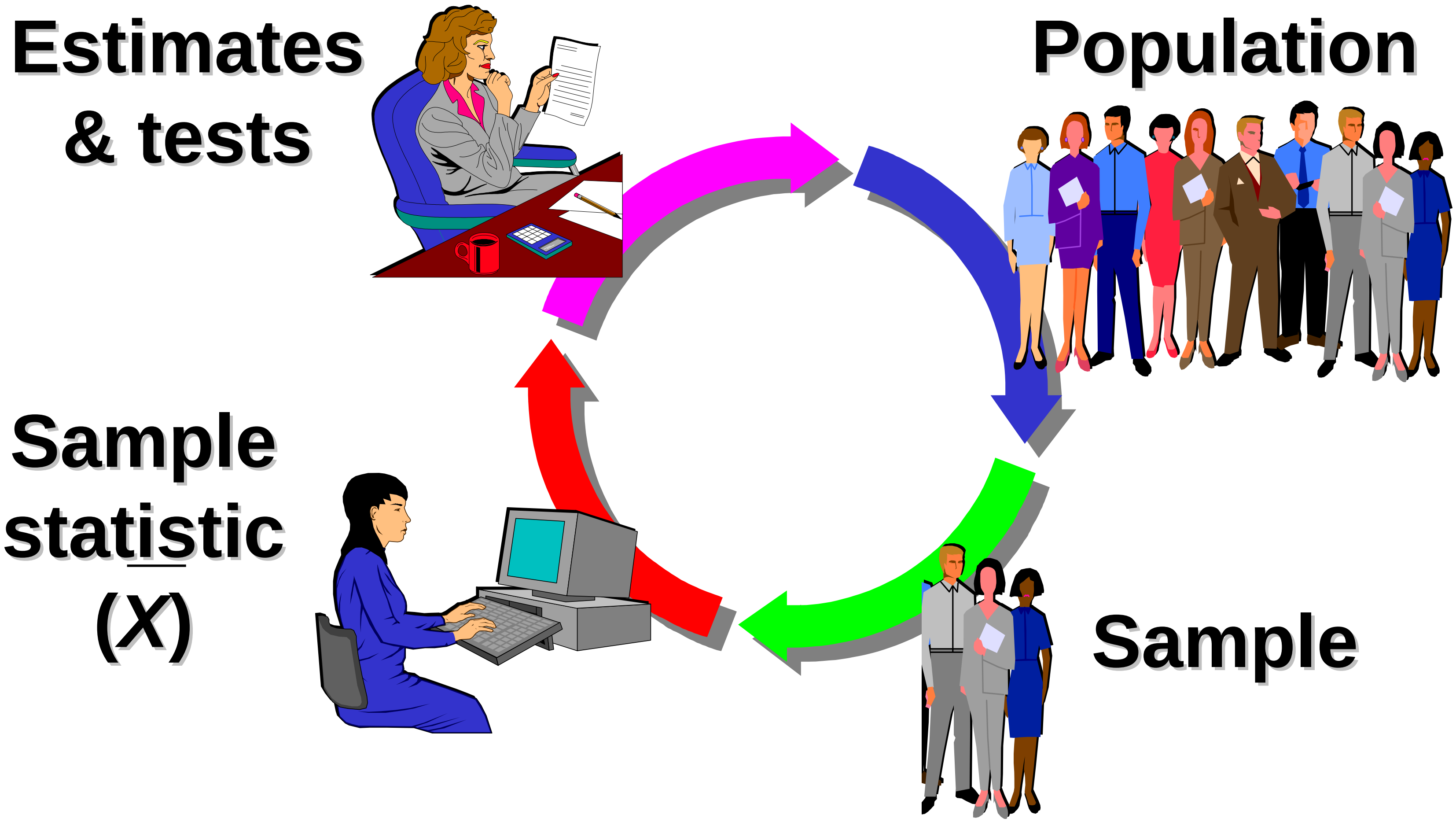
Lecture 16

The Diversity of Samples from the Same Population

Inference Process



Thought Question 1:

40% of large population disagree with new law.
In parts a and b, think about role of sample size.

a. If randomly **sample 10** people, will exactly four (40%) disagree with law? Surprised if only two in sample disagreed? How about if none disagreed?

b. If randomly **sample 1000** people, will exactly 400 (40%) disagree with law? Surprised if only 200 in sample disagreed? How about if none disagreed?

c. Explain how long-run relative-frequency interpretation of probability and gambler's fallacy helped you answer parts a and b.

Thought Question 2:

Mean weight of all women at large university is **135 pounds** with a **standard deviation** of **10 pounds**.

a. Recalling Empirical Rule for bell-shaped curves, in what range would you expect **95%** of women's weights to fall?

b. If randomly **sampled 10 women** at university, how close do you think their *average* weight would be to 135 pounds? If sampled 1000 women, would you expect *average* weight to be closer to 135 pounds than for the sample of only 10 women?

Thought Question 3:

Survey of 1000 randomly selected individuals has a **margin of error of about 3%**, so results accurate to within ± 3% most of the time.

Suppose **25% of adults believe in reincarnation**.

If **ten polls** are taken, each asking a different random sample of 1000 adults about belief in reincarnation, would you **expect each poll to find exactly 25%** of respondents expressing belief in reincarnation?

If not, into **what range** would you expect the ten sample proportions to reasonably fall?

