Statistical Inference

Statistics ?

- Statistics is the science of data

Individuals and Variables

- Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.

- A variable is any characteristic of an individual. A variable can take different values for different individuals.

Two types of variables

A variable can be either

- A categorical variable places an individual into one of several groups or categories. What can be counted is the count or proportion of individuals in each category.

or

- A quantitative variable takes numerical values for which arithmetic operations such as adding and averaging make sense.

- The distribution of a variable tells us what values it takes and how often it takes these values.

Example

- Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?

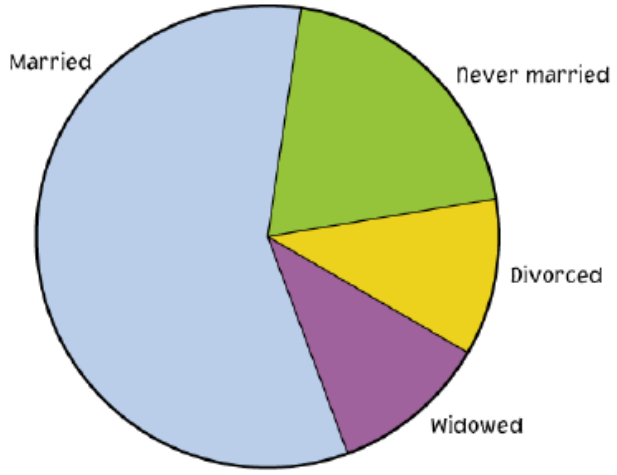
a) Gender (female or male) --- Categorical
b) Age (years) --- Quantitative
c) Race (black, white or other) --- Categorical
d) Smoker (yes or no) --- Categorical
e) Systolic blood pressure --- Quantitative
f) Level of calcium in the blood --- Quantitative

Ways to chart categorical data

Because the variable is categorical, the data in the graph can be ordered any way we want (alphabetical, by increasing value, by year, by personal preference, etc.).

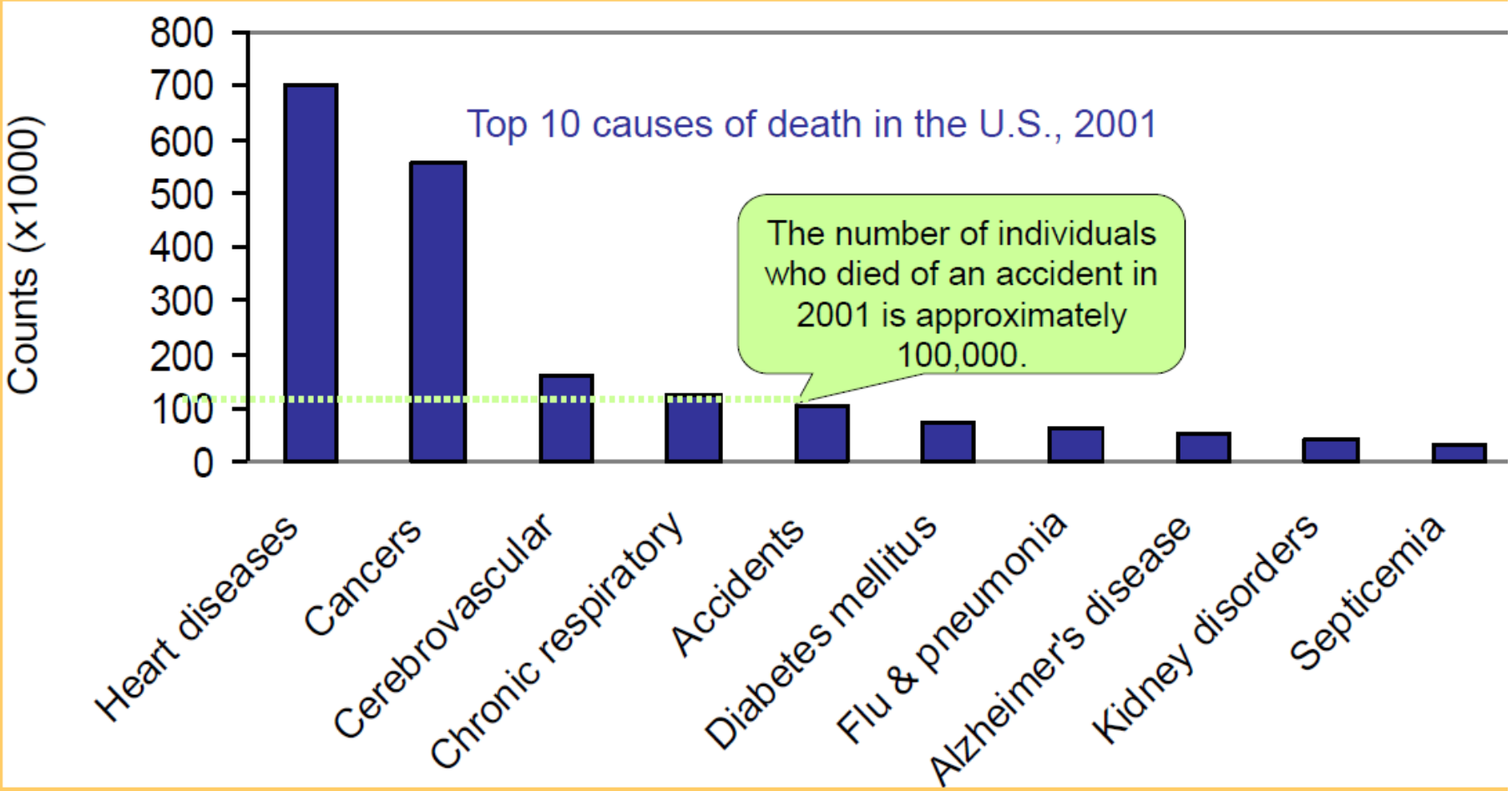
- **Bar graphs**
  Each category is represented by a bar.

- **Pie charts**
  Peculiarity: The slices must represent the parts of one whole.



Bar graphs

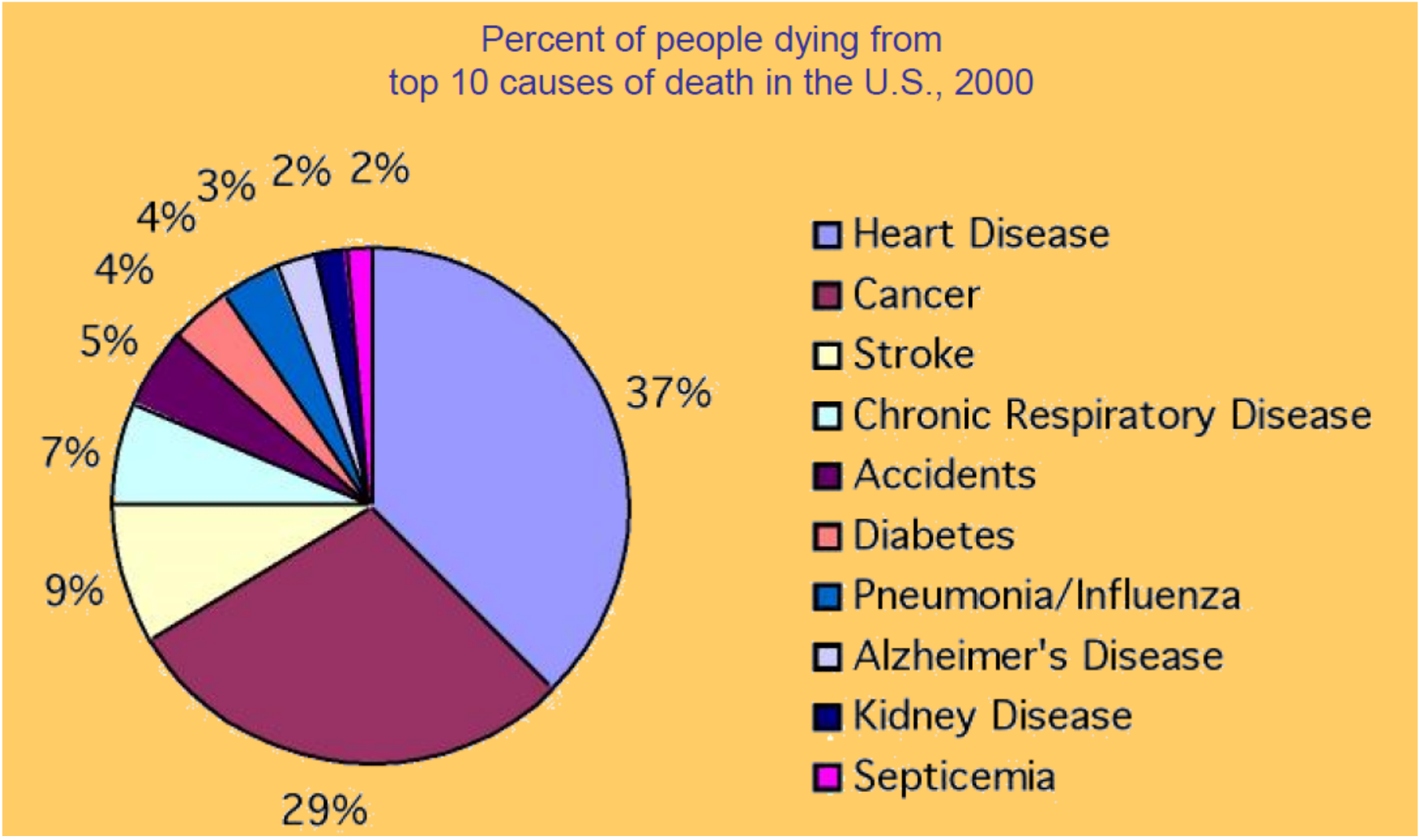
Each category is represented by one bar. The bar's height shows the count (or sometimes the percentage) for that particular category.



Pie charts

Each slice represents a piece of one whole.

The size of a slice depends on what percent of the whole this category represents.



Ways to chart quantitative data

- Histograms and stemplots

  These are summary graphs for a single variable. They are very useful to understand the pattern of variability in the data.

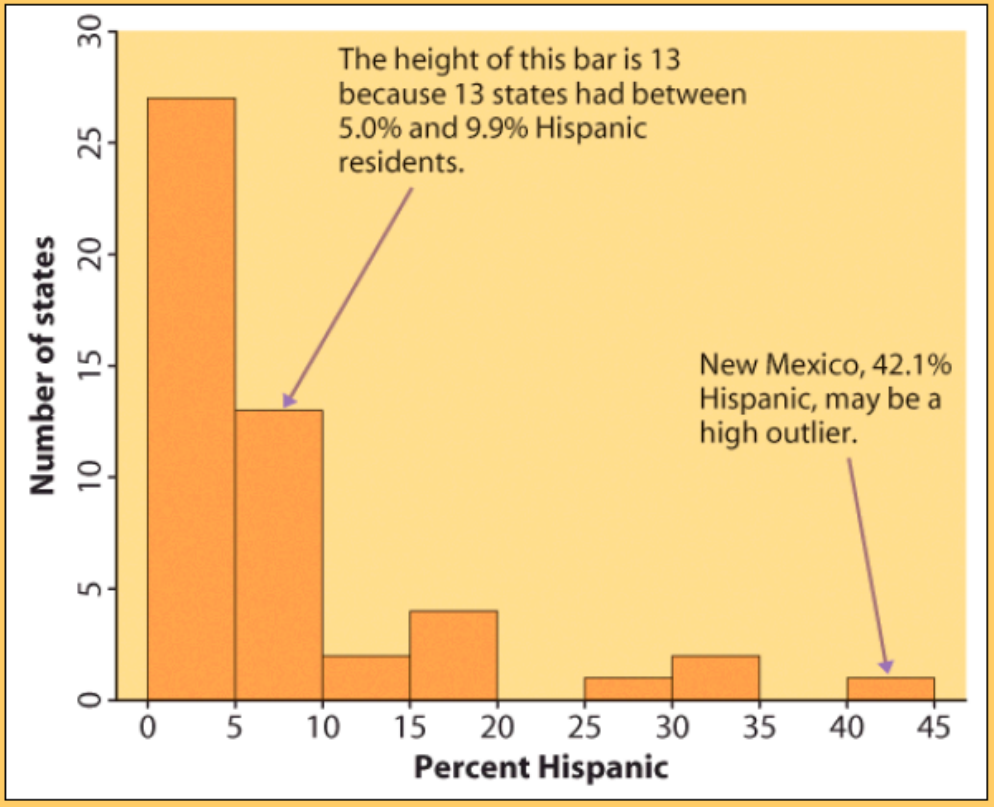
- Line graphs: time plots

  Use when there is a meaningful sequence, like time. The line connecting the points helps emphasize any change over time.

Histograms

The range of values that a variable can take is divided into equal-size intervals.

The histogram shows the number of individual data points that fall in each interval.

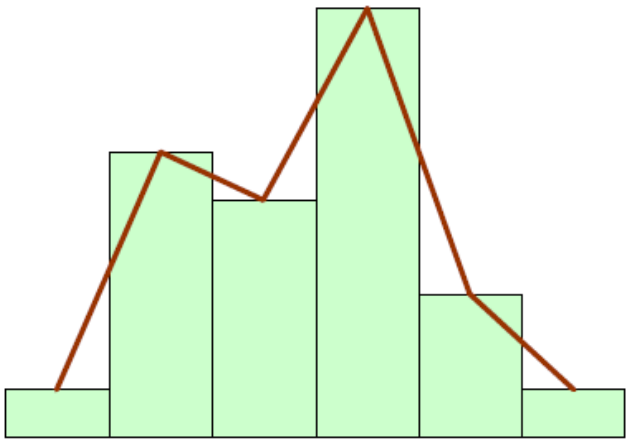


The first column represents all states with a percent Hispanic in their population between 0% and 4.99%. The height of the column shows how many states (27) have a percent Hispanic in this range.

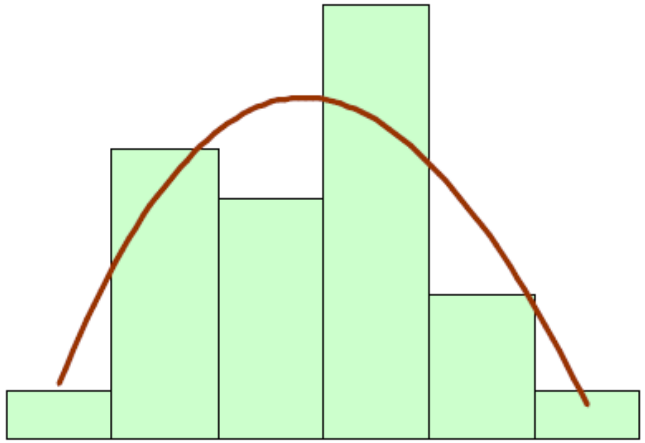
The last column represents all states with a percent Hispanic between 40% and 44.99%. There is only one such state: New Mexico, at 42.1% Hispanic.

