JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

# Section B

## The Theoretical Sampling Distribution of the Sample Mean and Its Estimate Based on a Single Sample

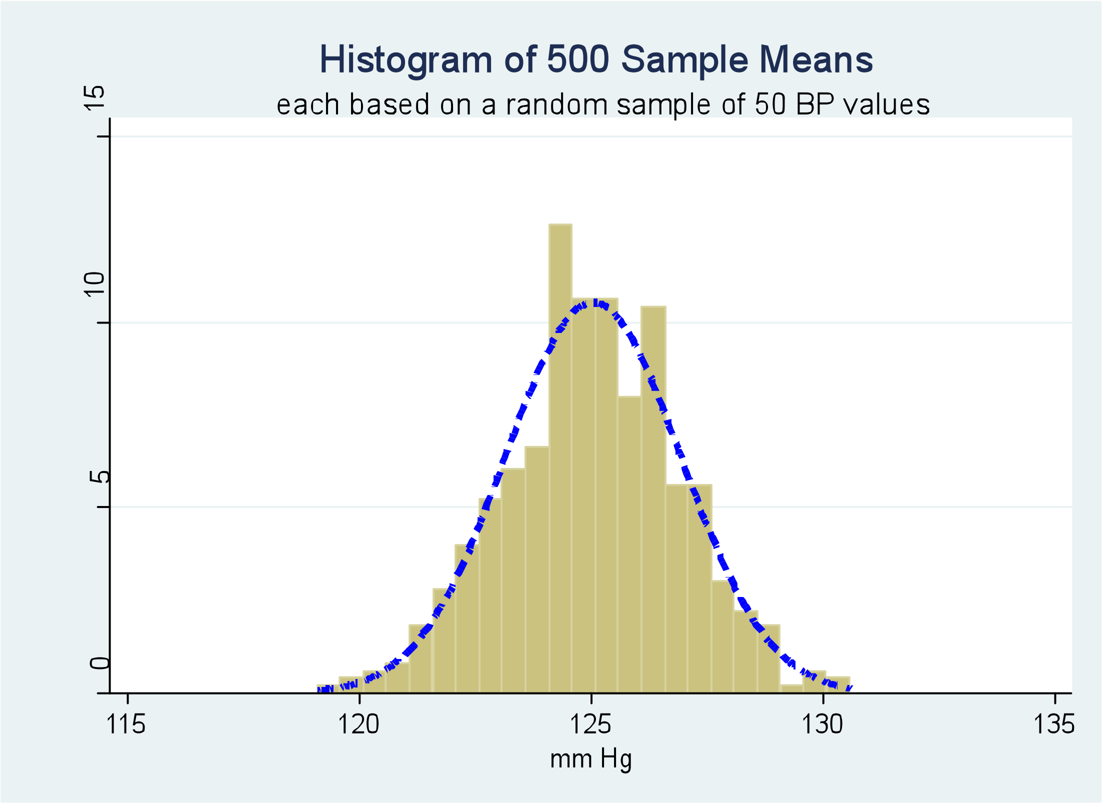
# Sampling Distribution of the Sample Mean

- In the previous section we reviewed the results of simulations that resulted in estimates of what's formally called the sampling distribution of a sample mean

- The sampling distribution of a sample mean is a theoretical probability distribution; it describes the distribution of all sample means from all possible random samples of the same size taken from a population

# Sampling Distribution of the Sample Mean

- For example: this histogram is an estimate of the sampling distribution of sample BP means based on random samples of *n* = 50 from the population of (BP measurements for) all men



# Sampling Distribution of the Sample Mean

- In real research it is impossible to estimate the sampling distribution of a sample mean by actually taking multiple random samples from the same population, no research would ever happen if a study needed to be repeated multiple times to understand this sampling behavior

- Simulations are useful to illustrate a concept, but not to highlight a practical approach!