19.1 Setting the Stage

Working Backward from Samples to Populations

- Start with question about population.
- Collect a sample from the population, measure variable.
- Answer question of interest for sample.
- With statistics, determine how close such an answer, based on a sample, would tend to be from the actual answer for the population.

Understanding Dissimilarity among Samples

- Suppose most samples are likely to provide an answer that is within 10% of the population answer.
- Then the population answer is expected to be within 10% of whatever value the sample gave.
- So, can make a good guess about the population value.

19.2 What to Expect of Sample Proportions

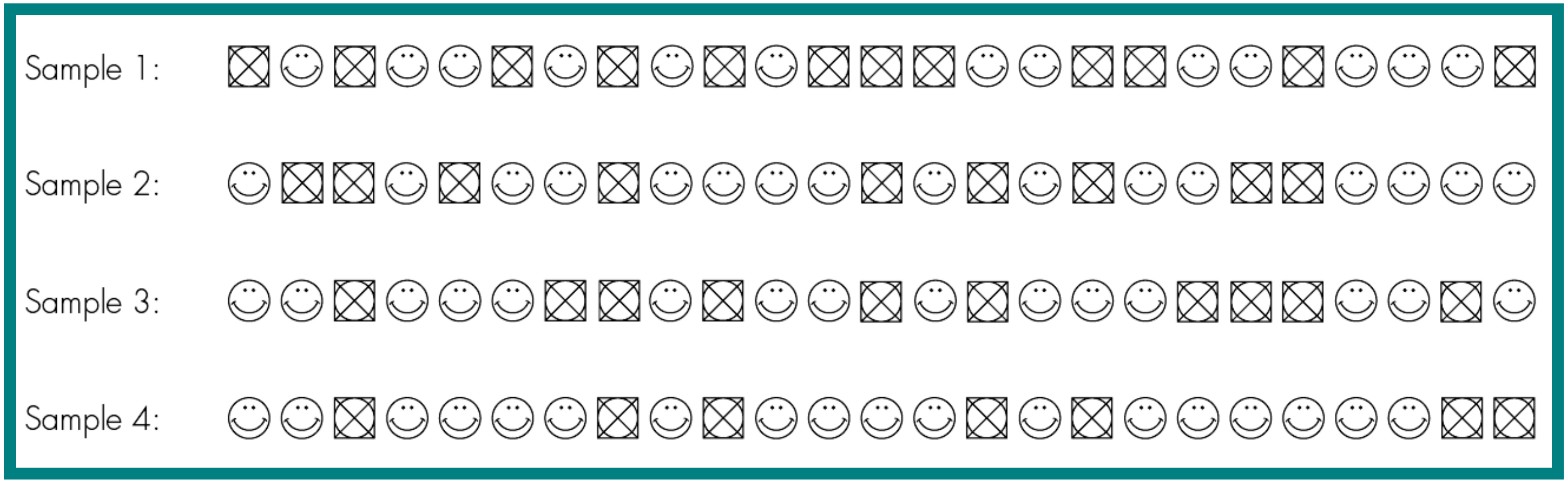
40% of population carry a certain gene

Do Not Carry Gene = ☺, Do Carry Gene = ☒

A slice of the population:

☺☒☺☺☺☒☒☺☒☺☺☺☒☒☺☺☺☺☺☺☺☺☺☺☺☺☺☺☒☺☺
☺☺☒☒☺☺☒☺☺☒☺☺☺☺☒☒☺☒☒☺☒☺☺☺☒☒☺☺☒☒☺☒☺
☺☺☺☺☺☺☺☺☺☺☺☒☒☒☒☒☺☒☺☺☺☒☒☺☒☺☺☺☒☒☺☺☺
☺☺☺☺☺☺☺☺☺☺☺☺☺☒☺☺☺☺☒☒☺☺☒☺☺☒☺☺☺☺☒☒☺
☒☒☺☒☺☺☺☒☒☺☺☒☒☺☒☺☺☺☺☺☺☺☺☺☺☺☺☒☒☒☒☒☒
☒☒☒☒☺☺☺☺☺☺☒☺☒☺☺☺☒☺☺☒☒☺☒☒☺☺☺☒☒☺☺☒☺
☺☺☺☒☒☺☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☒☺☺☺☺☺☺☺☺☺
☺☺☺☺☺☺☒☒☒☒☒☒☒☒☒☺☺☺☺☺☺☺☒☒☺☒☒☺☒☺☺☺☒
☒☺☺☒☒☺☒☺☺☺☺☺☒☒☒☒☒☺☺☺☺☺☺☒☺☒☺☺☺☒☺☺☒
☒☺☒☒☺☺☺☒☒☺☺☒☺☺☺☺☒☒☺☒☒☒☒☒☒☒☒☒☒☒☒☒☒
☒☒☒☒☺☺☺☺☺☺☺☺☺☺☺☺☺☺☺☒☒☒☒☒☒☒☒☒☺☺☺☺☺
☺☺☒☒☺☒☒☺☒☺☺☺☒☒☺☺☒☒☺☒☺☺☺☺☺☺☺☺☺☺☺☺☺
☺☺☺☺....

Possible Samples



Sample 1: Proportion with gene = 12/25 = 0.48 = 48%
Sample 2: Proportion with gene = 9/25 = 0.36 = 36%
Sample 3: Proportion with gene = 10/25 = 0.40 = 40%
Sample 4: Proportion with gene = 7/25 = 0.28 = 28%

- ***Each sample gave a different answer.***
- ***Sample answer may or may not match population answer.***

Conditions for Rule for Sample Proportions

1. **There exists an actual population with fixed proportion** who have a certain trait. *Or*
   There exists a repeatable situation for which a certain outcome is likely to occur with fixed probability.
2. **Random sample** selected from population (so probability of observing the trait is same for each sample unit). *Or*
   Situation repeated numerous times, with outcome each time independent of all other times.
3. **Size of sample** or number of repetitions is relatively **large** – large enough to see at least 5 of each of the two possible responses.