Interpreting histograms

When describing a quantitative variable, we look for the overall pattern and for striking deviations from that pattern. We can describe the *overall* pattern of a histogram by its **shape, center,** and **spread.**



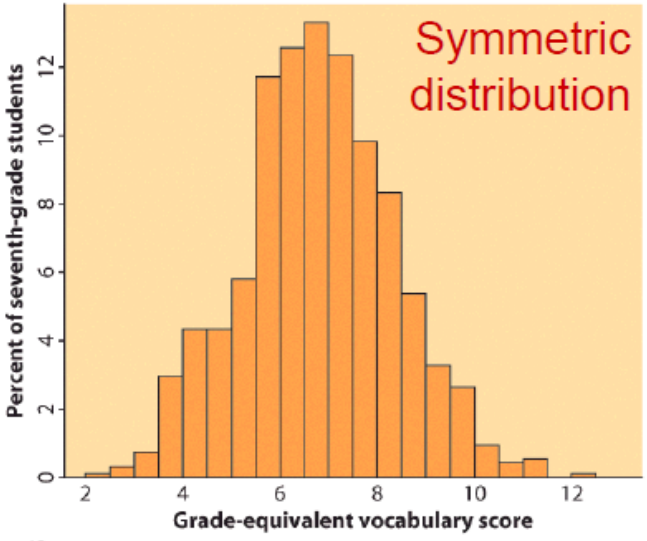
Histogram with a line connecting each column → too detailed

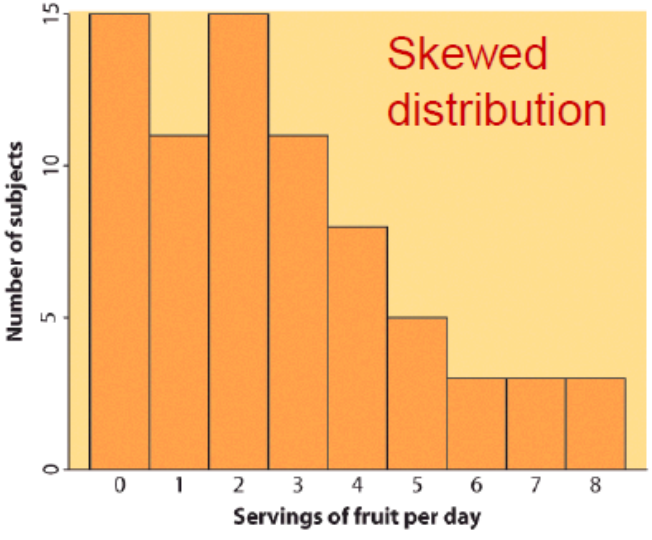


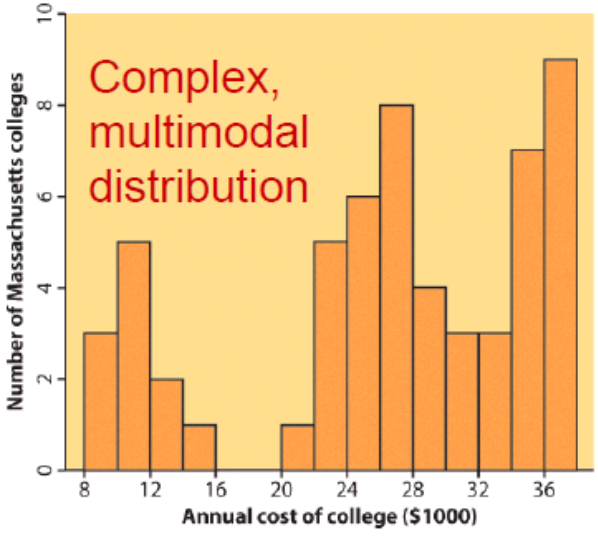
Histogram with a smoothed curve highlighting the overall pattern of the distribution

Most common distribution shapes

- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.







- Not all distributions have a simple overall shape, especially when there are few observations.

Measure of center: the **mean**

The mean or arithmetic average

To calculate the *average*, or *mean*, add all values, then divide by the number of individuals. It is the "center of mass."

Sum of heights is 1598.3
Divided by 25 women = 63.9 inches

58.2	64.0
59.5	64.5
60.7	64.1
60.9	64.8
61.9	65.2
61.9	65.7
62.2	66.2
62.2	66.7
62.4	67.1
62.9	67.8
63.9	68.9
63.1	69.6
63.9	

Mean

woman (i)	height (x)		woman (i)	height (x)
i = 1	x_1= 58.2		i = 14	x_{14}= 64.0
i = 2	x_2= 59.5		i = 15	x_{15}= 64.5
i = 3	x_3= 60.7		i = 16	x_{16}= 64.1
i = 4	x_4= 60.9		i = 17	x_{17}= 64.8
i = 5	x_5= 61.9		i = 18	x_{18}= 65.2
i = 6	x_6= 61.9		i = 19	x_{19}= 65.7
i = 7	x_7= 62.2		i = 20	x_{20}= 66.2
i = 8	x_8= 62.2		i = 21	x_{21}= 66.7
i = 9	x_9= 62.4		i = 22	x_{22}= 67.1
i = 10	x_{10}= 62.9		i = 23	x_{23}= 67.8
i = 11	x_{11}= 63.9		i = 24	x_{24}= 68.9
i = 12	x_{12}= 63.1		i = 25	x_{25}= 69.6
i = 13	x_{13}= 63.9		n=25	S=1598.3

Mathematical notation:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{1598.3}{25} = 63.9$$

Learn right away how to get the mean using your calculators.

Population vs Sample

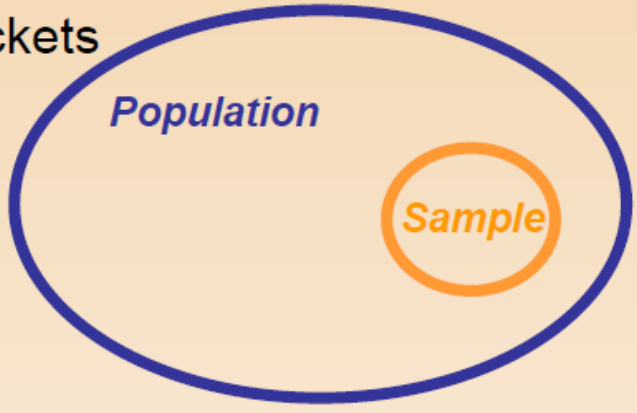
Population versus sample

- **Population:** The entire group of individuals in which we are interested but can't usually assess directly

  Example: All humans, all working-age people in California, all crickets

- **Sample:** The part of the population we actually examine and for which we do have data

  How well the sample represents the population depends on the sample design.



- A **parameter** is a number describing a characteristic of the population.

- A **statistic** is a number describing a characteristic of a sample.

Simple random samples

The **simple random sample (SRS)** is made of randomly selected individuals. Each individual in the population has the same probability of being in the sample. All possible samples of size *n* have the same chance of being drawn.

How to choose an SRS of size *n* from a population of size *N:*

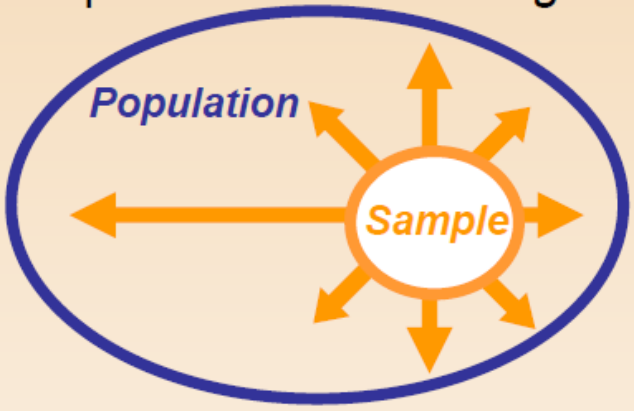
- **Label.** Give each member of the population a numerical label of the same length.

- **Table.** To choose an SRS, read from Table B successive groups of digits of the length you used as labels. Your sample contains the individuals whose labels you find in the table.

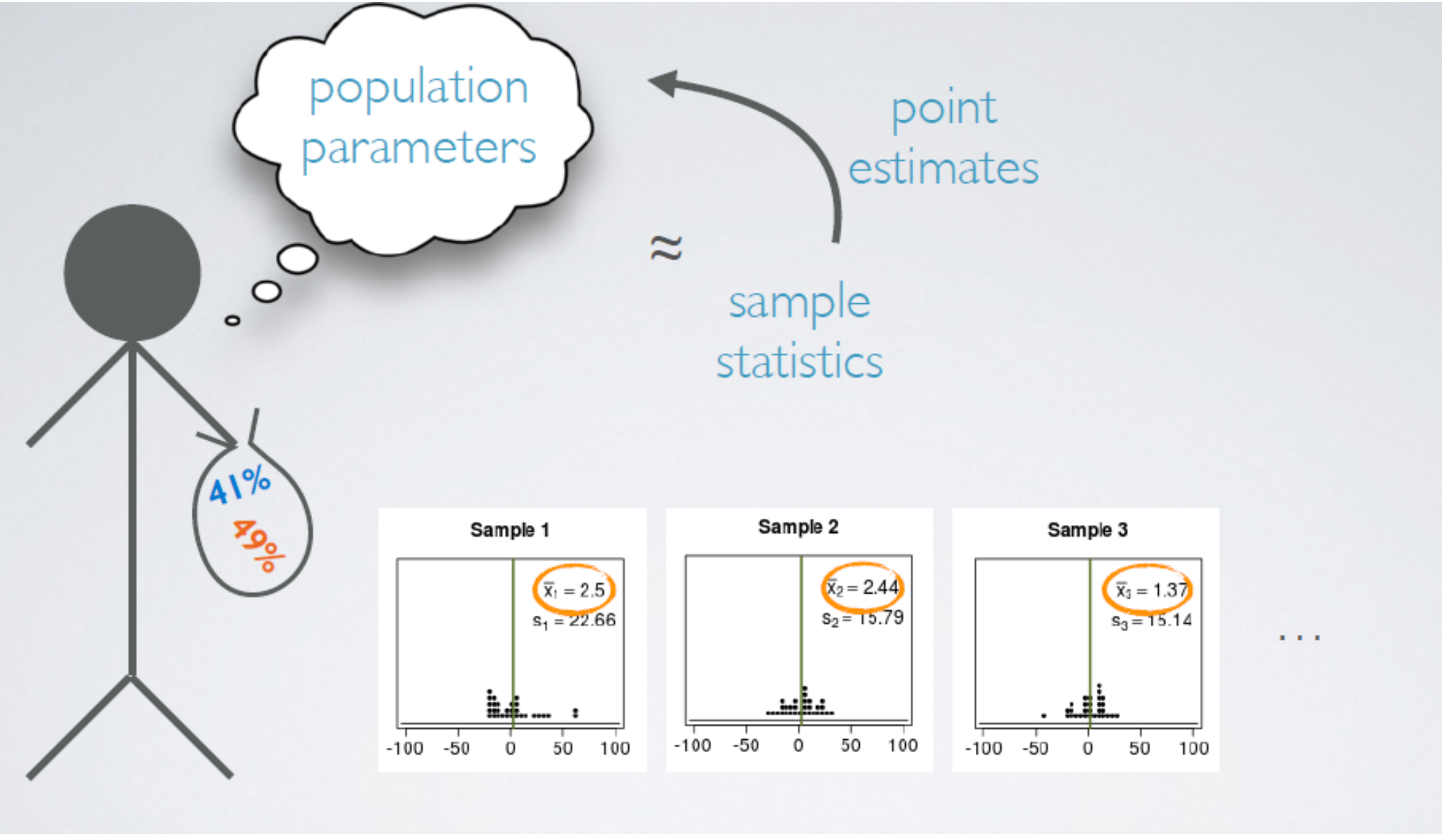
Learning about populations from samples

The techniques of inferential statistics allow us to draw inferences or conclusions about a population from a sample.

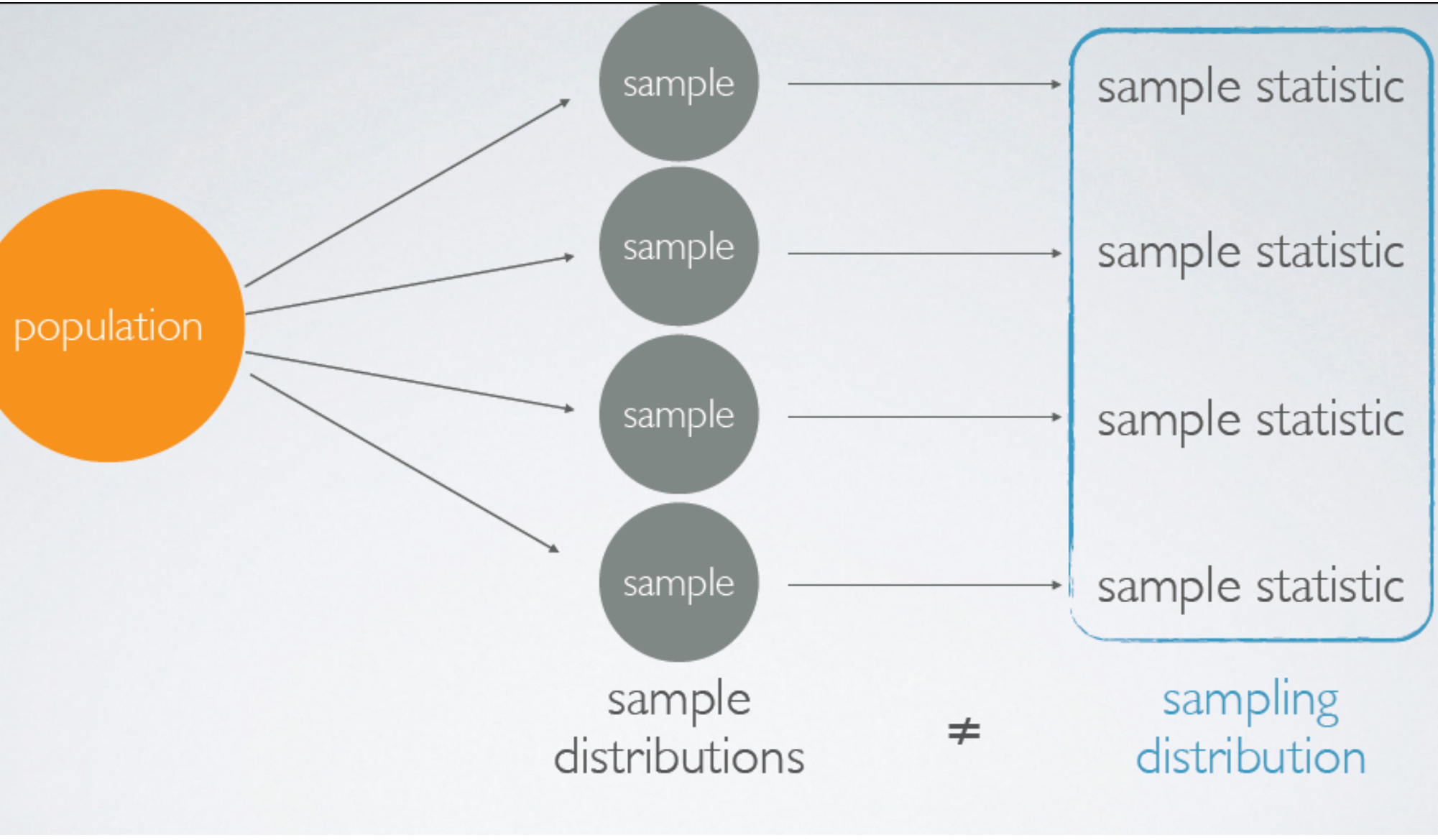
- Your estimate of the population is only as good as your sampling design → Work hard to eliminate biases.
- Your sample is only an estimate—and if you randomly sampled again, you would probably get a somewhat different result.
- The bigger the sample the better. *We'll get back to it in later chapters.*



