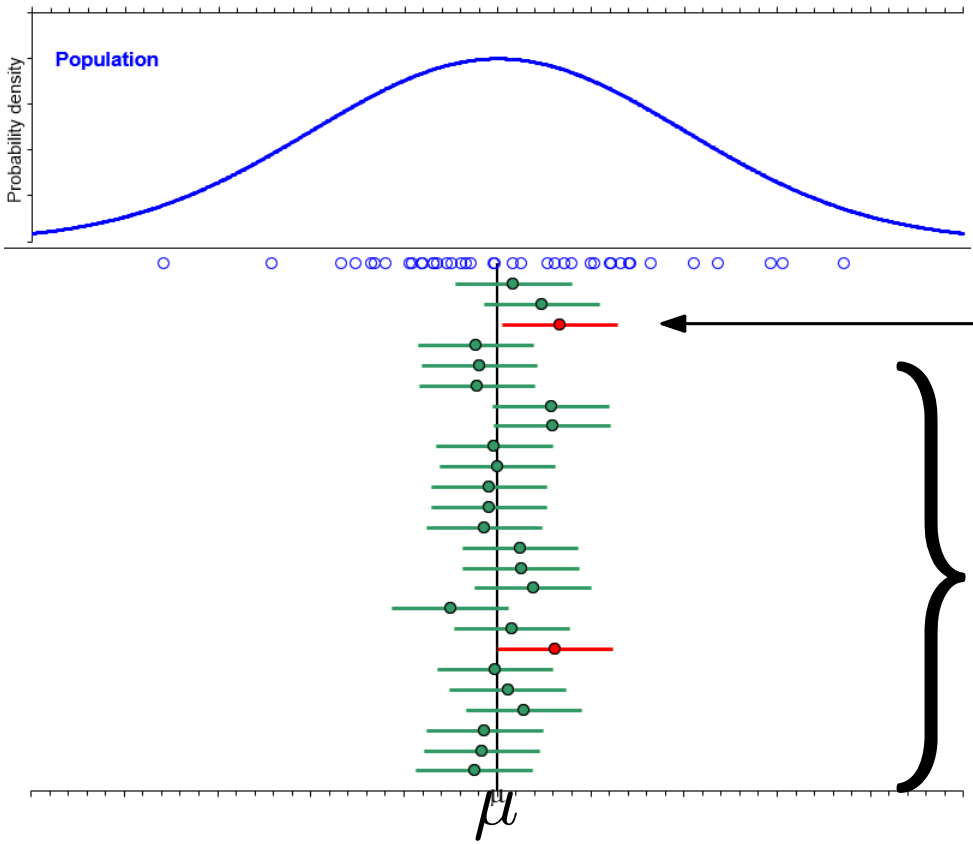
So 95% of all point estimates (green dots) are withing $\pm 2SE$ of μ.

Said differently:

- Stand at μ. Reach out about $2SE$'s.
  You will grab about 95% of sample means (green dots)
- Stand at a green dot. Reach out about $2SE$'s.
  If you did this at every green dot, about 95% of those reaches include μ.

Let's try that on ESCI.
(Make sure to select C.I.'s in 6, and deselect Mean Heap in 5)

Remember: Your Sample is One of Many!