Example 1: Election Polls

Pollster wants to estimate proportion of voters who favor a certain candidate. **Voters** are the ***population units***, and **favoring candidate** is ***opinion of interest***.

Example 2: Television Ratings

TV rating firm wants to estimate proportion of households with television sets tuned to a certain television program. Collection of **all households with television sets** makes up the ***population***, and being **tuned to program** is ***trait of interest***.

Example 3: Consumer Preferences

Manufacturer of soft drinks wants to know what proportion of consumers prefers new mixture of ingredients compared with old recipe. ***Population*** consists of **all consumers**, and ***response of interest*** is **preference** of new formula over old one.

Example 4: Testing ESP

Researcher wants to know the probability that people can successfully guess which of 5 symbols is on a hidden card. Each symbol is equally likely. ***Repeatable situation*** is a **guess**, and ***response of interest*** is **successful guess**.
Is the probability of correct guess higher than 20%?

Defining the Rule for Sample Proportions

If numerous samples or repetitions of the same size are taken, the frequency curve made from proportions from various samples will be **approximately bell-shaped**.

Mean will be true proportion from the population.

Standard deviation will be:

$$\sqrt{\frac{(\text{true proportion})(1 - \text{true proportion})}{\text{sample size}}}$$

Example 5: Using Rule for Sample Proportions

Suppose **40% of all voters** in U.S. favor candidate X. Pollsters take a sample of 2400 people. ***What sample proportion would be expected to favor candidate X?***

The sample proportion could be anything from a **bell-shaped** curve with **mean 0.40** and **standard deviation**:

$$\sqrt{\frac{(0.40)(1-0.40)}{2400}} = \mathbf{0.01}$$

For our sample of 2400 people:

- **68% chance** sample proportion is between 39% and 41%
- **95% chance** sample proportion is between 38% and 42%
- **almost certain** sample proportion is between 37% and 43%

19.3 What to Expect of Sample Means

- Want to estimate **average weight loss** for all who attend national weight-loss clinic for 10 weeks.
- Unknown to us, population mean weight loss is 8 pounds and standard deviation is 5 pounds.
- If weight losses are approximately bell-shaped, 95% of individual weight losses will fall between –2 (a gain of 2 pounds) and 18 pounds lost.

Possible Samples

Sample 1: 1,1,2,3,4,4,4,5,6,7,7,7,8,8,9,9,11,11,13,13,14,14,15,16,16
Sample 2: –2, 2,0,0,3,4,4,4,5,5,6,6,8,8,9,9,9,9,9,10,11,12,13,13,16
Sample 3: –4,–4,2,3,4,5,7,8,8,9,9,9,9,9,10,10,11,11,11,12,12,13,14,16,18
Sample 4: –3,–3,–2,0,1,2,2,4,4,5,7,7,9,9,10,10,10,11,11,12,12,14,14,14,19

Results:
Sample 1: Mean = 8.32 pounds, std dev = 4.74 pounds
Sample 2: Mean = 6.76 pounds, std dev = 4.73 pounds
Sample 3: Mean = 8.48 pounds, std dev = 5.27 pounds
Sample 4: Mean = 7.16 pounds, std dev = 5.93 pounds

- ***Each sample gave a different sample mean, but close to 8.***
- ***Sample standard deviation also close to 5 pounds.***

Conditions for Rule for Sample Means

1. **Population** of measurements is **bell-shaped**, and a **random sample** of any size is measured.

OR

2. **Population** of measurements of interest is **not bell-shaped**, but a **large random sample** is measured. Sample of size 30 is considered "large," but if there are extreme outliers, better to have a larger sample.

Sampling from Normal Populations

- **Central Tendency**

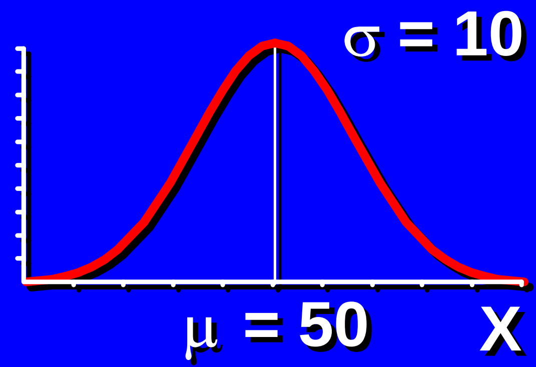
$$\mu_{\bar{x}} = \mu$$

- **Dispersion**

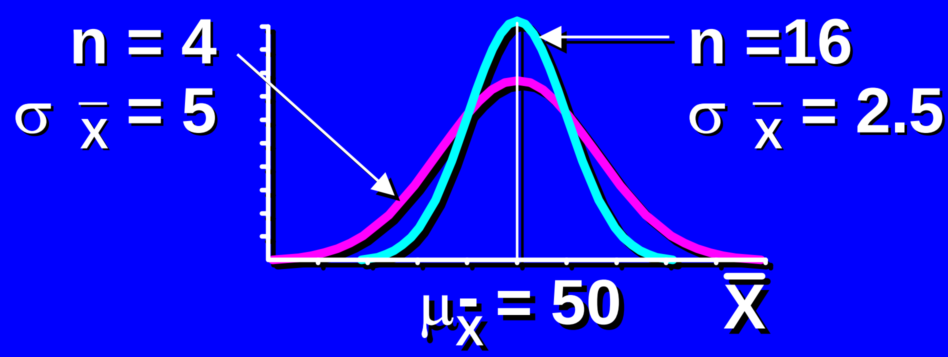
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sampling with replacement