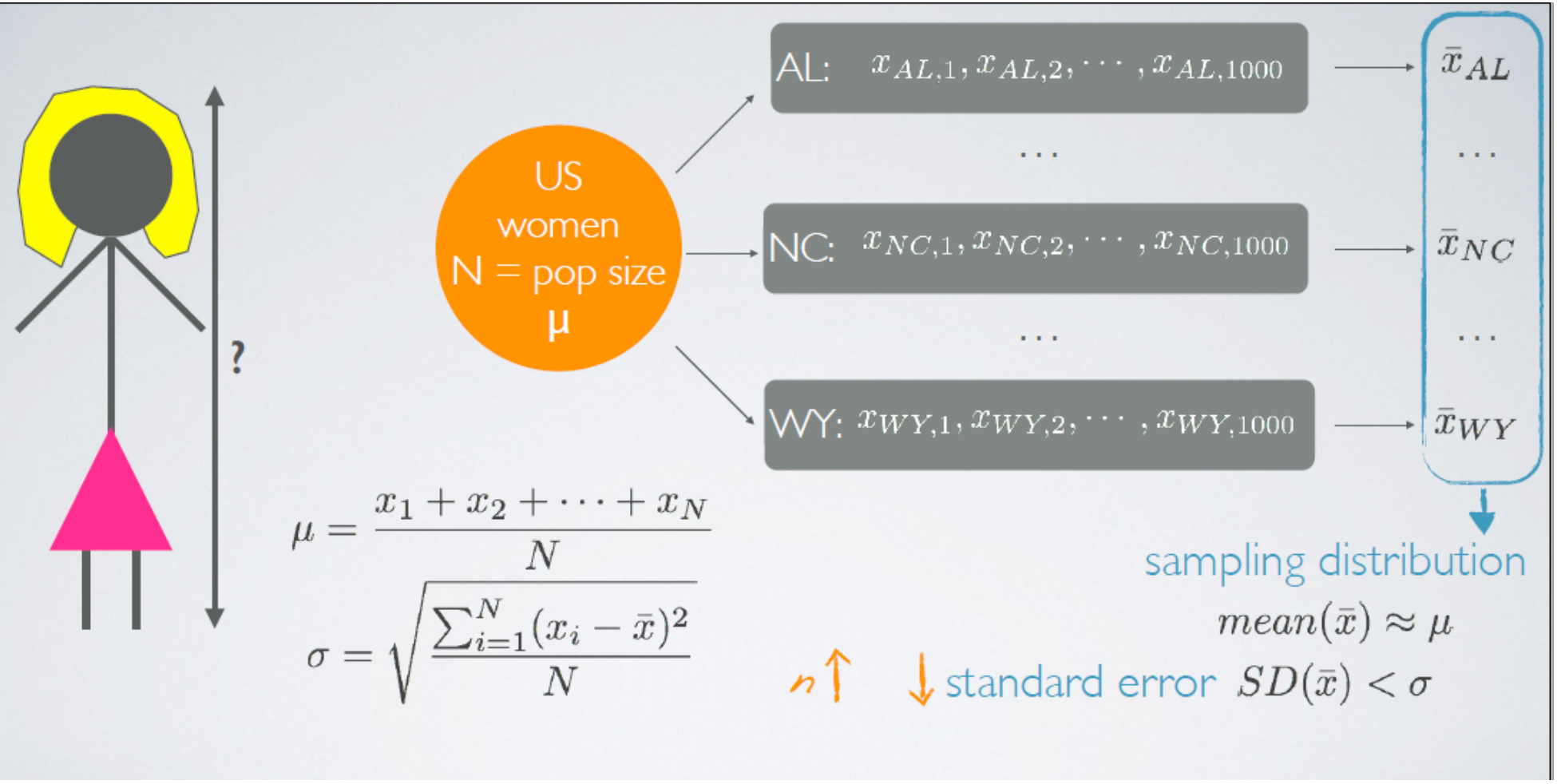
Sample statistics



Sample statistics



Sample statistics



$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

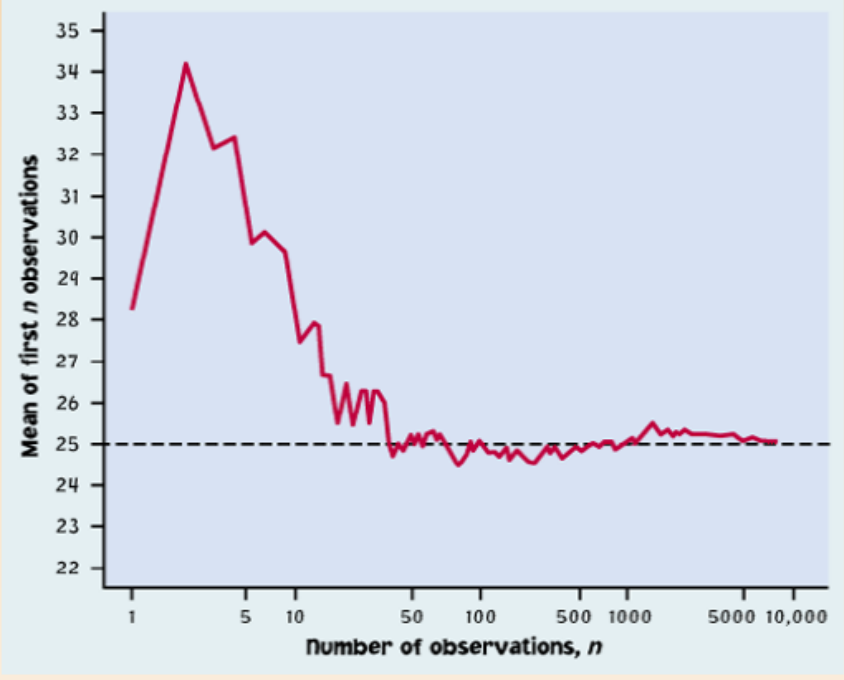
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

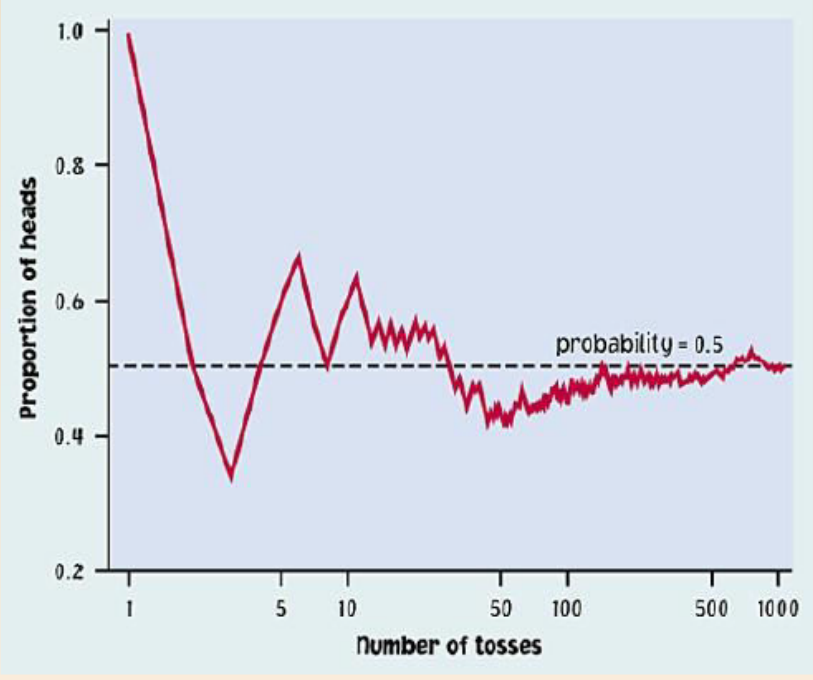
The law of large numbers

Law of large numbers: As the number of randomly-drawn observations (n) in a sample increases,

the mean of the sample ($\overline{x}$) gets closer and closer to the population mean μ (quantitative variable).

the sample proportion ($\hat{p}$) gets closer and closer to the population proportion p (categorical variable).





What is a sampling distribution?

The **sampling distribution of a statistic** is the distribution of all possible values taken by the statistic when all possible samples of a fixed size *n* are taken from the population. It is a theoretical idea—we do not actually build it.

The sampling distribution of a statistic is the **probability distribution** of that statistic.

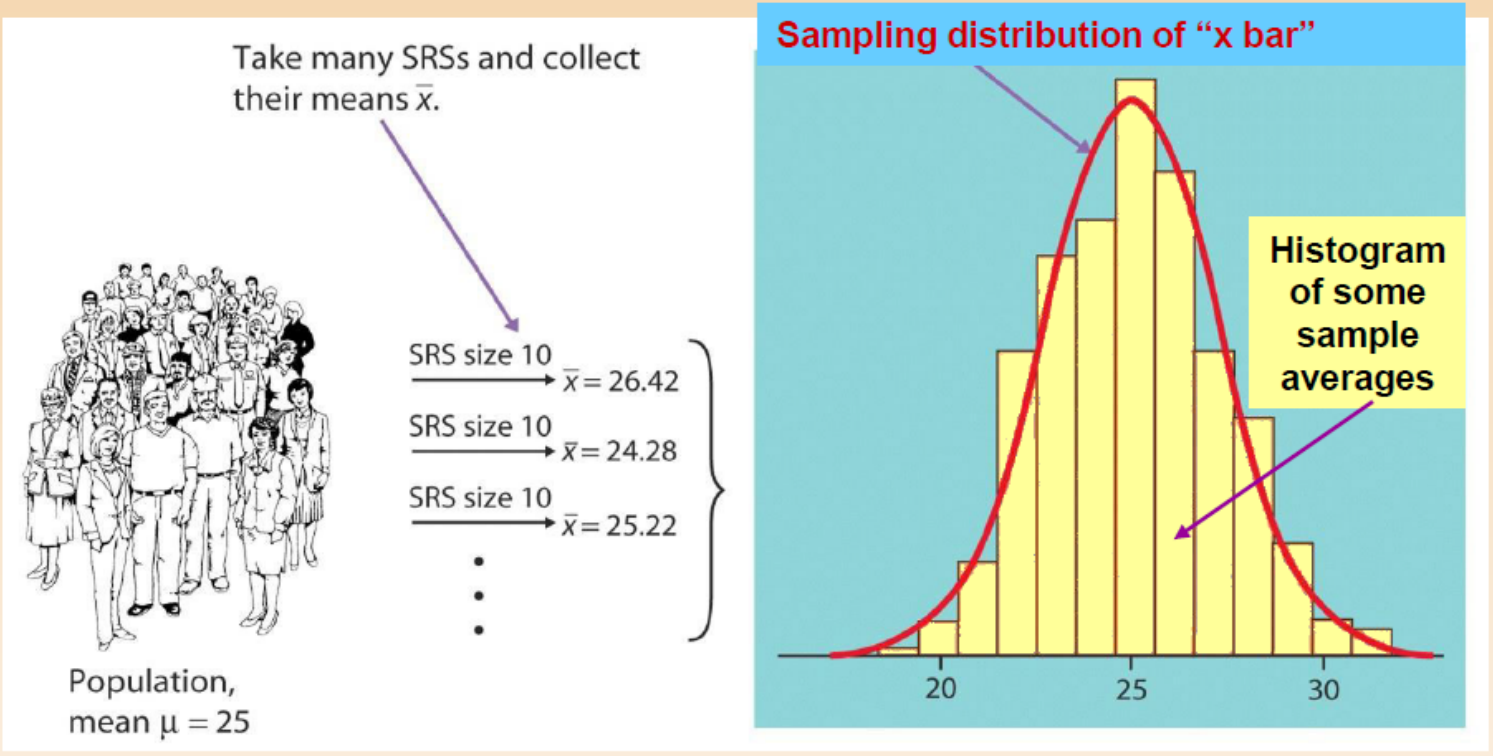
Note: When sampling randomly from a given population,

- *the law of large numbers describes what happens when the sample size* n *is gradually increased.*
- *The sampling distribution describes what happens when we take all possible random samples of a fixed size* n.

Sampling distribution of $\overline{x}$ *(the sample mean)*

We take many random samples of a given size *n* from a population with mean μ and standard deviation σ.

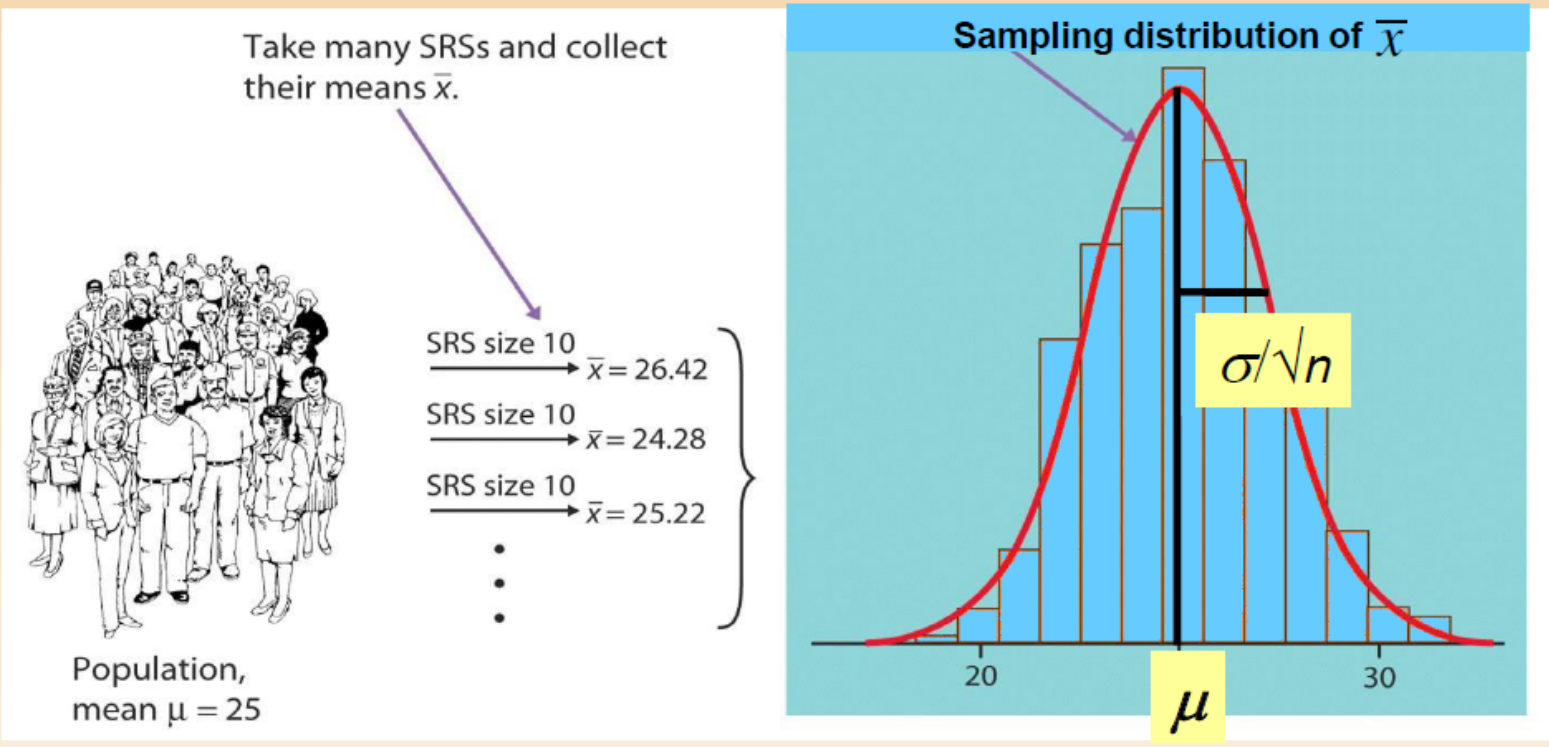
Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.



Sampling Distribution

For any population with mean μ and standard deviation σ:

- The **mean,** or center of the sampling distribution of $\overline{x}$, is equal to the population mean μ.
- The **standard deviation** of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size.