- Luckily, there is some mathematical machinery that generalizes some of the patterns we saw in the simulation results

# The Central Limit Theorem (CLT)

- The Central Limit Theorem (CLT) is a powerful mathematical tool that gives several useful results
  - The sampling distribution of sample means based on all samples of same size *n* is approximately normal, regardless of the distribution of the original (individual level) data in the population/samples
  - The mean of all sample means in the sampling distribution is the true mean of the population from which the samples were taken, μ
  - Standard deviation in the sample means of size *n* is equal to $\frac{\sigma}{\sqrt{n}}$: this is often called the standard error of the sample mean and sometimes written as $SE(\overline{x})$

# Example: Blood Pressure of Males

- Population distribution of individual BP measurements for males: normal
- True mean μ = 125 mmHg: σ = 14 mmHg

| Sample Sizes | Means of 500 Sample Means | Means of 5000 Sample Means | SD of 500 Sample Means | SD of 5000 Sample Means | SD of Sample Means (SE) by CLT |
|---|---|---|---|---|---|
| n = 20 | 124.98 mmHg | 125.05 mmHg | 3.31 mmHg | 3.11 mmHg | 3.13 mmHg |
| n = 50 | 125.03 mmHg | 125.01 mmHg | 1.89 mmHg | 1.96 mmHg | 1.98 mmHg |
| n = 100 | 124.99 mmHg | 125.01 mmHg | 1.43 mmHg | 1.39 mmHg | 1.40 mmHg |

# Example: Blood Pressure of Males

- Population distribution of individual BP measurements for males: normal

- True mean μ = 125 mmHg: σ = 14 mmHg

| Sample Sizes | Means of 500 Sample Means | Means of 5000 Sample Means | SD of 500 Sample Means | SD of 5000 Sample Means | SD of Sample Means (SE) by CLT |
|---|---|---|---|---|---|
| n = 20 | 124.98 mmHg | 125.05 mmHg | **3.31 mmHg** | **3.11 mmHg** | **3.13 mmHg** |
| n = 50 | 125.03 mmHg | 125.01 mmHg | 1.89 mmHg | 1.96 mmHg | 1.98 mmHg |
| n = 100 | 124.99 mmHg | 125.01 mmHg | 1.43 mmHg | 1.39 mmHg | 1.40 mmHg |

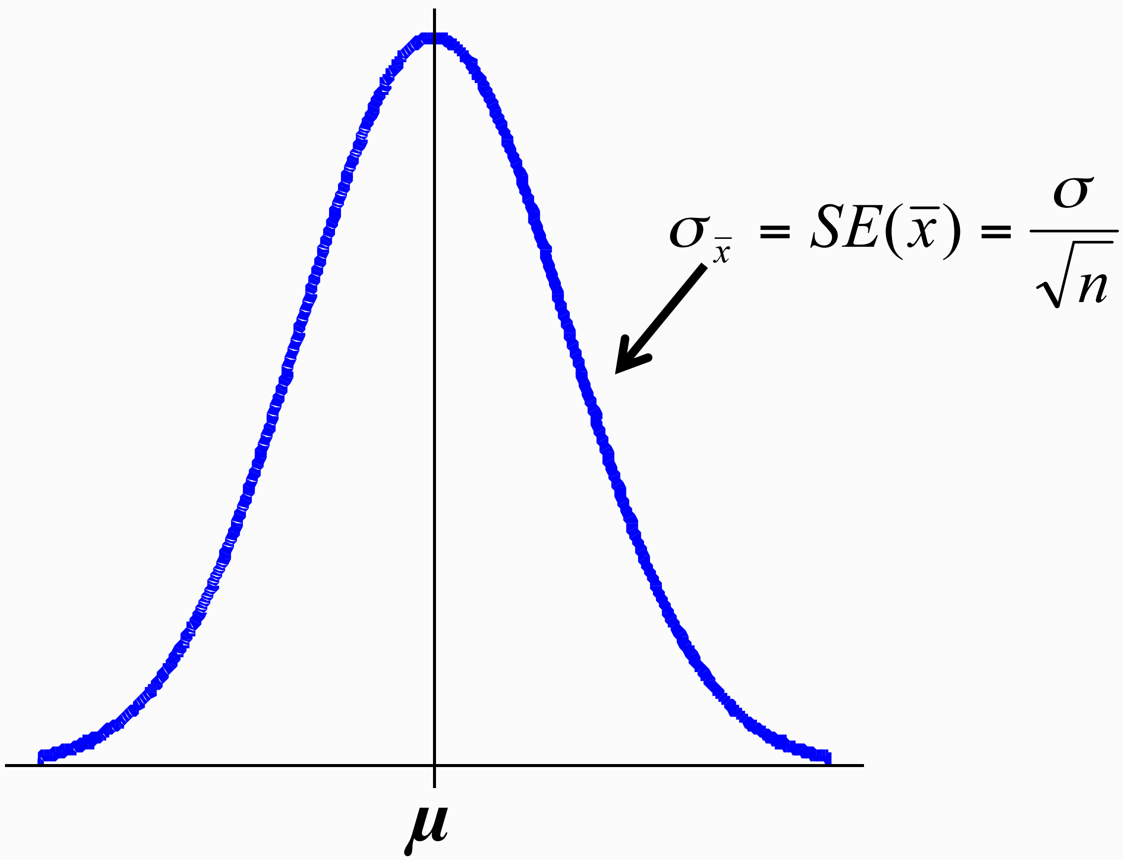
# Example: Blood Pressure of Males

- Population distribution of individual BP measurements for males: normal
- True mean μ = 125 mmHg: σ = 14 mmHg