Population Distribution



Sampling Distribution



Sampling from Non-Normal Populations

- **Central Tendency**

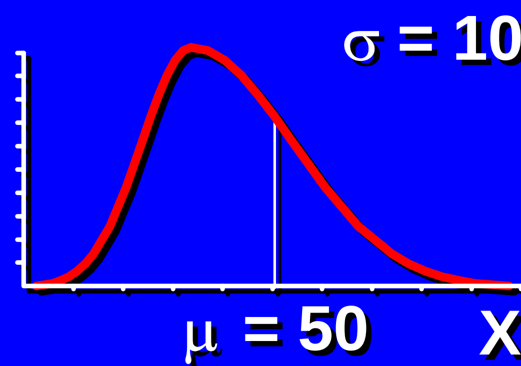
$$\mu_{\bar{x}} = \mu$$

- **Dispersion**

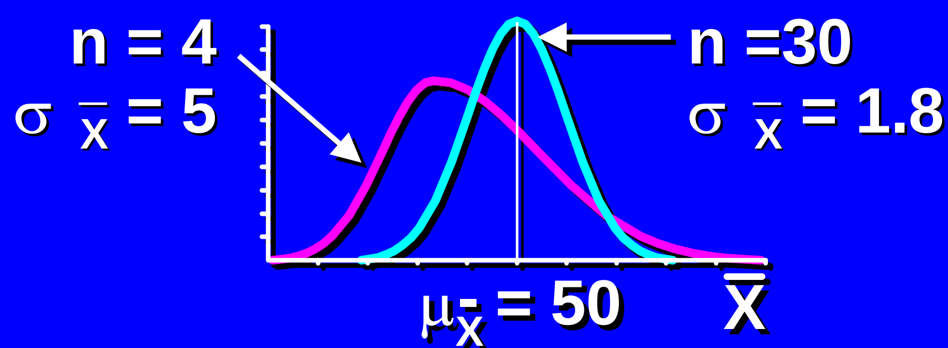
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  - Sampling with replacement