Sampling Distribution

- Mean of a sampling distribution of $\overline{x}$:

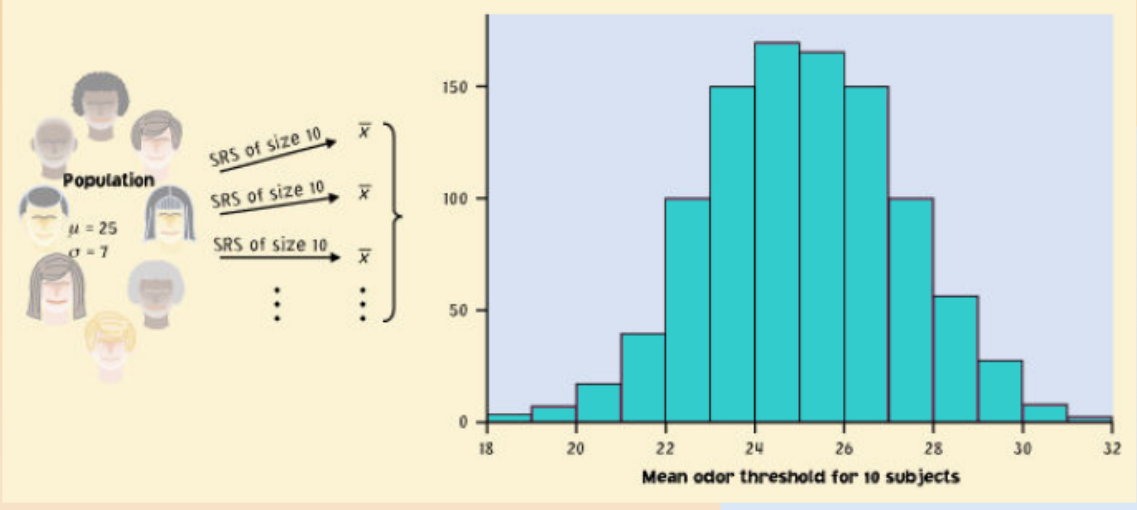
  There is no tendency for a sample mean to fall systematically above or below μ, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution of $\overline{x}$ is an **unbiased estimate** of the population mean μ —it will be "correct on average" in many samples.

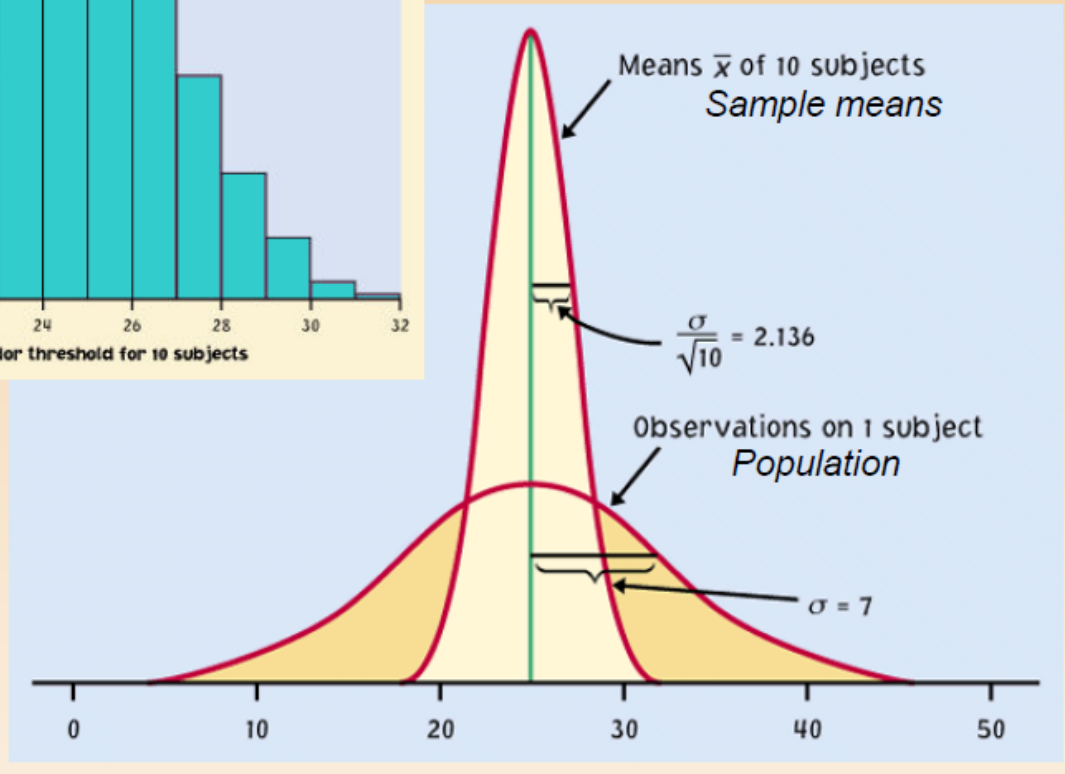
- Standard deviation of a sampling distribution of $\overline{x}$:

  The standard deviation of the sampling distribution measures how much the sample statistic $\overline{x}$ varies from sample to sample. It is smaller than the standard deviation of the population by a factor of $\sqrt{n}$. ➔ **Averages are less variable than individual observations.**

For normally distributed populations

When a variable in a population is normally distributed, then the sampling distribution of $\overline{x}$ for all possible samples of size *n* is also normally distributed.





If the population is $N(\mu,\sigma)$, then the sample means distribution is $N(\mu,\sigma/\sqrt{n})$.

Example

IQ scores: population vs. sample

In a large population of adults, the mean IQ is 112 with standard deviation 20. Suppose 200 adults are randomly selected for a market research campaign.

- The distribution of the sample mean IQ is

A) exactly normal, mean 112, standard deviation 20.

B) approximately normal, mean 112, standard deviation 20.

C) approximately normal, mean 112 , standard deviation 1.414.

D) approximately normal, mean 112, standard deviation 0.1.

C) approximately normal, mean 112, standard deviation 1.414.

Population distribution: $N(\mu = 112; \sigma = 20)$

Sampling distribution for $n = 200$ is $N(\mu = 112; \sigma/\sqrt{n} = 1.414)$

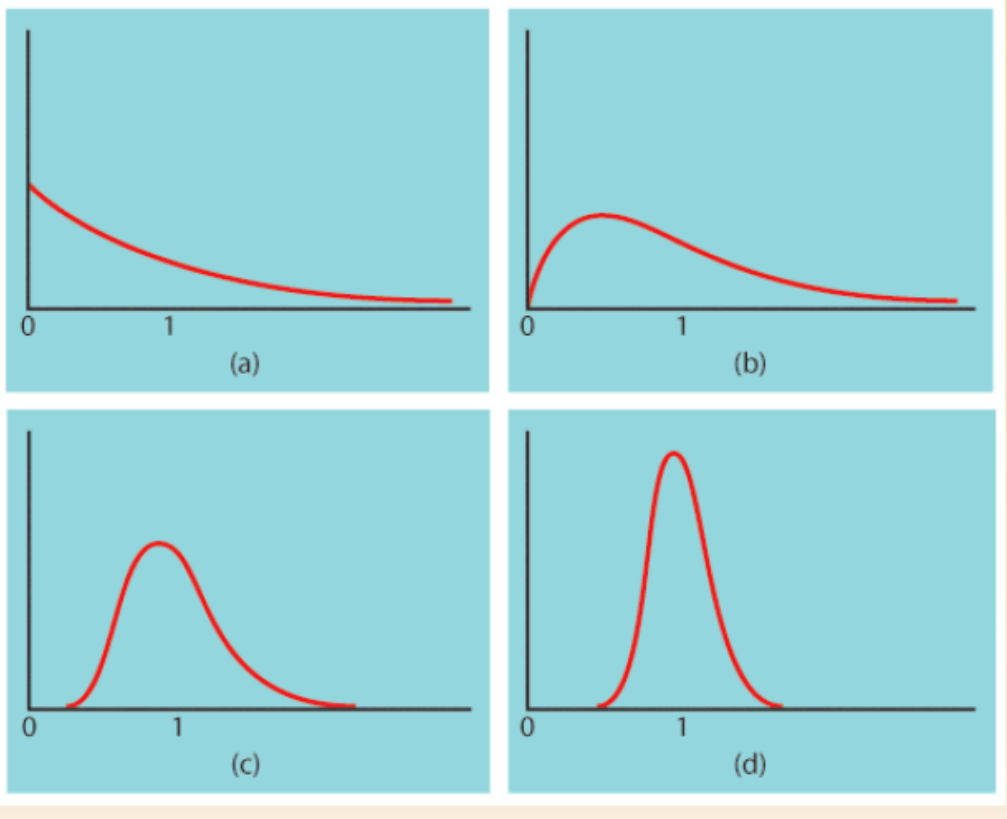
Practical note

- Large samples are not always attainable.
  - Sometimes the cost, difficulty, or preciousness of what is studied limits drastically any possible sample size.
  - Blood samples/biopsies: no more than a handful of repetitions acceptable. Often we even make do with just one.
  - Opinion polls have a limited sample size due to time and cost of operation. During election times, though, sample sizes are increased for better accuracy.
- Not all variables are normally distributed.
  - Income is typically strongly skewed for example.
  - Is $\overline{x}$ still a good estimator of μ then?

The central limit theorem

Central Limit Theorem: When randomly sampling from any population with mean μ and standard deviation σ, **when *n* is large enough,** the sampling distribution of $\bar{x}$ is approximately normal: $N(\mu, \sigma/\sqrt{n})$.

Population with strongly skewed distribution



Sampling distribution of $\bar{x}$ for n = 2 observations

Sampling distribution of $\bar{x}$ for n = 10 observations

Sampling distribution of $\bar{x}$ for n = 25 observations

CLT

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

shape center spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

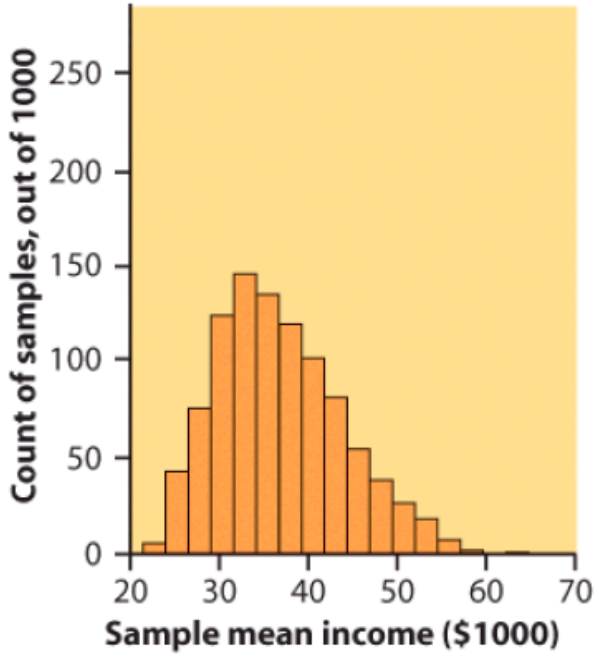
Example

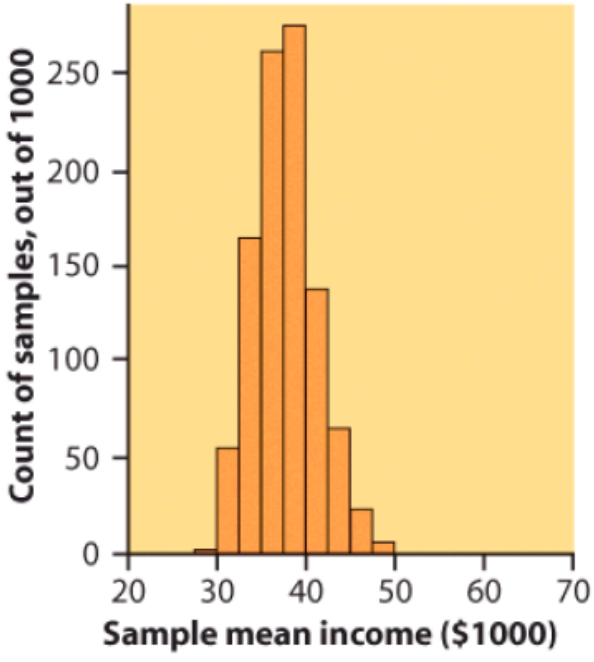
Income distribution

Let's consider the very large database of individual incomes from the Bureau of Labor Statistics as our population. It is strongly right-skewed.

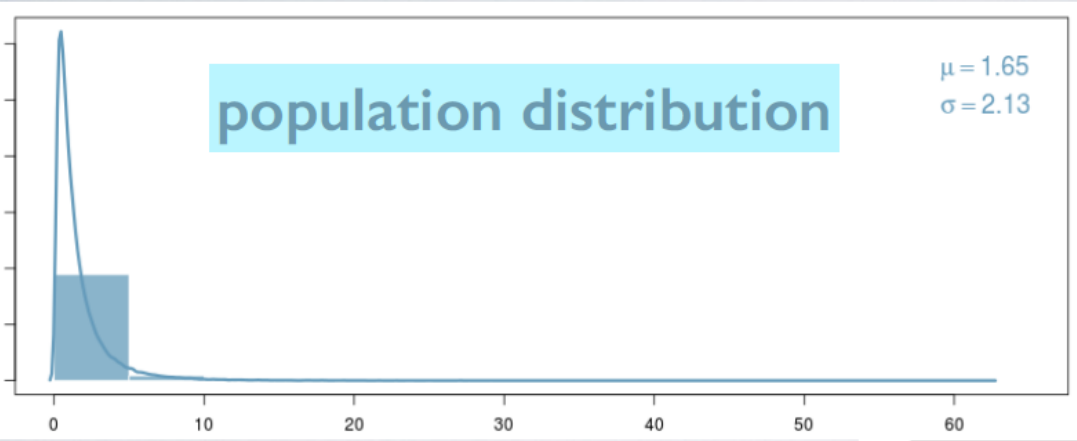
- We take 1000 SRSs of 100 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.
- We also take 1000 SRSs of 25 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.

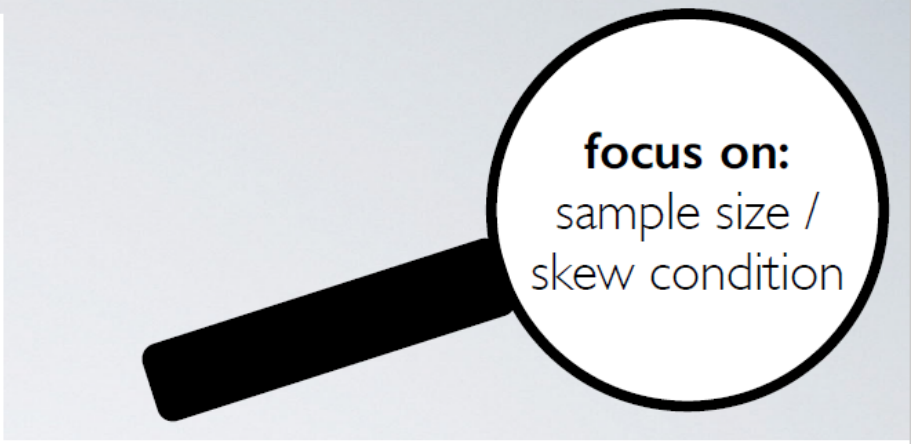
Which histogram corresponds to the samples of size 100? 25?