| Sample Sizes | Means of 500 Sample Means | Means of 5000 Sample Means | SD of 500 Sample Means | SD of 5000 Sample Means | SD of Sample Means (SE) by CLT |
|---|---|---|---|---|---|
| n = 20 | 124.98 mmHg | 125.05 mmHg | 3.31 mmHg | 3.11 mmHg | 3.13 mmHg |
| n = 50 | 125.03 mmHg | 125.01 mmHg | **1.89 mmHg** | **1.96 mmHg** | **1.98 mmHg** |
| n = 100 | 124.99 mmHg | 125.01 mmHg | 1.43 mmHg | 1.39 mmHg | 1.40 mmHg |

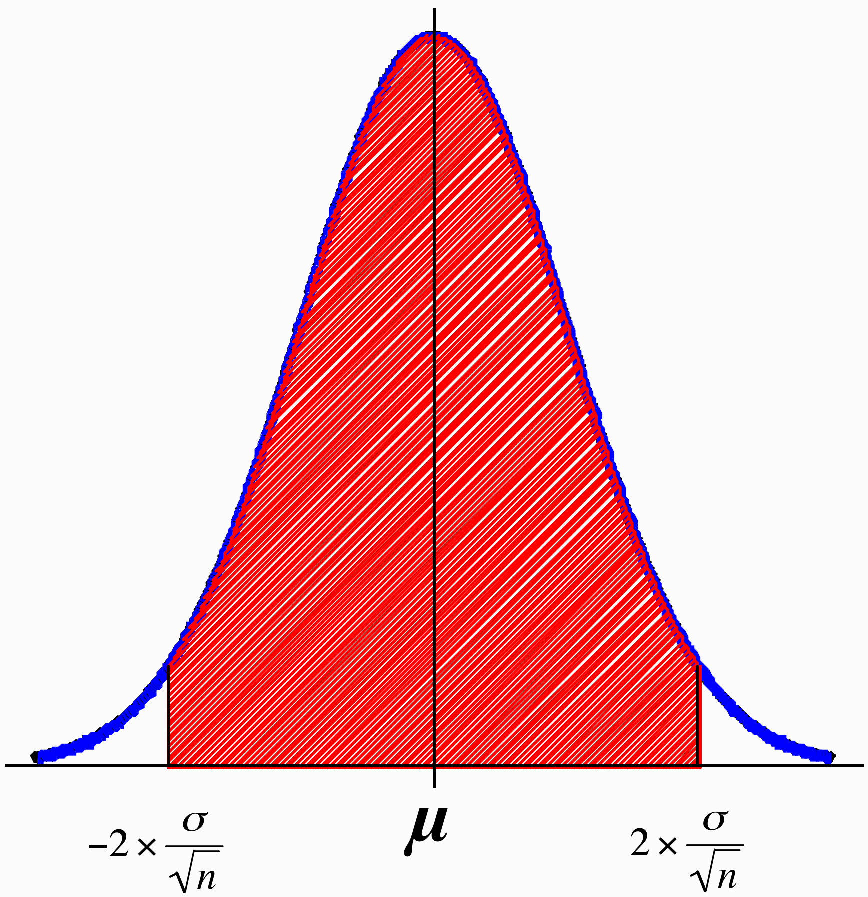
# Recap: CLT

- So the CLT tells us the following:
  - When taking a random sample of continuous measures of size *n* from a population with true mean $\mu$ and true sd $\sigma$ the theoretical sampling distribution of sample means from all possible random samples of size *n* is as follows:



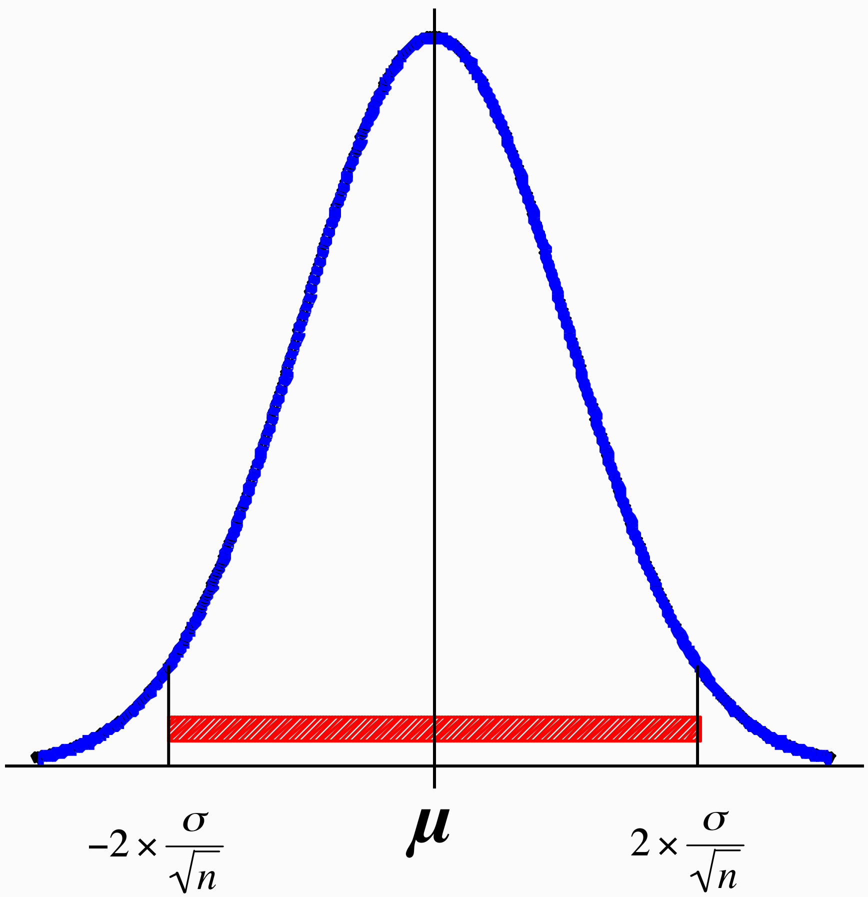
# CLT: So What?

- So what good is this info?
  - Well using the properties of the normal curve, this shows that for most random samples we can take (95%), the sample mean will fall within 2 SEs of the true mean μ: $\overline{x}$



# CLT: So What?

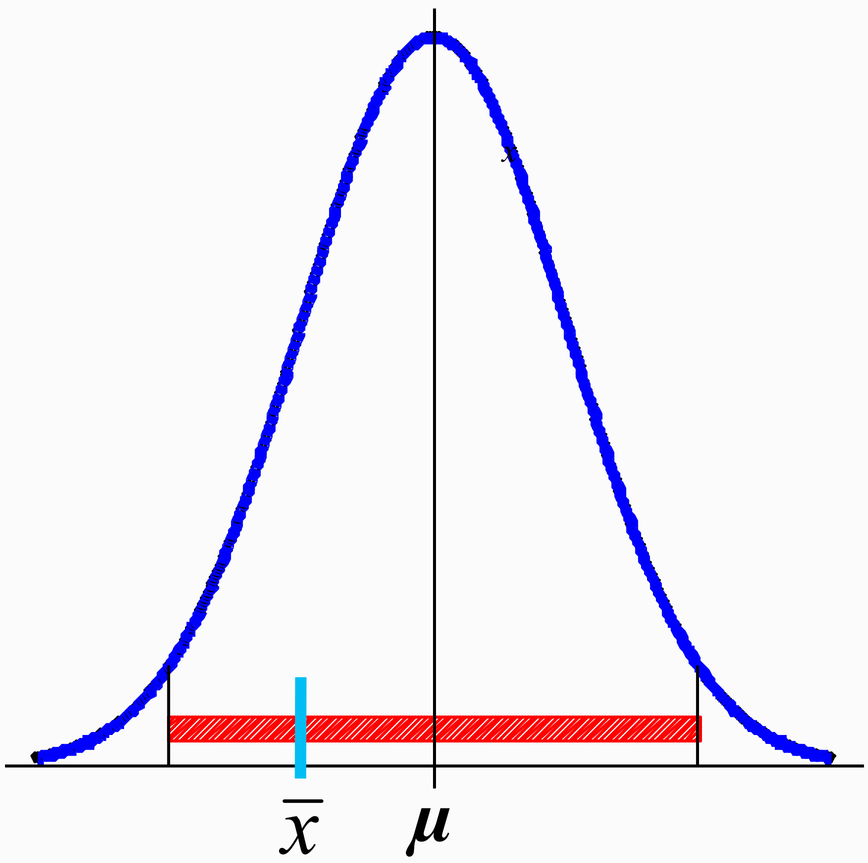
- So AGAIN what good is this info?
  - We are going to take a single sample of size *n* and get one $\overline{x}$
  - So we won't know μ, and if we did know μ why would we care about the distribution of estimates of μ from imperfect subsets of the population?