Population Distribution

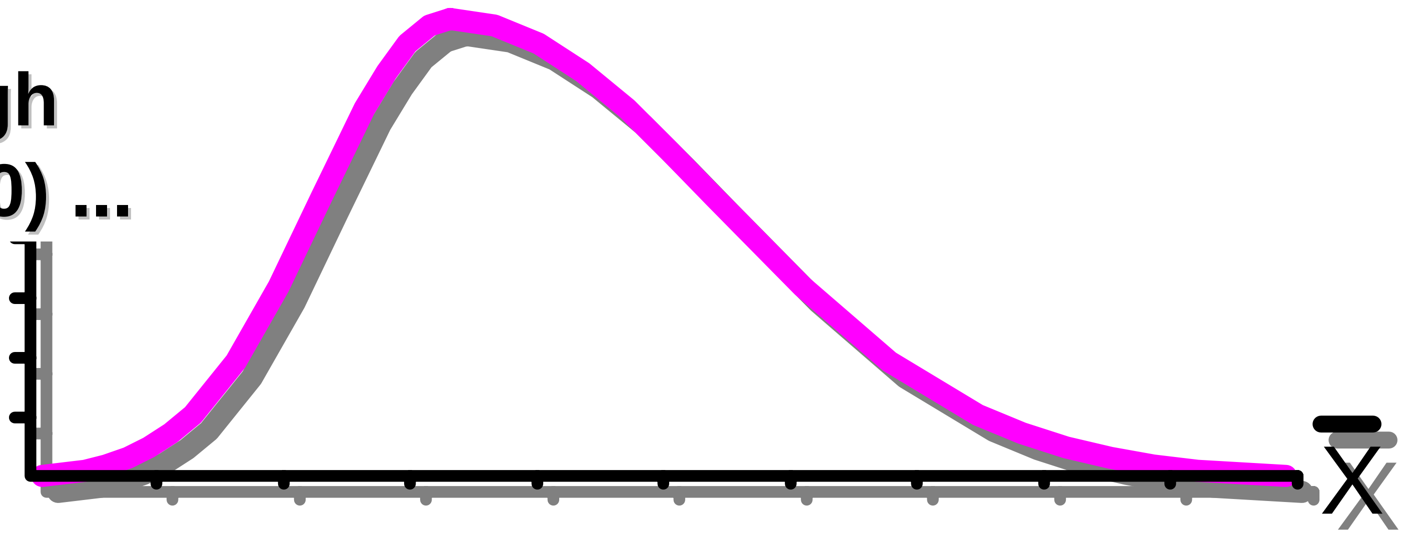


Sampling Distribution



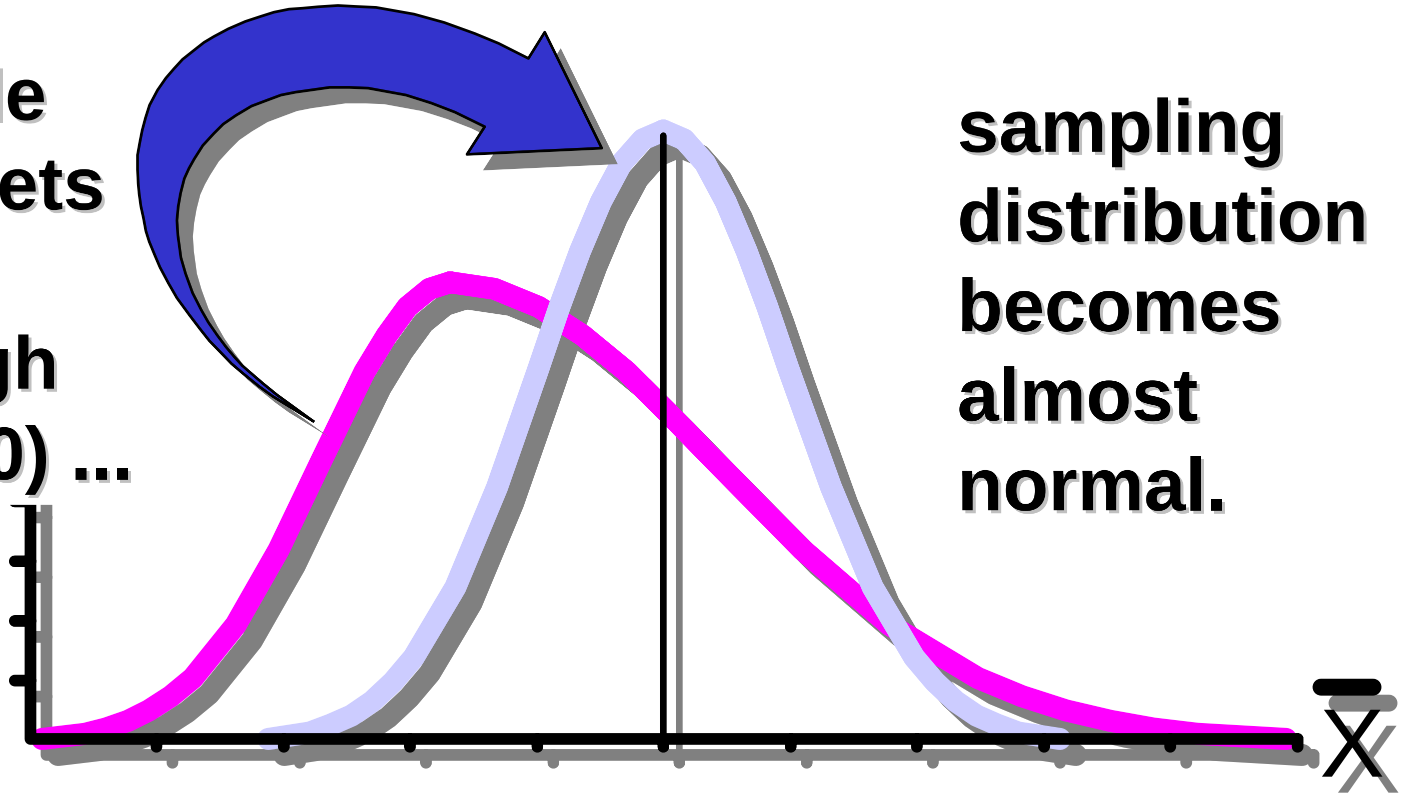
Central Limit Theorem

As sample size gets large enough (n ≥ 30) ...



Central Limit Theorem

As sample size gets large enough ($n \geq 30$) ...



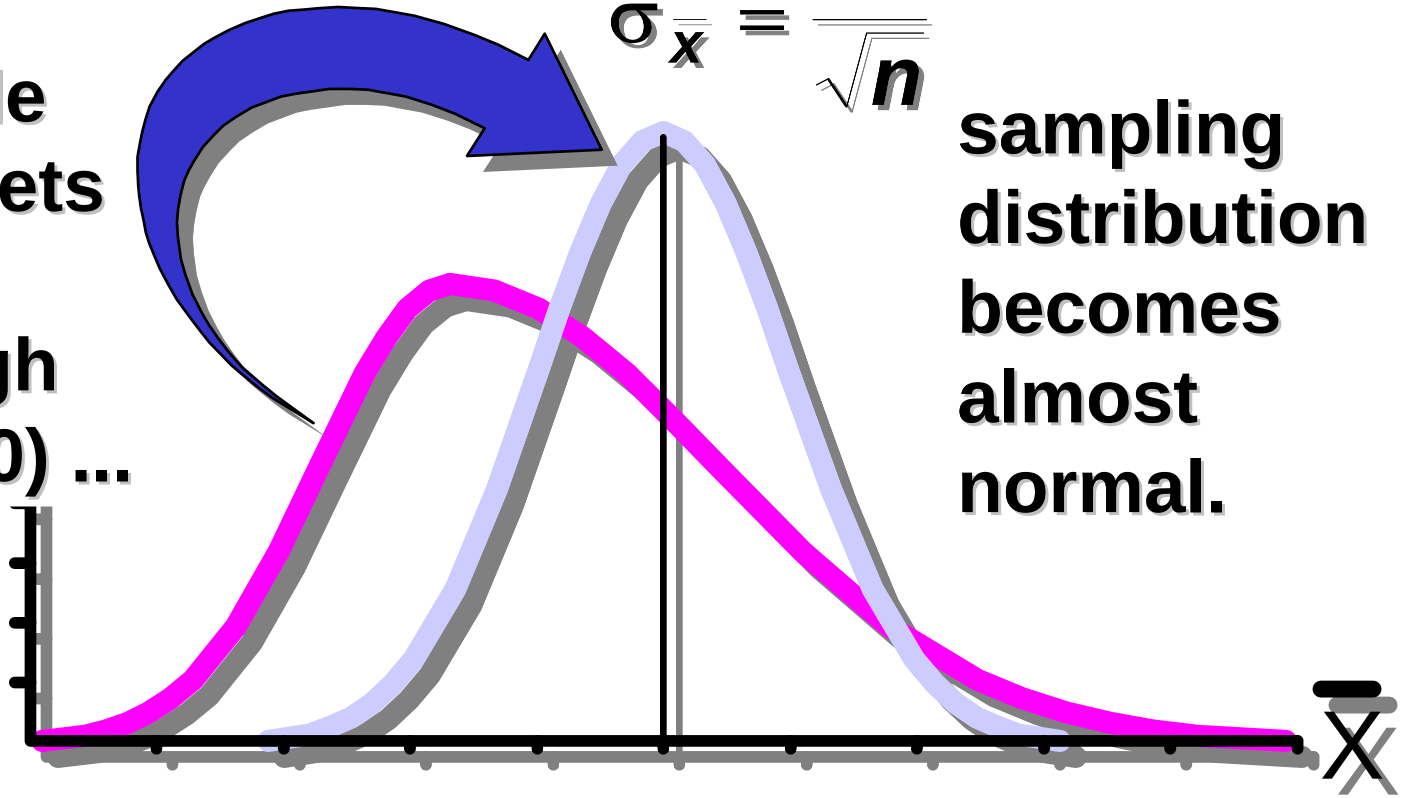
sampling distribution becomes almost normal.