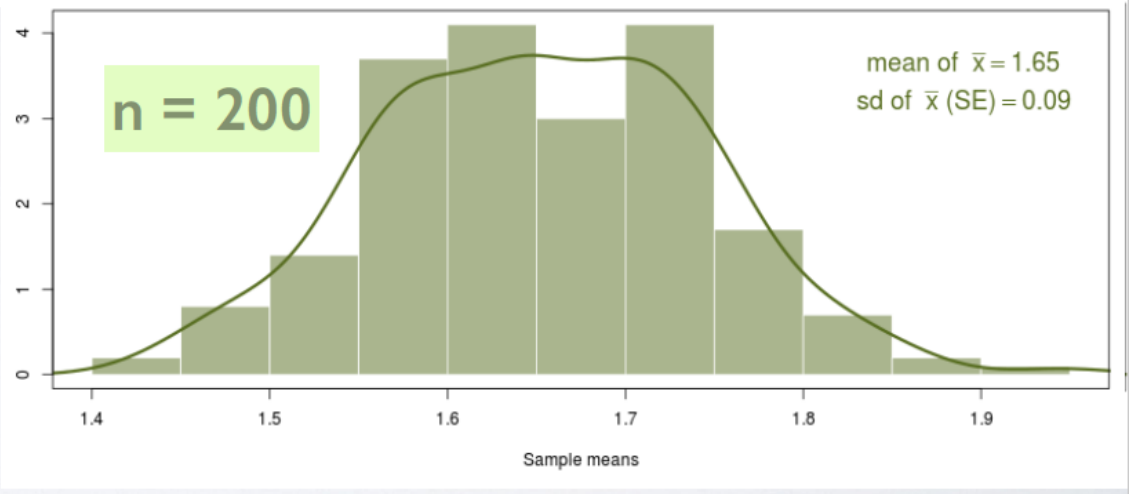




Example


population distribution
μ = 1.65
σ = 2.13
0
10
20
30
40
50
60


focus on:
sample size /
skew condition

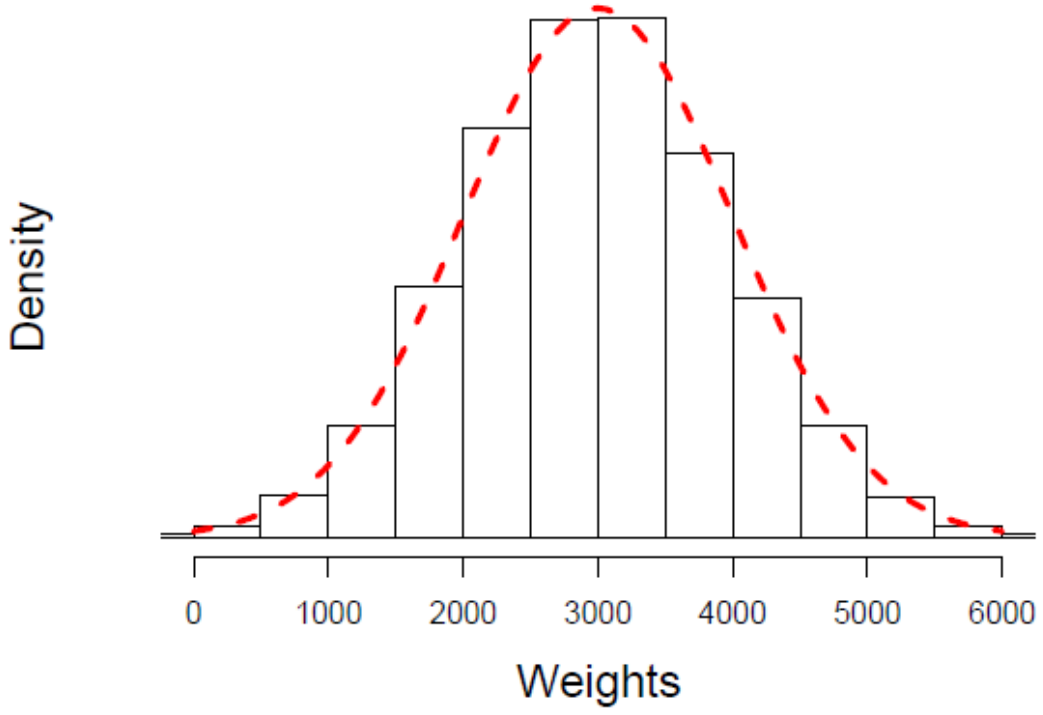

n = 200
mean of x̄ = 1.65
sd of x̄ (SE) = 0.09
0
1
2
3
4
1.4
1.5
1.6
1.7
1.8
1.9
Sample means

Z-Scores

- We use z-score to convert a "general normal" to a "standard normal" random variable.
- $Z = \frac{X-\mu}{\sigma}$ is a 'standard normal" random variable.
- For any real number t, We can find $P(Z < t)$ or $P(Z > t)$ using z-score table (see z score tables).
- $P(Z < t)$ is the same as CDF of a standard normal random variable at point t.
- $P(Z > t) = 1 - P(Z < t)$

Z-Scores Example

- Birthweights (in grams) histogram:



$$\mu = 3000$$

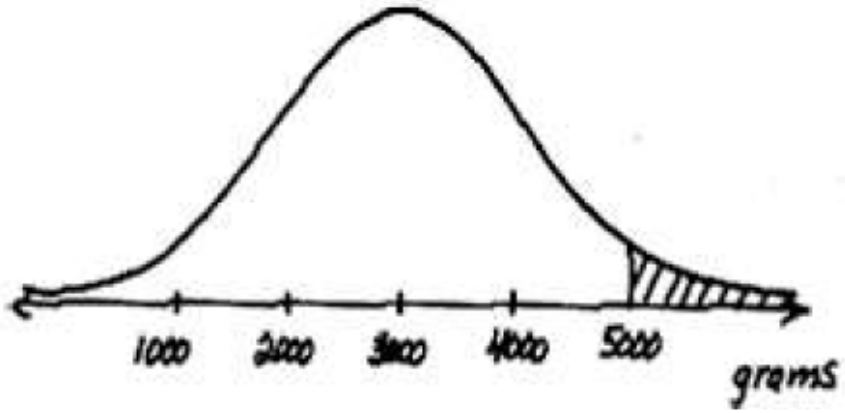
$$\sigma = 1000$$

$$X = \textit{birthweight}$$

$$Z = \frac{X - \mu}{\sigma}$$

Z-Scores Example

What is the probability of an infant weighing more than 5000g?



$$
\begin{aligned}
P(X > 5000) &= P\left(\frac{X-\mu}{\sigma} > \frac{5000-3000}{1000}\right) \\
&= P(Z > 2) \\
&= 0.0228
\end{aligned}
$$

- Using the table: $P(Z < 2) = 0.97725$.
- Therefore $P(Z > 2) = 1 - 0.97725 = 0.02275 \approx 0.0228$

Z-Scores Example

What is the probability of an infant weighing less than 3500g?

$$\begin{aligned} P(X < 3500) &= P(\frac{X-\mu}{\sigma} < \frac{3500-3000}{1000}) \\ &= P(Z < 0.5) \\ &= 0.6915 \end{aligned}$$

Z-Scores Example

What is the probability of an infant weighing between 2500 and 4000g?

$$
\begin{aligned}
P(2500 < X < 4000) &= P\left(\frac{2500 - 3000}{1000} < \frac{X - \mu}{\sigma} < \frac{4000 - 3000}{1000}\right) \\
&= P(-0.5 < Z < 1) \\
&= 1 - P(Z > 1) - P(Z < -0.5) \\
&= 1 - 0.1587 - 0.3085 \\
&= 0.5328
\end{aligned}
$$

Example

Application

Hypokalemia is diagnosed when blood potassium levels are low, below 3.5mEq/dl. Let's assume that we know a patient whose measured potassium levels vary daily according to a normal distribution $N(\mu = 3.8, \sigma = 0.2)$.

If only one measurement is made, what's the probability that this patient will be misdiagnosed hypokalemic?

$$z = \frac{(x - \mu)}{\sigma} = \frac{3.5 - 3.8}{0.2}$$

$z = -1.5$, $P(z < -1.5) = 0.0668 \approx 7\%$

If instead measurements are taken on four separate days, what is the probability of such a misdiagnosis?

$$z = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} = \frac{3.5 - 3.8}{0.2/\sqrt{4}}$$

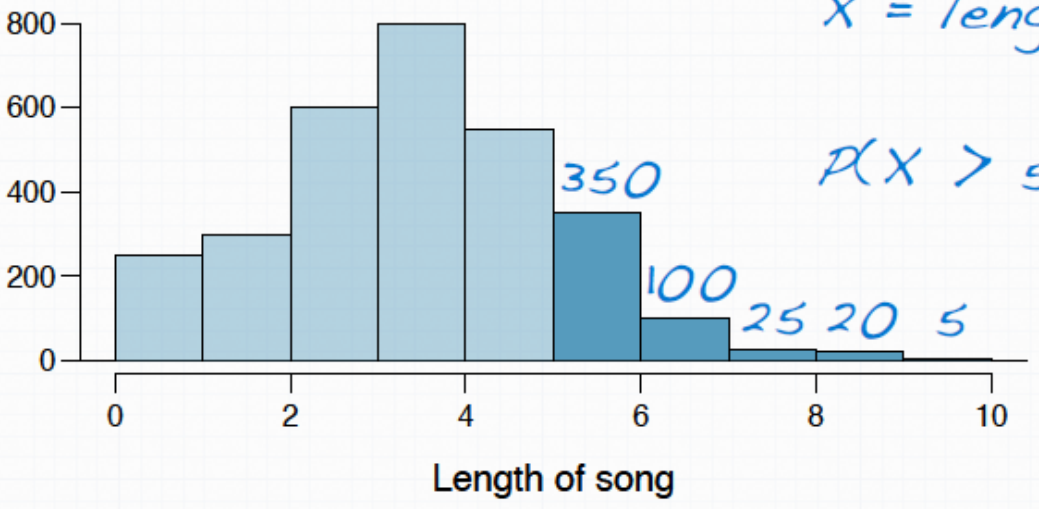
$z = -3$, $P(z < -1.5) = 0.0013 \approx 0.1\%$

Note:

Make sure to standardize (z) using the standard deviation for the sampling distribution.

Example

Suppose my iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than 5 minutes.



X = length of one song

$$P(X > 5) = \frac{350 + 100 + 25 + 20 + 5}{3000}$$

$$= 500 / 3000$$

$$\approx 0.17$$

Example

I'm about to take a trip to visit my parents and the drive is 6 hours. I make a random playlist of 100 songs. What is the probability that my playlist lasts the entire drive?

6 hours = 360 minutes

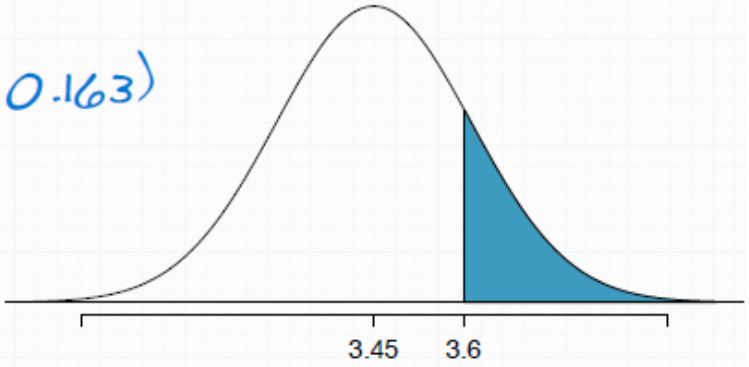
$P(X_1 + X_2 + \ldots + X_{100} > 360 \text{ min}) = ?$

$P(\bar{X} > 3.6) = ?$

$\bar{X} \sim N(\text{mean} = \mu = 3.45, SE = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{100}} = 0.163)$

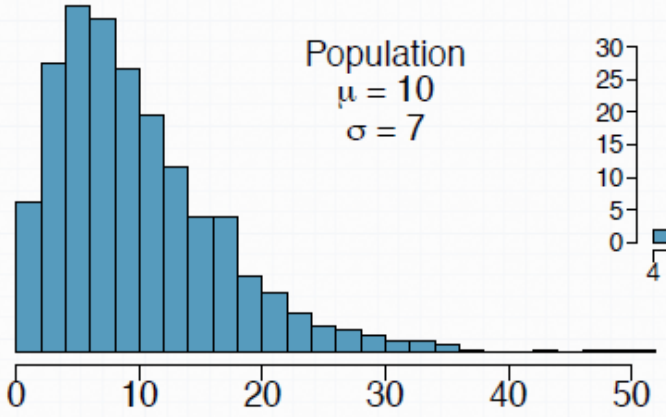
$Z = \frac{3.6 - 3.45}{0.163} = 0.92$

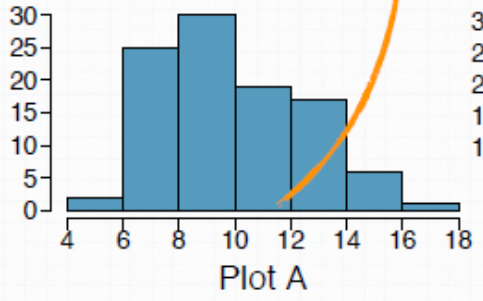
$P(Z > 0.92) = 0.179$

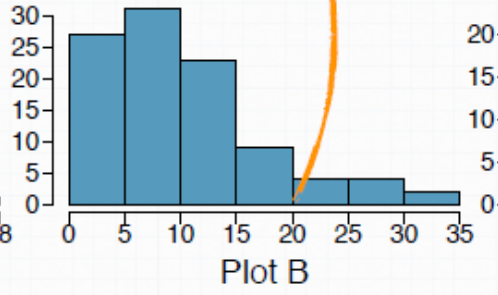


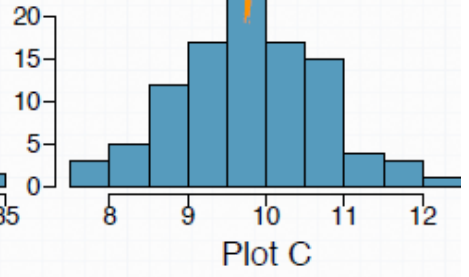
Example

Four plots: Determine which plot (A, B, or C) is which.
(1) The distribution for a population ($\mu = 10, \sigma = 7$),
(2) a single random sample of 100 observations from this population,
(3) a distribution of 100 sample means from random samples with size 7, and
(4) a distribution of 100 sample means from random samples with size 49.





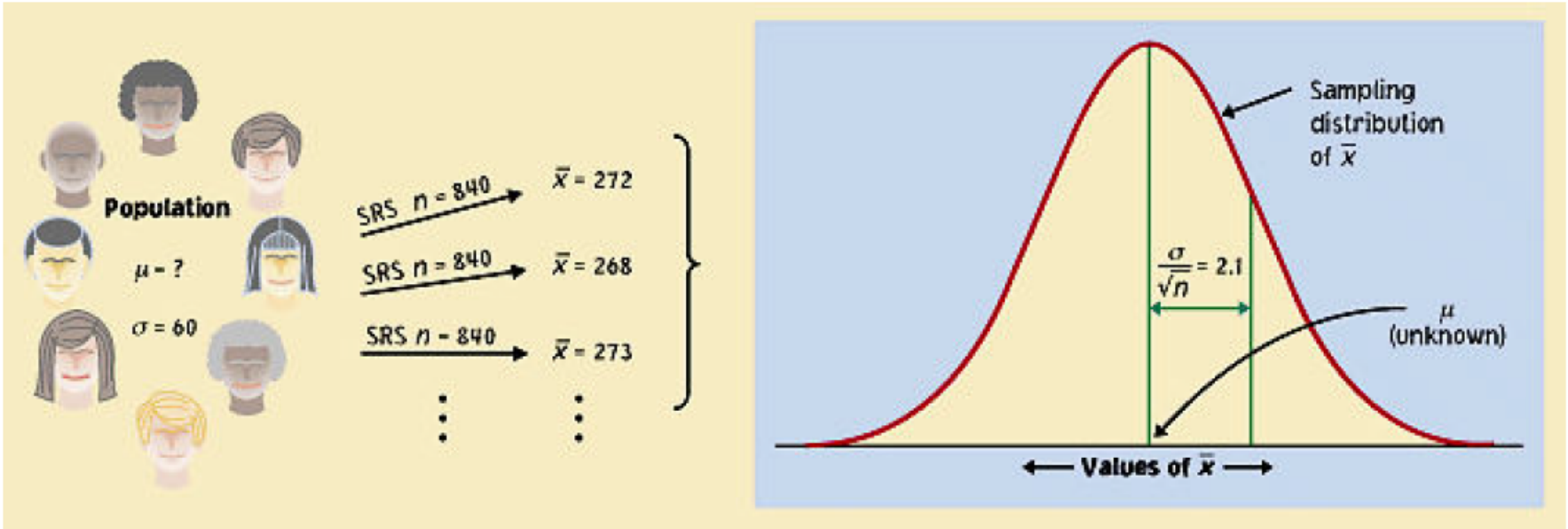




Estimating with confidence

Although the sample mean, $\bar{x}$, is a unique number for any particular sample, if you pick a different sample, you will probably get a different sample mean.