# CLT: So What?

- We are going to take a single sample of size *n* and get one $\bar{x}$

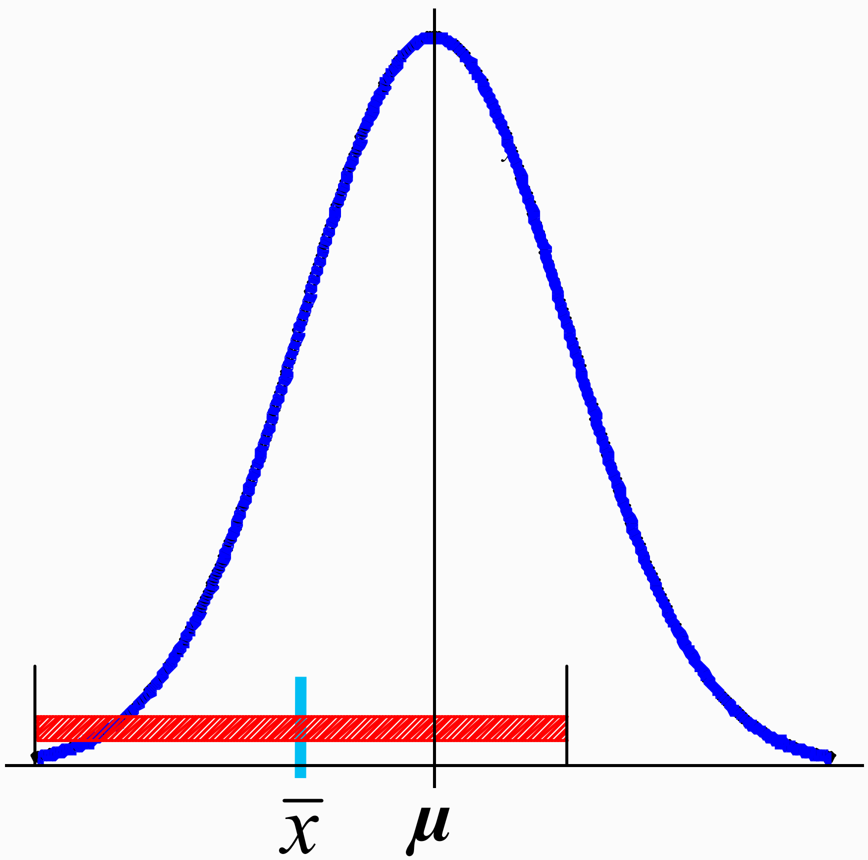
- But for most (95%) of the random samples we can get, our $\bar{x}$ will fall within +/- 2SEs of μ



# CLT: So What?

- We are going to take a single sample of size *n* and get one $\overline{x}$

- So if we start at $\overline{x}$ and go 2SEs in either direction, the interval created will contain μ most (95 out of 100) of the time



# Estimating a Confidence Interval

- Such an interval is a called a 95% confidence interval for the population mean $\mu$
- Interval given by $\bar{x} \pm 2SE(\bar{x}) \rightarrow \bar{x} \pm 2 * \frac{\sigma}{n}$
- Problem: we don't know $\sigma$ either
  - Can estimate with $s$, will detail this in next section
- What is interpretation of a confidence interval?