Central Limit Theorem

As sample size gets large enough (n ≥ 30) ...

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

sampling distribution becomes almost normal.



$$\mu_{\bar{x}} = \mu$$

An example…

- Suppose that we want to compare the crime rate in San Diego with the crime rate in the rest of the country.
  - Is there more or less crime in San Diego than the national average?
  - We do not have the funds to measure the crime rate in every neighborhood in San Diego, so we take a random sample of streets from the entire population of streets in San Diego.

An example…

– We then calculate how much crime, on average, the streets in our sample have experienced and compare it to the national average.
– If the mean of our sample differs from the national average, what can we conclude?

An example…

- Ideally, if we wanted to know the value of a population parameter, we would gather all the observations in the population, and, using descriptive statistics, directly calculate the parameter.
- But if it is not feasible to measure the entire population, this is when inferential statistics is of use.

An example…

- The population of interest to us is how much crime, on average, *all* the neighborhoods of San Diego experience.
- From the study of our sample of streets, we want to draw an inference as to whether the mean of our population is the same as the national average.
- However, our sample of streets is just one of numerous random samples we could have drawn from our population of San Diego streets.

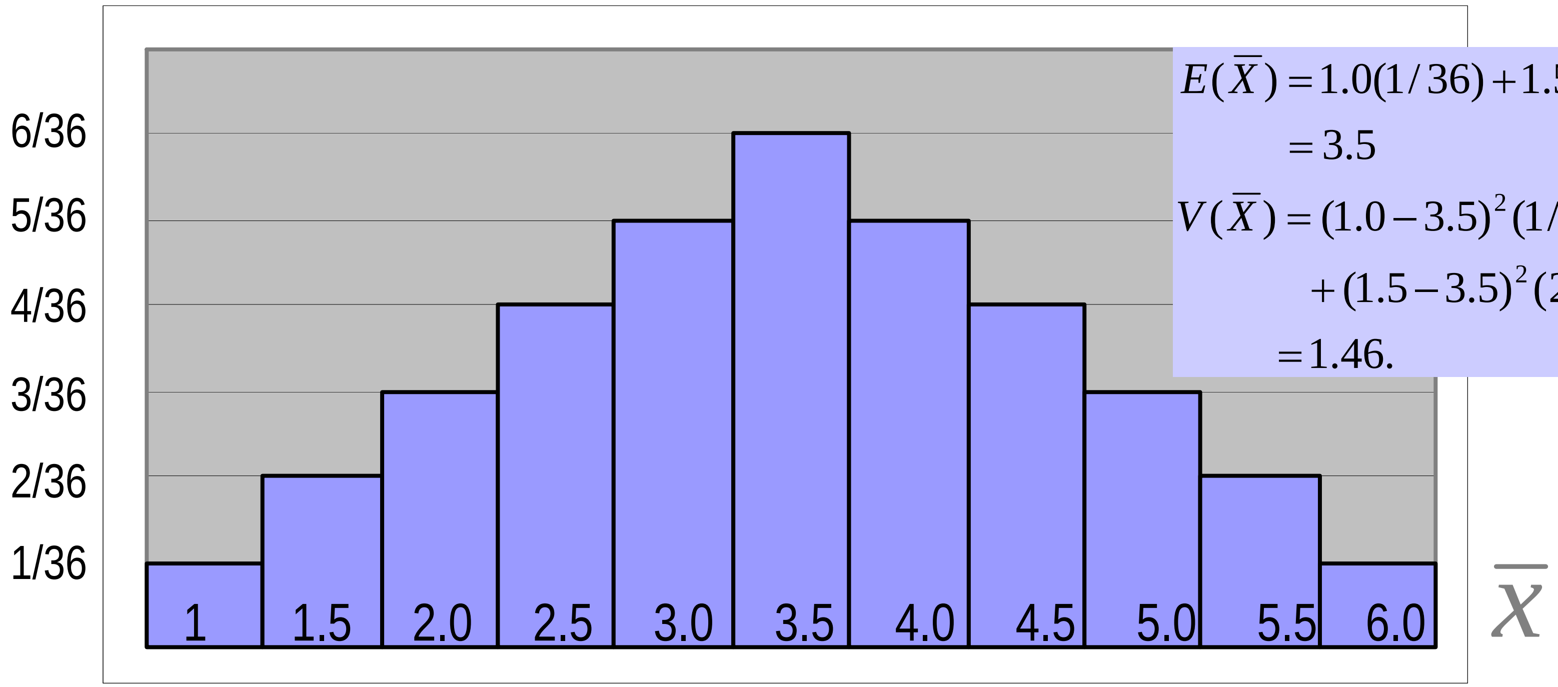
What do you compare your sample to?