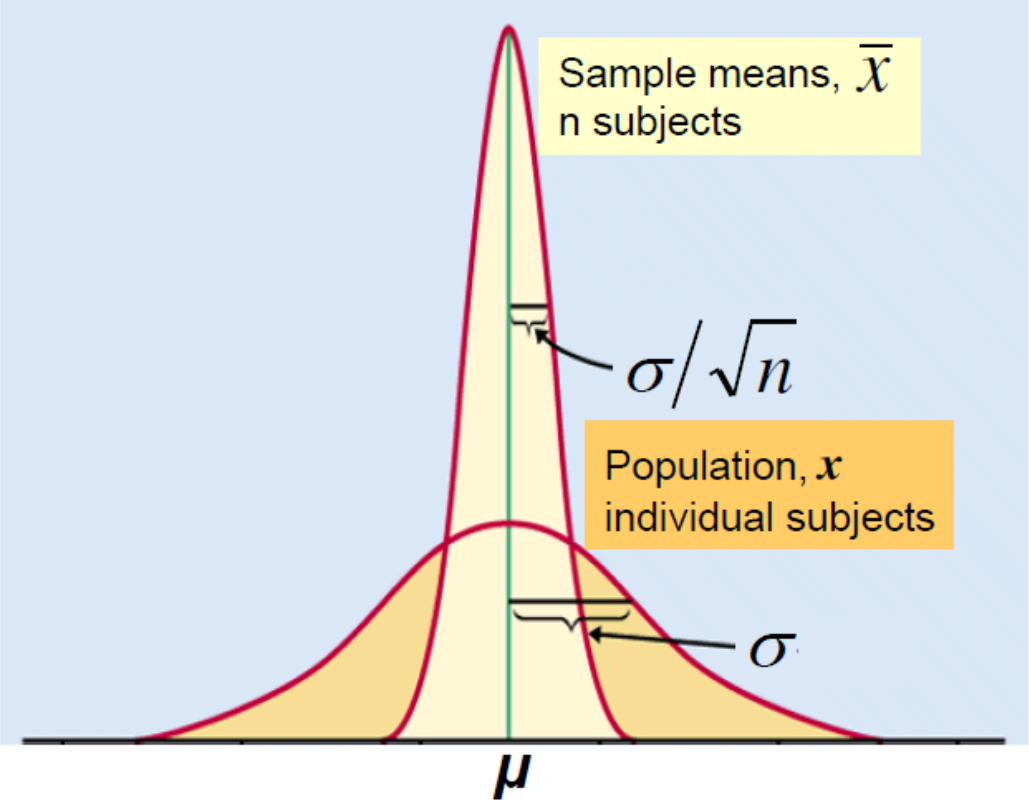
In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, □.



Confidence Intervals

But the sample distribution is narrower than the population distribution, by a factor of $\sqrt{n}$.

Thus, the estimates $\bar{x}$ gained from our samples are always relatively close to the population parameter μ.

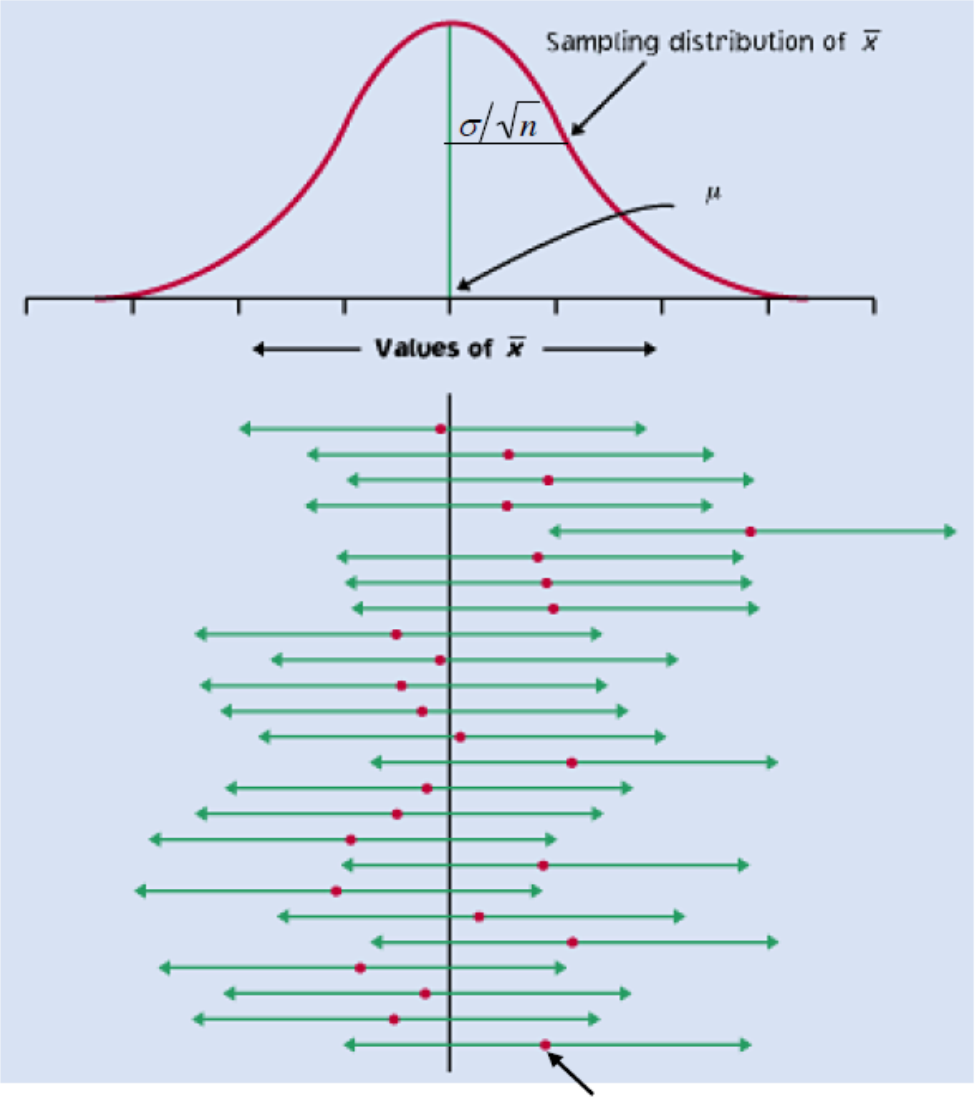


If the population is normally distributed $N(\mu,\sigma)$, so will the sampling distribution $N(\mu,\sigma/\sqrt{n})$.

Confidence Intervals

95% of all sample means will be within roughly 2 standard deviations ($2^*\sigma/\sqrt{n}$) of the population parameter *m*.

Because distances are symmetrical, this implies that **the population parameter μ must be within roughly 2 standard deviations from the sample average $\overline{x}$, in 95% of all samples.**



This reasoning is the essence of statistical inference.

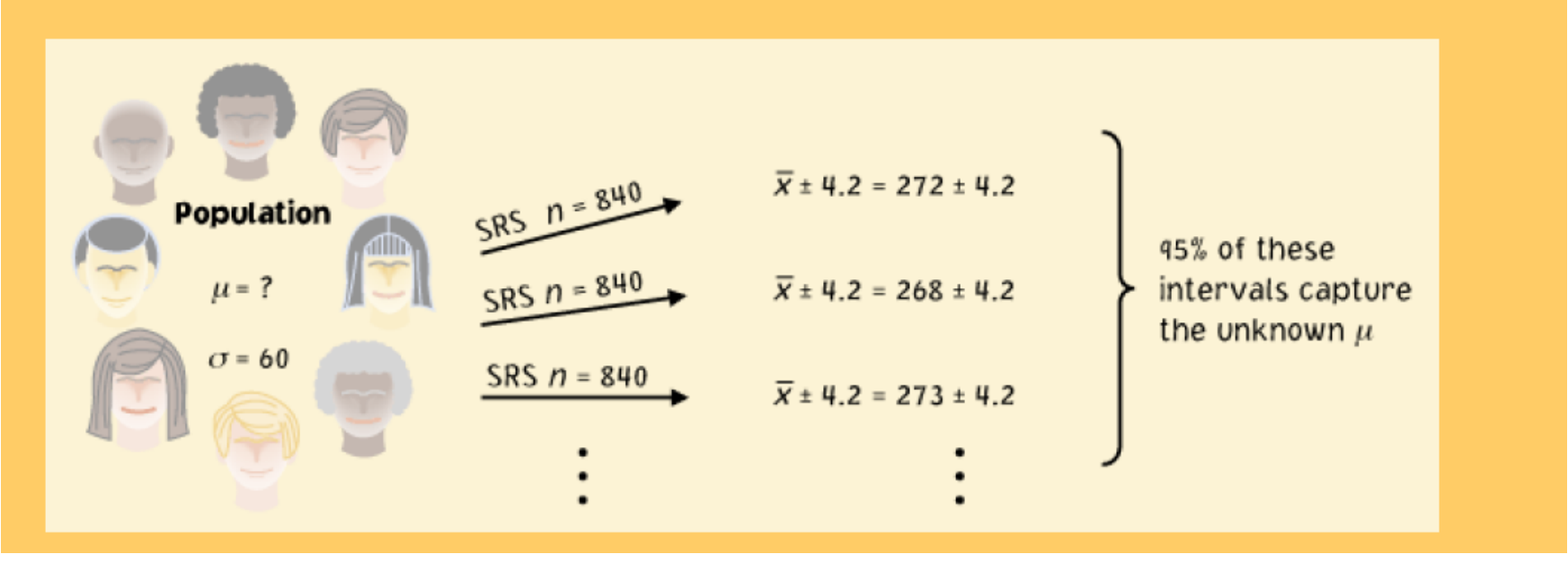
Confidence interval

A level C confidence interval for a parameter has two parts:

- An interval calculated from the data, usually of the form

estimate ±margin of error

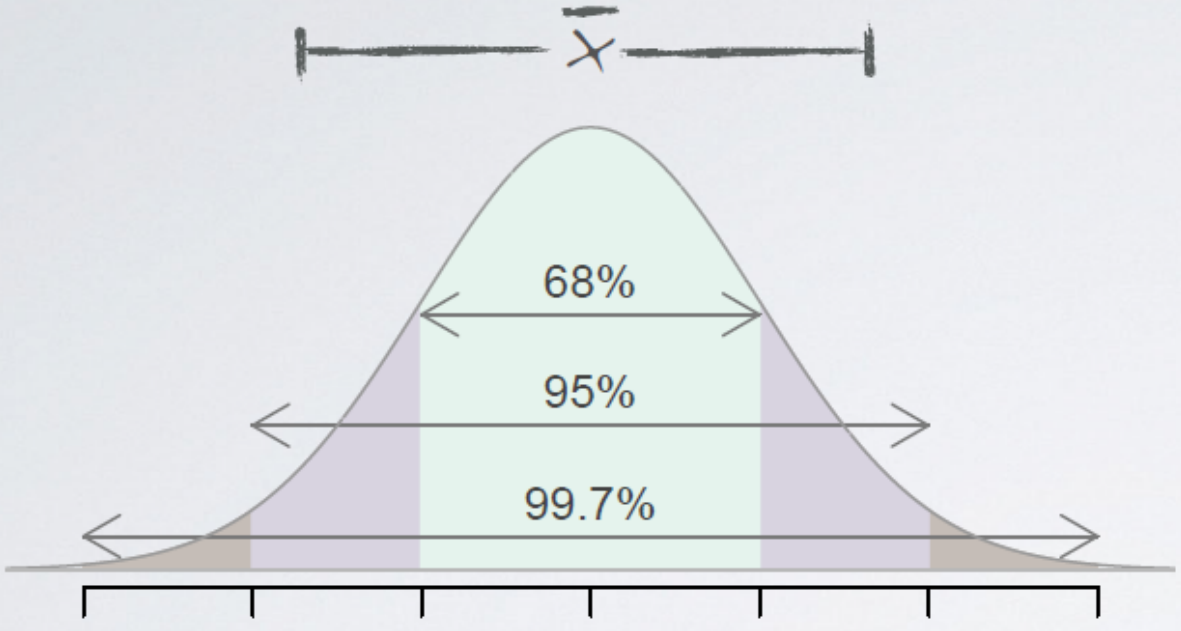
- A confidence level C, which gives the probability that the interval will capture the true parameter value in repeated samples, or the success rate for the method.



Confidence Intervals

Central Limit Theorem (CLT):

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

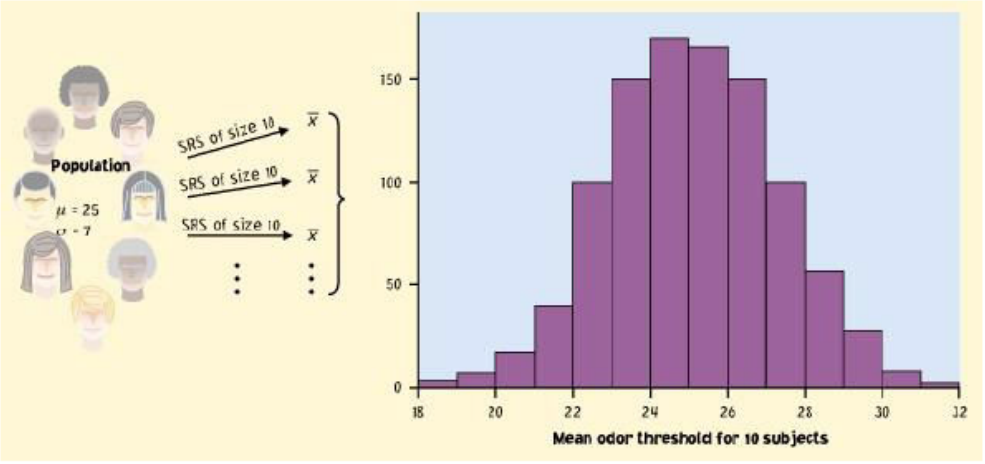


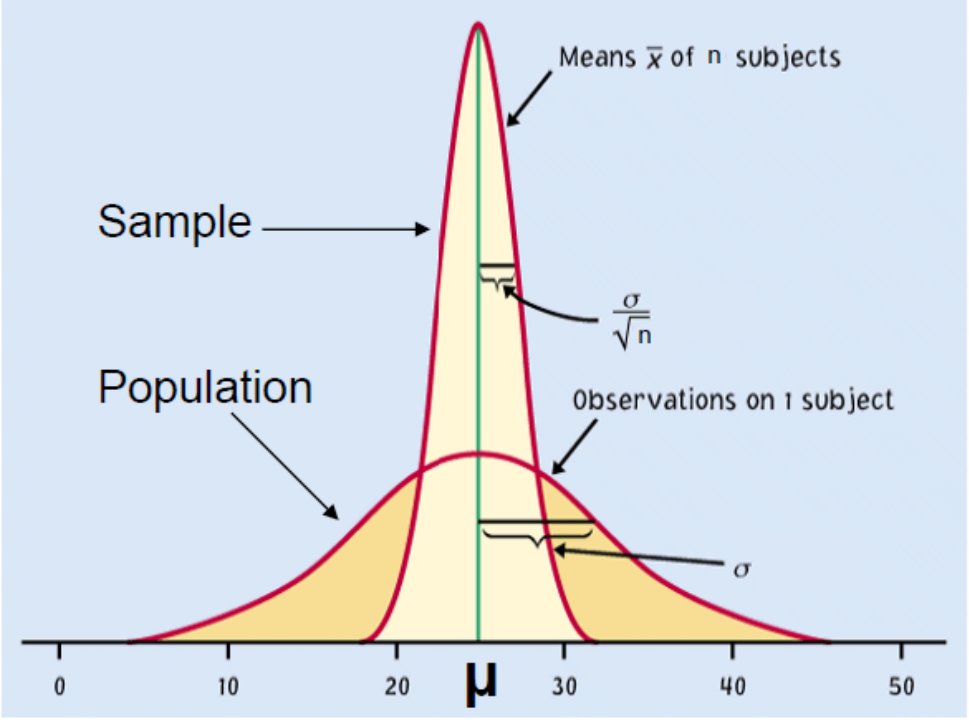
approximate 95% CI: $\bar{x} \pm 2SE$

margin of error (ME)

Implications

We don't need to take lots of random samples to "rebuild" the sampling distribution and find μ at its center.