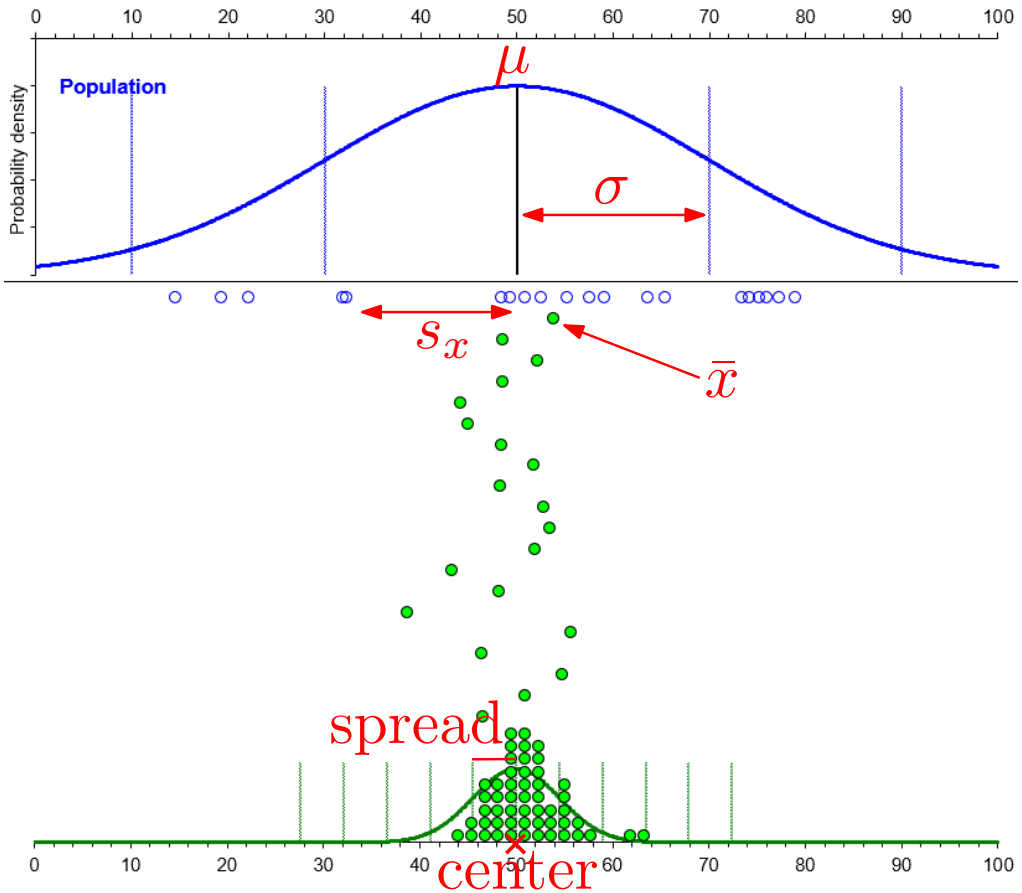
A C.I. that doesn't capture the true mean

Most C.I.'s do capture μ

If we fix a C.I. width of $\pm 2SE$, we know that about 95% of these intervals capture the (unknown true) mean μ.

Your first Confidence Interval

On the alien planet, you draw a sample of $n = 50$ aliens and in that sample calculate a mean of 59 kg with a standard deviation of 1.4 kg. What notation should we assign to these numbers?



We have:

$$\bar{x} = 59$$
$$s_x = 1.4$$

While our best guess for μ is $\bar{x} = 59$, we choose to report a 95% C.I. Find it. The C.I. in this case is

$$\bar{x} \pm 1.96SE$$
$$= \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Remark: The truth is that 95% of data is withing 1.96 SD (not 2 SD).

If You Know It, Use It. If Not, Approximate It

We often do not know σ, but it is well approximated by s_x.

Our C.I. is roughly $\bar{x} \pm 1.96\frac{s_x}{\sqrt{n}} = 59 \pm 1.96\frac{1.4}{\sqrt{50}}$

The C.I. is $[58.61, 59.38]$.

What language should we use to report our answer?

> We are 95% confident that the true mean is in $[58.61, 59.38]$.

What Does "95% Confident" Mean?

Statement:

> "We are 95% confident that the true mean is in $[58.61, 59.38]$."

Incorrect interpretation: There's a 95% chance the true mean lies in this interval.

The probability the true mean lies in the interval is either 0% or 100%. There is nothing random about the true mean or our interval, once it is built.

Correct interpretation: Our C.I. is one of infinitely-many C.I's (built from all the size n samples we could ever draw). We've chosen a width for these so that 95% of these will capture the true (fixed) population mean.

Questions

100 different research teams visit the alien planet. Each team draws a random sample, calculates the average and forms a 95% C.I.
About how many of these C.I.'s would we expect to contain the true average height of all aliens?

(i) 0

(ii) 5

(iii) 95

(iv) 100

Answer: (iii) by construction of the C.I.'s!

Which do you expect to be the smallest?

(i) Standard deviation of heights in the population

(ii) Standard deviation of heights in a sample of 30 aliens

(iii) Standard deviation of the height averages in samples of size 30

(iv) All will be rouhly equal

Answer: (iii), since (i) is σ, (ii) is $s_x \simeq \sigma$, and (iii) is $SE = \sigma/\sqrt{n}$.

Questions

Our 100 research teams get together and decide they will be better: now they want about 99 of the 100 teams to produce C.I.'s that contain the true mean.
How will this choice affect the size of the C.I.'s they all generate?

(i) The C.I.'s get smaller

(ii) The C.I.'s get wider

(iii) The C.I.'s stay roughly the same size

(iv) Cannot be determined

Answer: (ii), since if you want to be more sure catching the true mean, you must increase the width of your confidence interval.

That's the irony: The more sure you want to be, the wider the confidence interval.
But the wider the confidence interval, the less accurate it is (i.e less informative).