- Random samples rarely represent the underlying population exactly.
- What is the probability of obtaining the sample mean that you did?
  - Compare your sample to other samples of the same size from the same population.
    - Sampling Distribution of the Mean

- Suppose we want to estimate μ from the mean $\bar{x}$ of a sample of size $n = 2$.
- What is the distribution that $\bar{x}$ will follow?

Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
4	1,4	2.5	16	3,4	3.5	28	5,4	4.5
5	1,5	3	17	3,5	4	29	5,5	5
6	1,6	3.5	18	3,6	4.5	30	5,6	5.5
7	2,1	1.5	19	4,1	2.5	31	6,1	3.5
8	2,2	2	20	4,2	3	32	6,2	4
9	2,3	2.5	21	4,3	3.5	33	6,3	4.5
10	2,4	3	22	4,4	4	34	6,4	5
11	2,5	3.5	23	4,5	4.5	35	6,5	5.5
12	2,6	4	24	4,6	5	36	6,6	6

Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
4	1,4	2.5	16	3,4	3.5	28	5,4	4.5
5	1,5	3	17	3,5	4	29	5,5	5
6	1,6	3.5	18	3,6	4.5	30	5,6	5.5
7	2,1	1.5	19	4,1	2.5	31	6,1	3.5
8	2,2	2	20	4,2	3	32	6,2	4
9	2,3	2.5	21	4,3	3.5	33	6,3	4.5
10	2,4	3	22	4,4	4	34	6,4	5
11	2,5	3.5	23	4,5	4.5	35	6,5	5.5
12	2,6	4	24	4,6	5	36	6,6	6

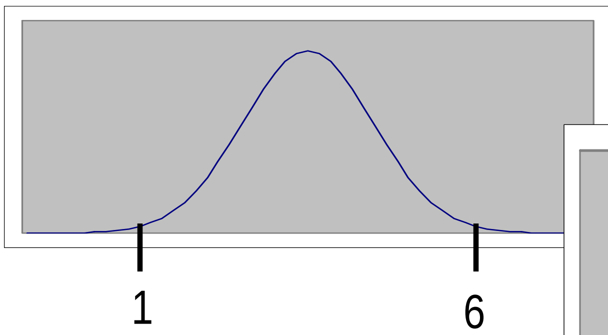


$$E(\bar{X}) = 1.0(1/36) + 1.5(2/36) + \ldots$$
$$= 3.5$$
$$V(\bar{X}) = (1.0 - 3.5)^2(1/36)$$
$$+ (1.5 - 3.5)^2(2/36) + \ldots$$
$$= 1.46.$$

$n = 5$

$\mu_{\bar{x}} = 3.5$

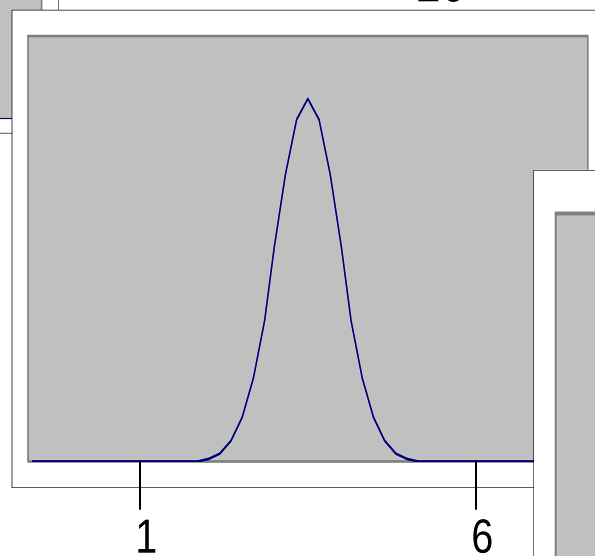
$\sigma^2_{\bar{x}} = .5833 \ (= \frac{\sigma^2_x}{5})$