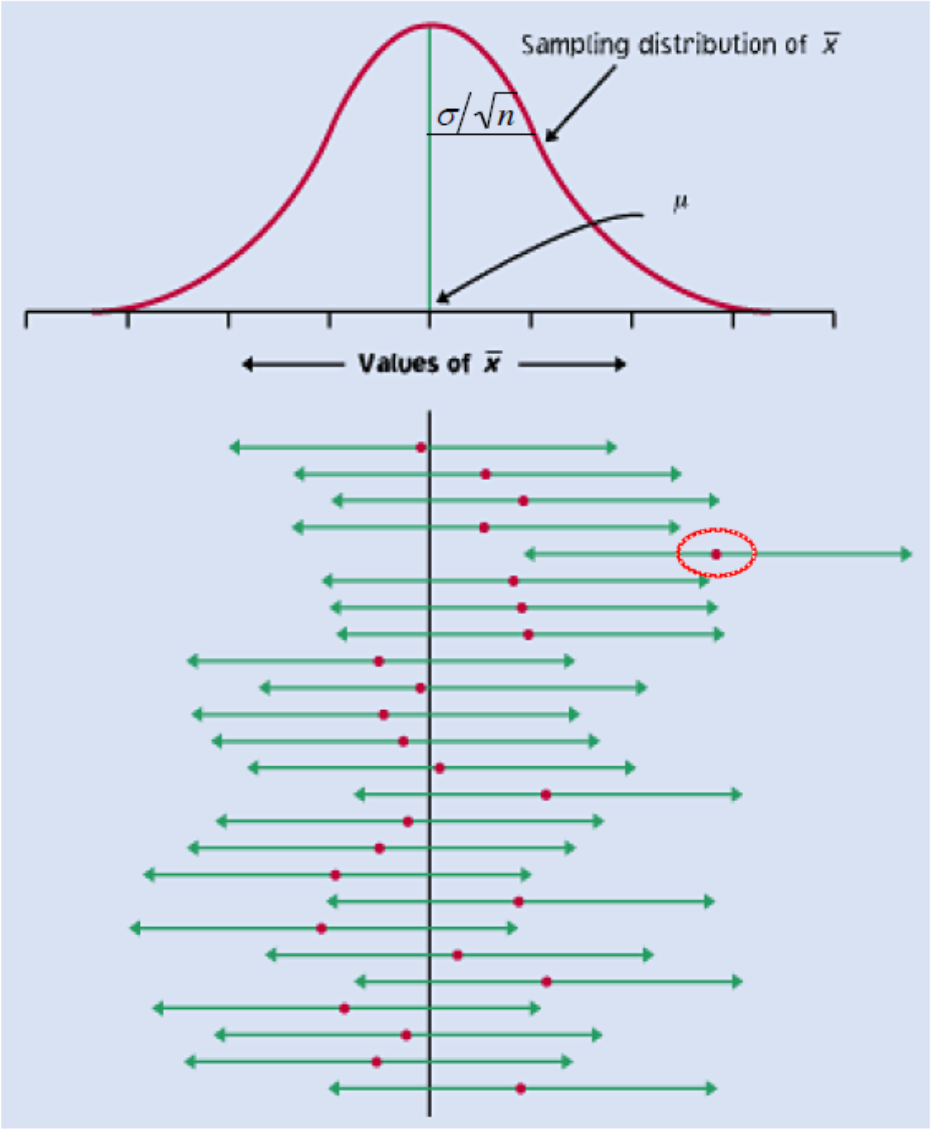
All we need is one SRS of size *n*, and relying on the properties of the sample means distribution to infer the population mean *μ*.

Confidence Intervals

Reworded

With 95% confidence, we can say that μ should be within roughly 2 standard deviations ($2^*\sigma/\sqrt{n}$) from our sample mean $\overline{x}$ bar.

- In 95% of all possible samples of this size n, μ will indeed fall in our confidence interval.
- In only 5% of samples would $\overline{x}$ be farther from μ.



Interpreting a confidence interval for a mean

A **confidence interval** can be expressed as:

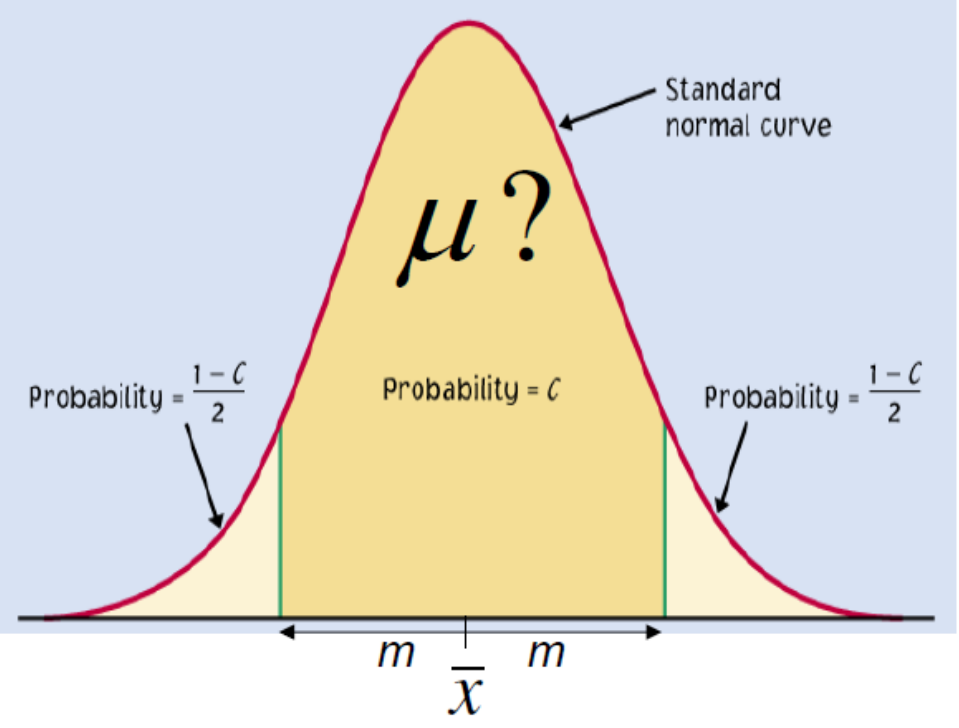
- $\overline{x} \pm m$
  m is called the **margin of error**

- Two endpoints of an interval:
  m possibly within $(\overline{x} - m)$ to $(\overline{x} + m)$

  Example: 114 to 126

A **confidence level** ***C*** (in %) indicates the success rate of the method that produces the interval.

It represents the area under the normal curve within $\pm m$ of the center of the curve.



Varying confidence levels

Confidence intervals contain the population mean μ in C% of samples.

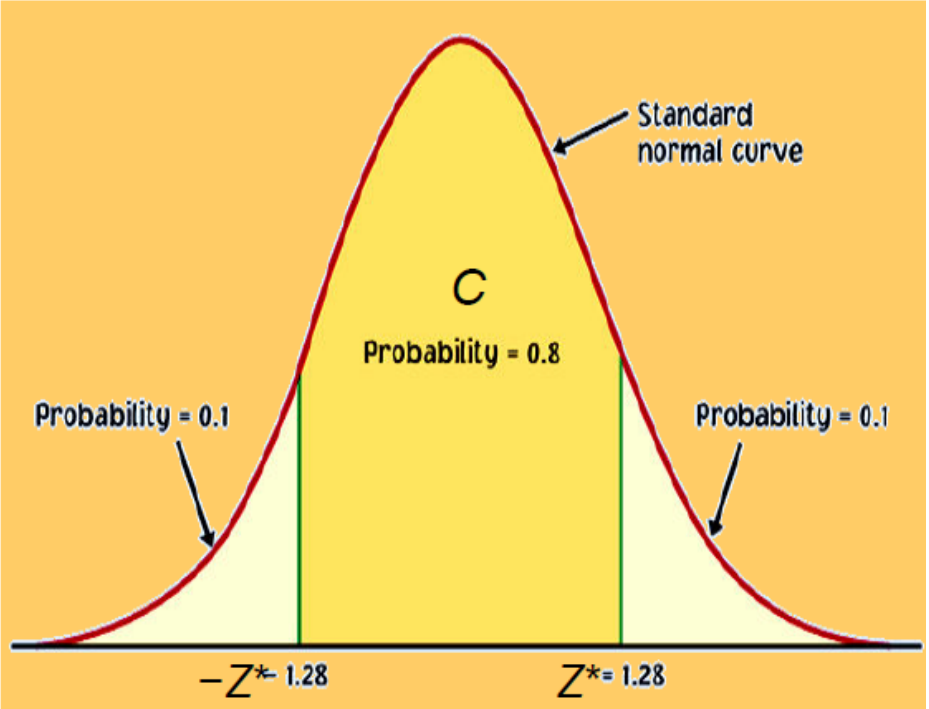
Different areas under the curve give different confidence levels C.

Practical use of z: z^*

- z^* is related to the chosen confidence level C.
- C is the area under the standard normal curve between $-z^*$ and z^*.

The confidence interval is thus:

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



Example: For an 80% confidence level C, 80% of the normal curve's area is contained in the interval.

Confidence Intervals

Confidence interval for a population mean: Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution).

$$\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$$

Conditions for this confidence interval:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** n ≥ 30, larger if the population distribution is very skewed.

How do we find specific z^* values?

We can use a table of z/t values (Table C). For a particular confidence level C, the appropriate z^* value is just above it.

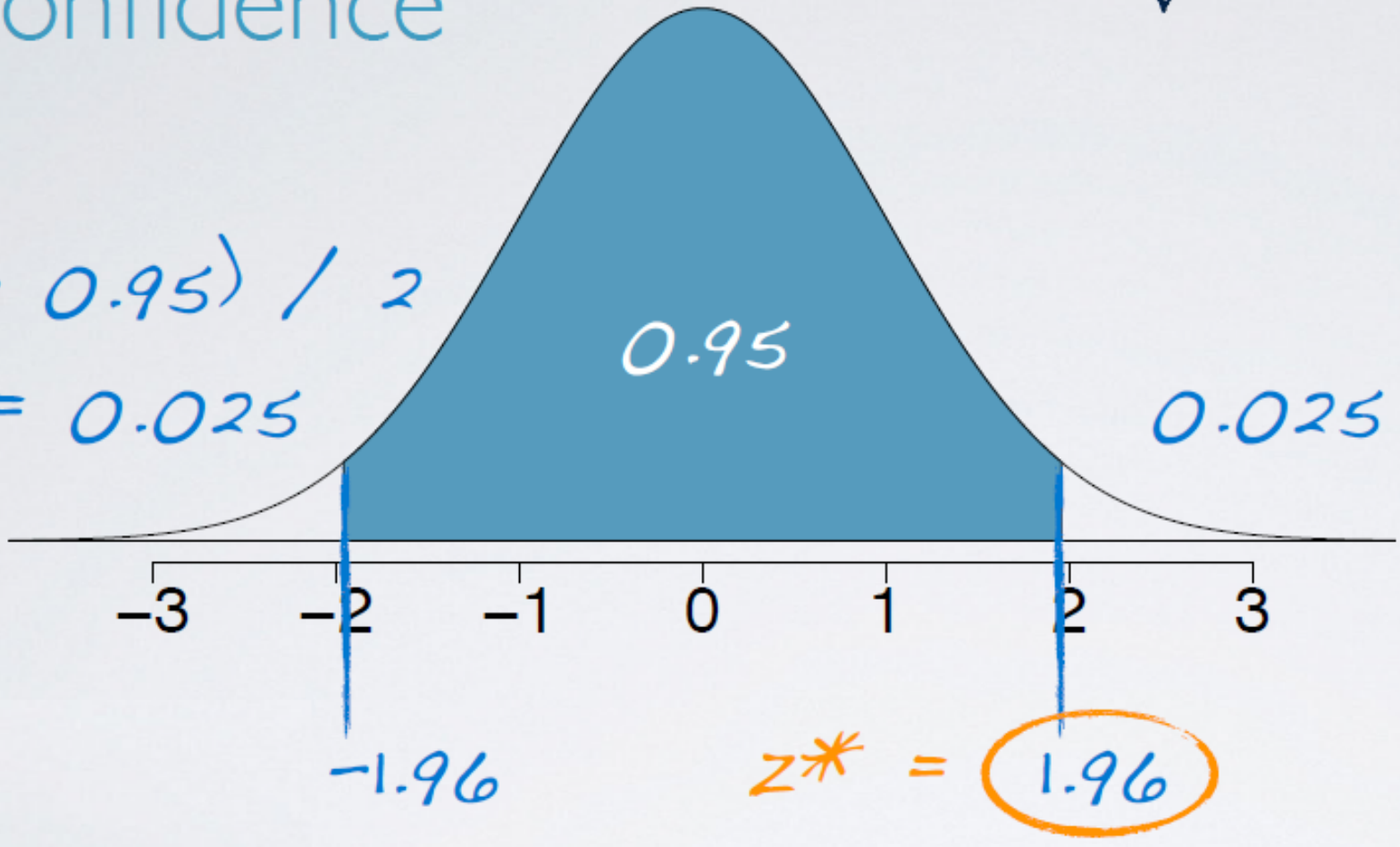
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Ex. For a 98% confidence level, z*=2.326

Example

finding the critical value
95% confidence

$$\bar{x} \pm z^{\star} \frac{s}{\sqrt{n}}$$



Z-score table

	Second decimal place				
0.07	0.06	0.05	0.04	0.00	Z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

Link between confidence level and margin of error

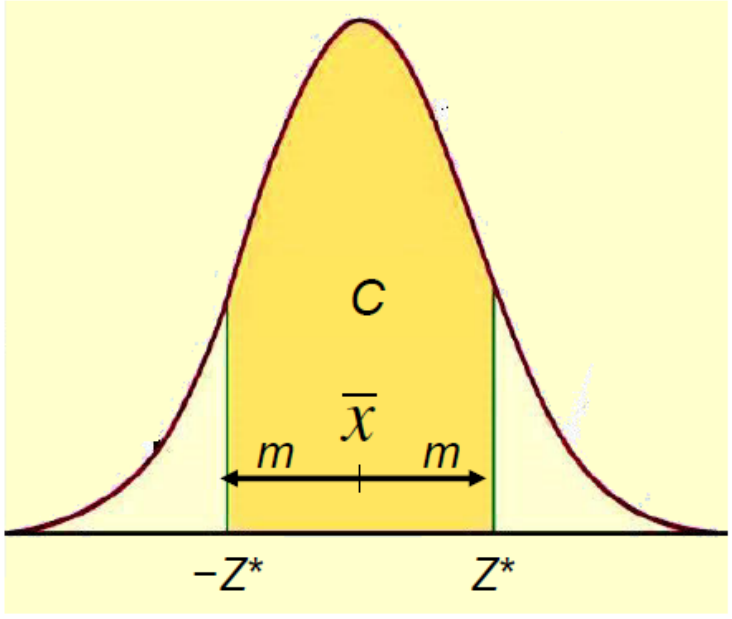
The confidence level *C* determines the value of z^* (in Table *C*).

The margin of error also depends on z^*.

$$m = z^* \sigma / \sqrt{n}$$

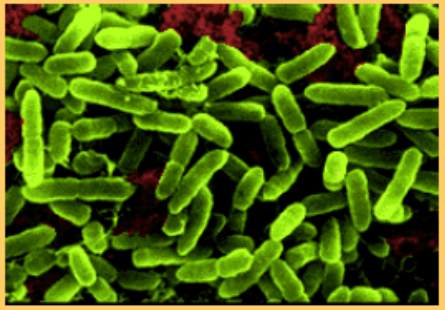
Higher confidence ***C*** implies a larger margin of error ***m*** (thus less precision in our estimates).