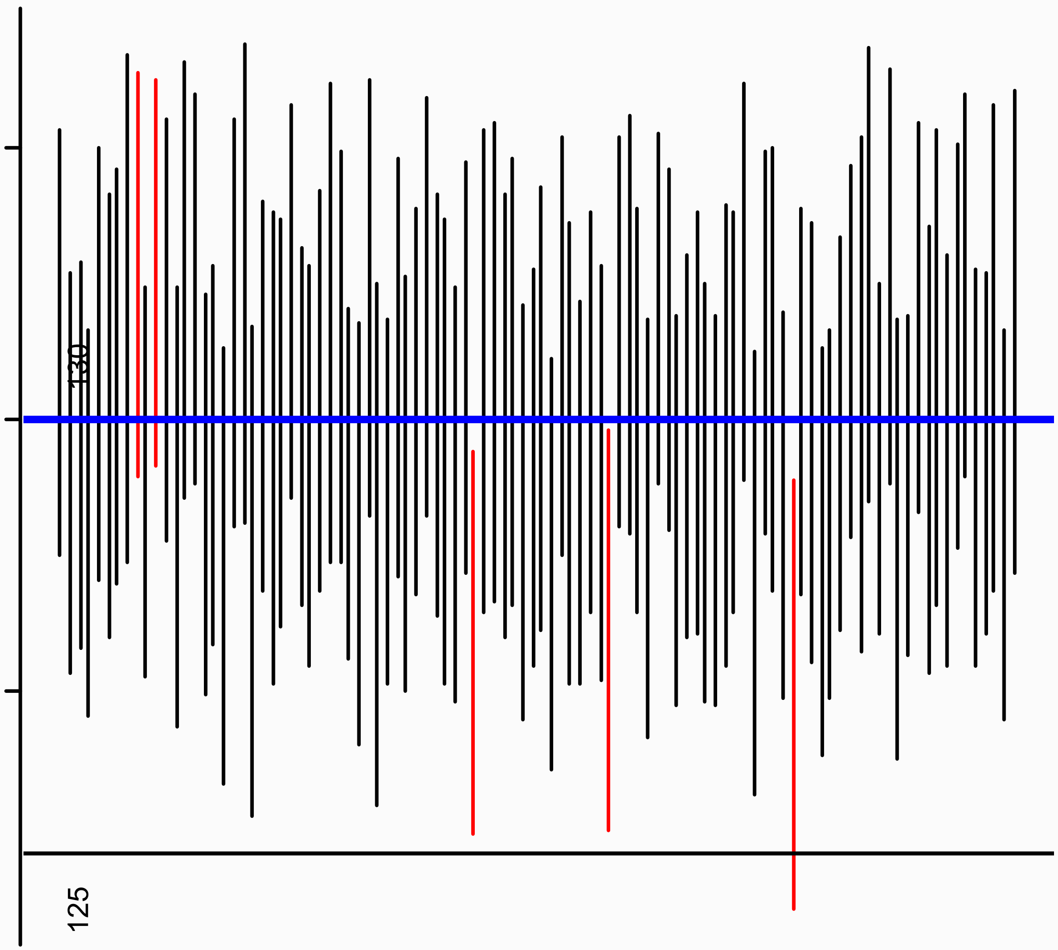
# Interpretation of a 95% Confidence Interval (CI)

- Laypersons' range of "plausible" values for true mean
  - Researcher never can observe true mean μ
  - $\bar{x}$ is the best estimate based on a single sample
  - The 95% CI starts with this best estimate and additionally recognizes uncertainty in this quantity

- Technical
  - Were 100 random samples of size *n* taken from the same population, and 95% confidence intervals computed using each of these 100 samples, 95 of the 100 intervals would contain the values of true mean μ within the endpoints

# Technical Interpretation

- One hundred 95% confidence intervals from 100 random samples of size $n$ = 50: Blood Pressure for Males



# Notes on Confidence Intervals

- Random sampling error
  - Confidence interval only accounts for random sampling error—not other systematic sources of error or bias

# Examples of Systematic Bias

- BP measurement is always +5 too high (broken instrument)
- Only those with high BP agree to participate (non-response bias)

# Notes on Confidence Intervals

- Are all CIs 95%?
  - No
  - It is the most commonly used
  - A 99% CI is wider
  - A 90% CI is narrower

- To change level of confidence adjust number of SE added and subtracted from $\bar{x}$
  - For a 99% CI, you need ± 2.6 SE
  - For a 95% CI, you need ± 2 SE
  - For a 90% CI, you need ± 1.65 SE

# Semantic:Standard Deviation vs. Standard Error

- The term “standard deviation” refers to the variability in individual observations in a single sample (*s*) or population ($\sigma$)

- The standard error of the mean is also a measure of standard deviation, but not of individual values, rather variation in multiple sample means computed on multiple random samples of the same size, taken from the same population