A lower confidence level ***C*** produces a smaller margin of error ***m*** (thus better precision in our estimates).



Example

Different confidence intervals for the same set of measurements



Density of bacteria in solution:

Measurement equipment has standard deviation $\sigma = 1*10^6$ bacteria/ml fluid.

3 measurements: 24, 29, and $31*10^6$ bacteria/ml fluid

Mean: $\bar{x} = 28*10^6$ bacteria/ml. Find the 96% and 70% CI.

- 96% confidence interval for the true density, $z^* = 2.054$, and write

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 28 \pm 2.054(1/\sqrt{3})$$

$$= 28 \pm 1.19*10^6 \text{ bacteria/ml}$$

- 70% confidence interval for the true density, $z^* = 1.036$, and write

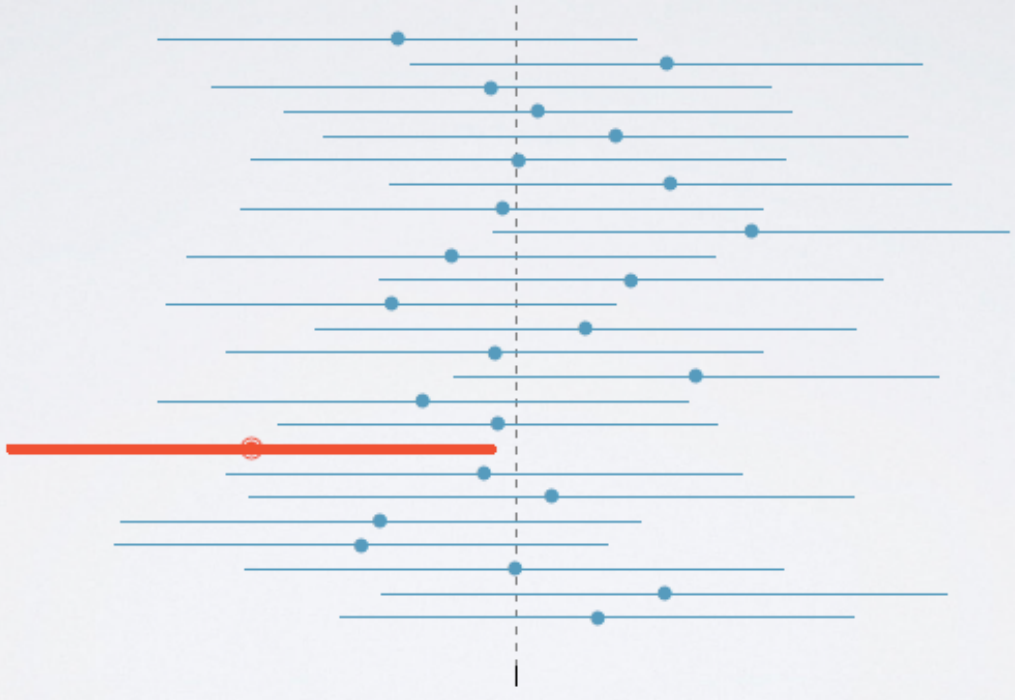
$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 28 \pm 1.036(1/\sqrt{3})$$

$$= 28 \pm 0.60*10^6 \text{ bacteria/ml}$$

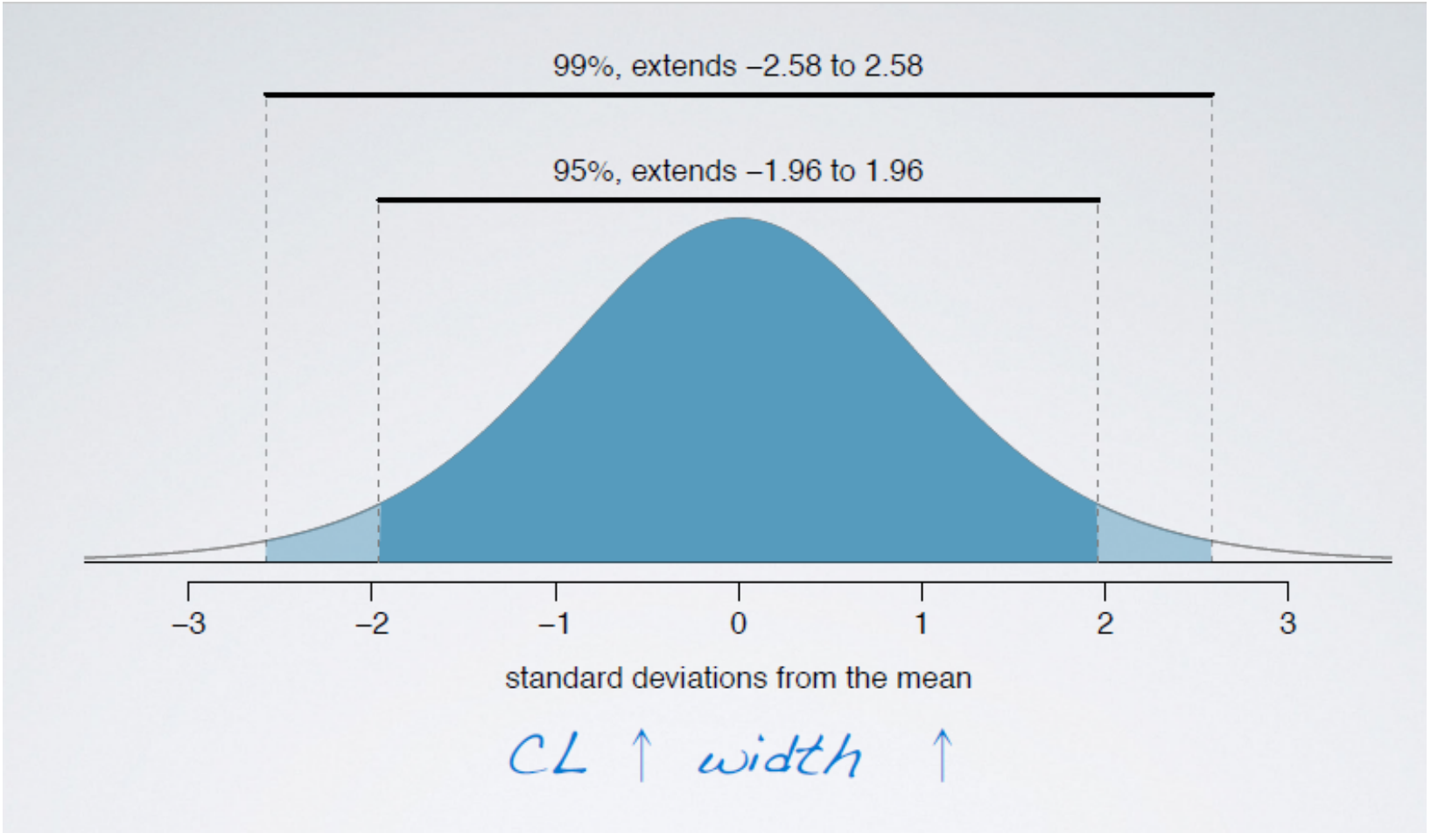
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Accuracy vs Precision

If we want to be very certain that we capture the population parameter, should we use a wider interval or a narrower interval?

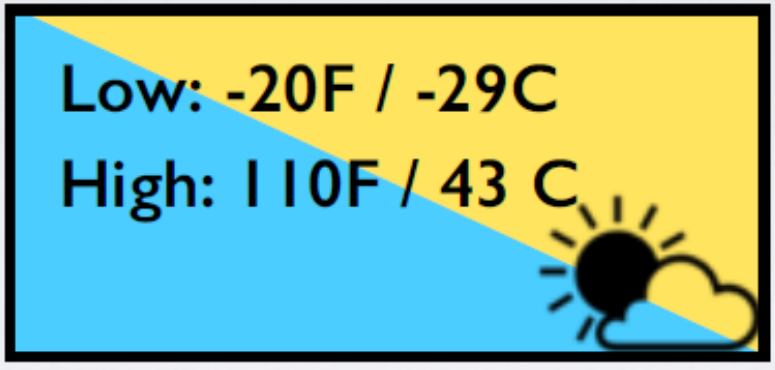


Accuracy vs Precision



Accuracy vs Precision

What drawbacks are associated with using a wider interval?



CL ↑ width ↑ accuracy ↑

precision ↓

Accuracy vs Precision

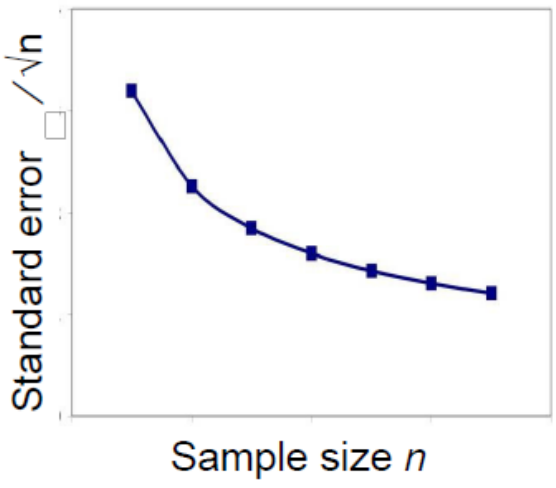
How can we get the best of both worlds — higher precision and higher accuracy?

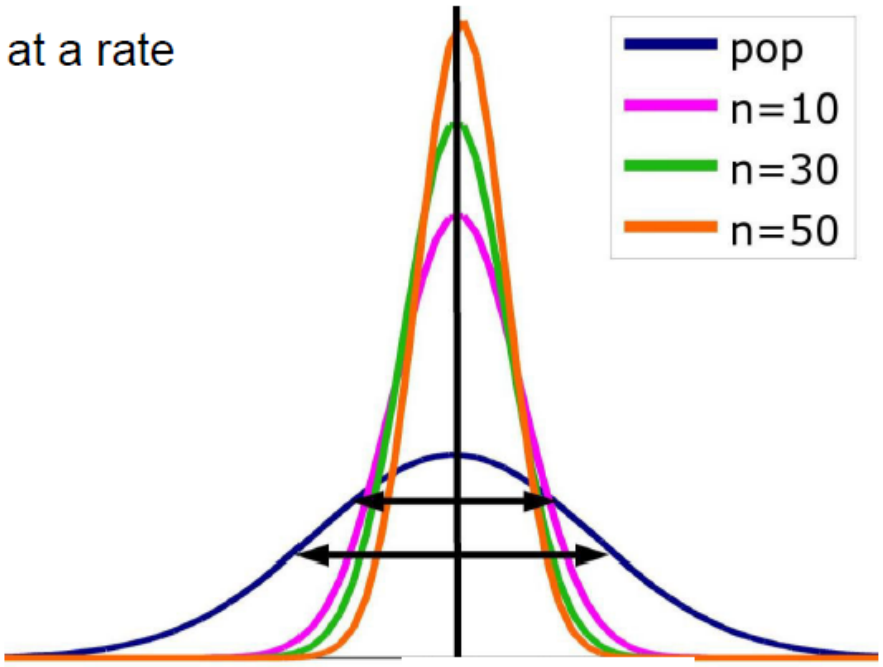
increase sample size

Impact of sample size

The spread in the sampling distribution of the mean is a function of the number of individuals per sample.

- The larger the sample size, the smaller the standard deviation (spread) of the sample mean distribution.
- But the spread only decreases at a rate equal to $\sqrt{n}$.





Sample size and experimental design

You may need a certain margin of error (e.g., drug trial, manufacturing specs). In many cases, the population variability (σ) is fixed, but we can choose the number of measurements (n).

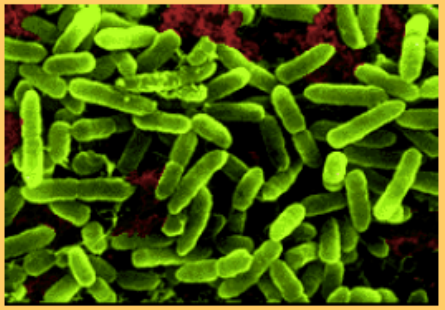
So plan ahead what sample size to use to achieve that margin of error.

$$m = z^* \frac{\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad n = \left(\frac{z^* \sigma}{m} \right)^2$$

Remember, though, that sample size is not always stretchable at will. There are typically costs and constraints associated with large samples. The best approach is to use the smallest sample size that can give you useful results.

Margin of Error

What sample size for a given margin of error?



Density of bacteria in solution:

Measurement equipment has standard deviation $\sigma = 1*10^6$ bacteria/ml fluid.

How many measurements should you make to obtain a margin of error of at most $0.5*10^6$ bacteria/ml with a confidence level of 90%?

For a 90% confidence interval, $z^*= 1.645$.

$$n = \left(\frac{z^*\sigma}{m}\right)^2 \Rightarrow n = \left(\frac{1.645*1}{0.5}\right)^2 = 3.29^2 = 10.8241$$

Using only 10 measurements will not be enough to ensure that m is no more than 0.5*106. Therefore, we need at least 11 measurements.

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Margin of Error

Suppose that the depths of female kangaroo pouches are normally distributed with unknown mean μ, and known standard deviation σ = 1.75 inches. For a random sample of 30 of the female kangaroos, we calculate a sample mean pouch depth of 9.3 inches.

- a) Construct a 95% confidence interval for **μ**

$$\bar{X} \sim N\left(\mu, \frac{1.75}{\sqrt{30}}\right)$$

$$\bar{x} = 9.3$$

$$z^* = 1.96$$

$$\Rightarrow \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 9.3 \pm 1.96 \frac{1.75}{\sqrt{30}} = [8.67377, 9.92623]$$

Margin of Error

b) How large of a sample size should you use if you want your margin of error to be no more than 0.5 inches with **90% confidence?**

$$m = 0.5$$

$$z^* = 1.645$$

$$n = \left(\frac{z^* \sigma}{m} \right)^2 = \left(\frac{1.645 \times 1.75}{0.5} \right)^2 = 33.14881 \approx 34$$

Example

A group of researchers want to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this medication during pregnancy.

Previous studies suggest that the SD of IQ scores of three-year-old children is 18 points.

How many such children should the researchers sample in order to obtain a 90% confidence interval with a margin of error less than or equal to 4 points?

$ME \leq 4$ pts

$CL = 90\%$

$z^* = 1.65$

$\sigma = 18$

$$4 = 1.65 \frac{18}{\sqrt{n}} \longrightarrow n = \left(\frac{1.65 \times 18}{4}\right)^2 = 55.13$$

We need at least 56 such children in the sample obtain a maximum margin of error of 4 points.

Example

We found that we needed at least 56 children in the sample to achieve a maximum margin of error of 4 points. How would the required sample size change if we want to further decrease the margin of error to 2 points?

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{n}} \frac{1}{2}$$

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{4n}}$$

$$4n = 56 \times 4 = 224$$

Example

A sample of 50 college students were asked how many exclusive relationships they've been in so far. The students in the sample had an average of 3.2 exclusive relationships, with a standard deviation of 1.74. In addition, the sample distribution was only slightly skewed to the right. Estimate the true average number of exclusive relationships based on this sample using a 95% confidence interval.

$n = 50$

$\bar{x} = 3.2$

$s = 1.74$

1. random sample & 50 < 10% of all college students

We can assume that the number of exclusive relationships one student in the sample has been in is independent of another.

2. $n > 30$ & not so skewed sample

We can assume that the sampling distribution of average number of exclusive relationships from samples of size 50 will be nearly normal.

Example

$n = 50$

$\bar{x} = 3.2$

$s = 1.74$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.246$$

$$\bar{x} \pm z^* \, SE = 3.2 \pm 1.96\,(0.246)$$

$$= 3.2 \pm 0.48$$

$$= (2.72, 3.68)$$

We are 95% confident that college students on average have been in 2.72 to 3.68 exclusive relationships.