$n = 10$

$\mu_{\bar{x}} = 3.5$

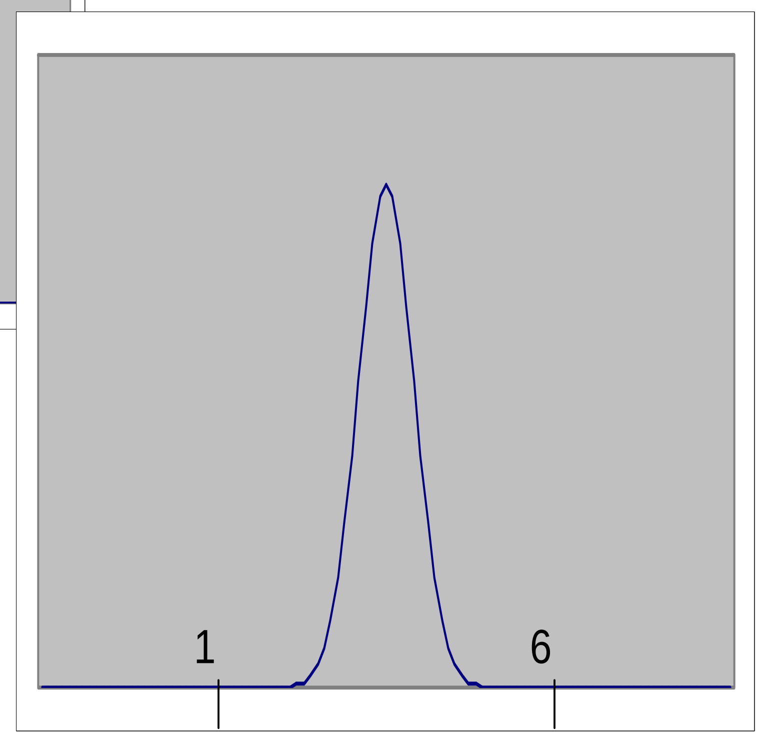
$\sigma^2_{\bar{x}} = .2917 \ (= \frac{\sigma^2_x}{10})$



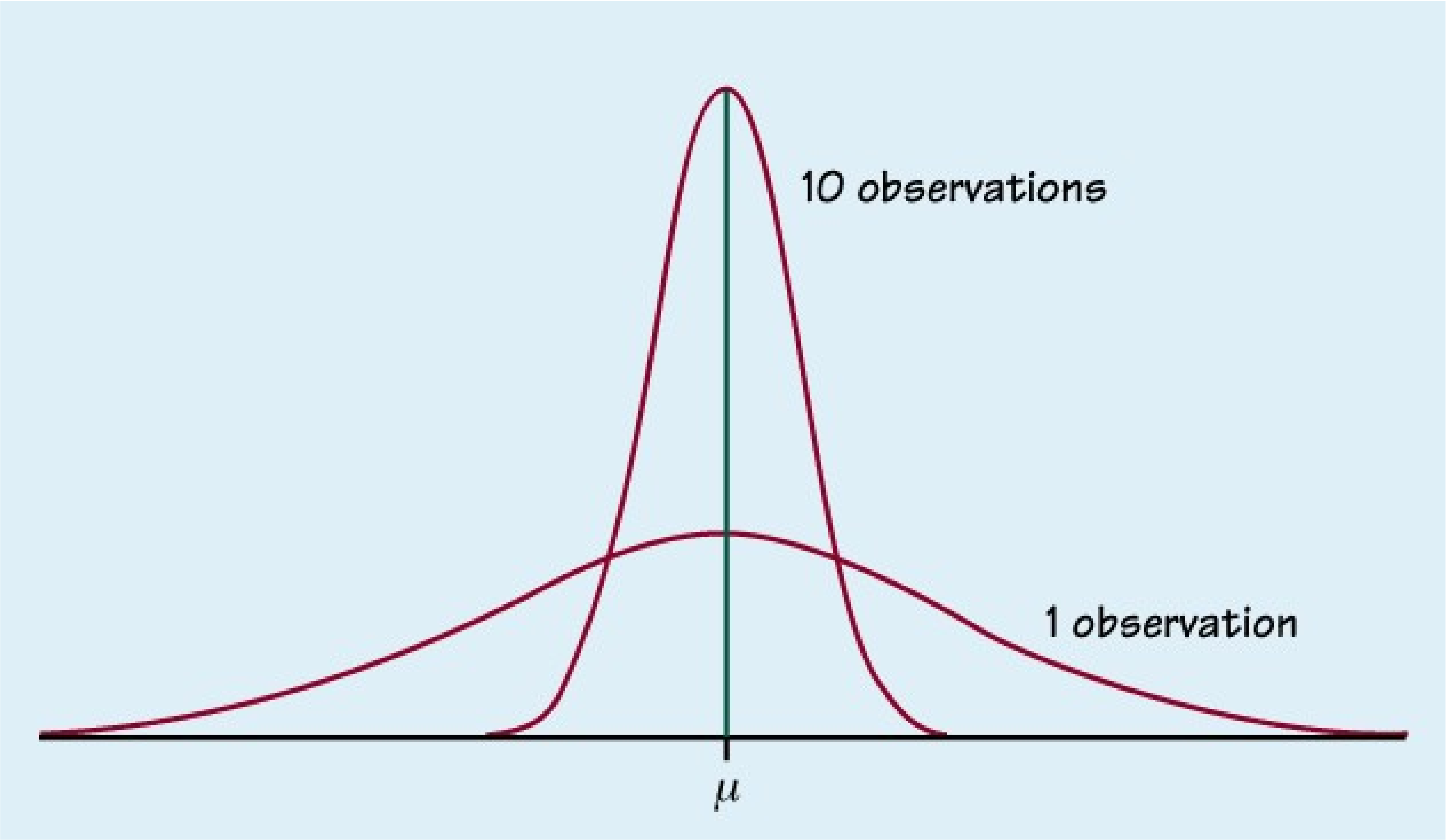
$n = 25$

$\mu_{\bar{x}} = 3.5$

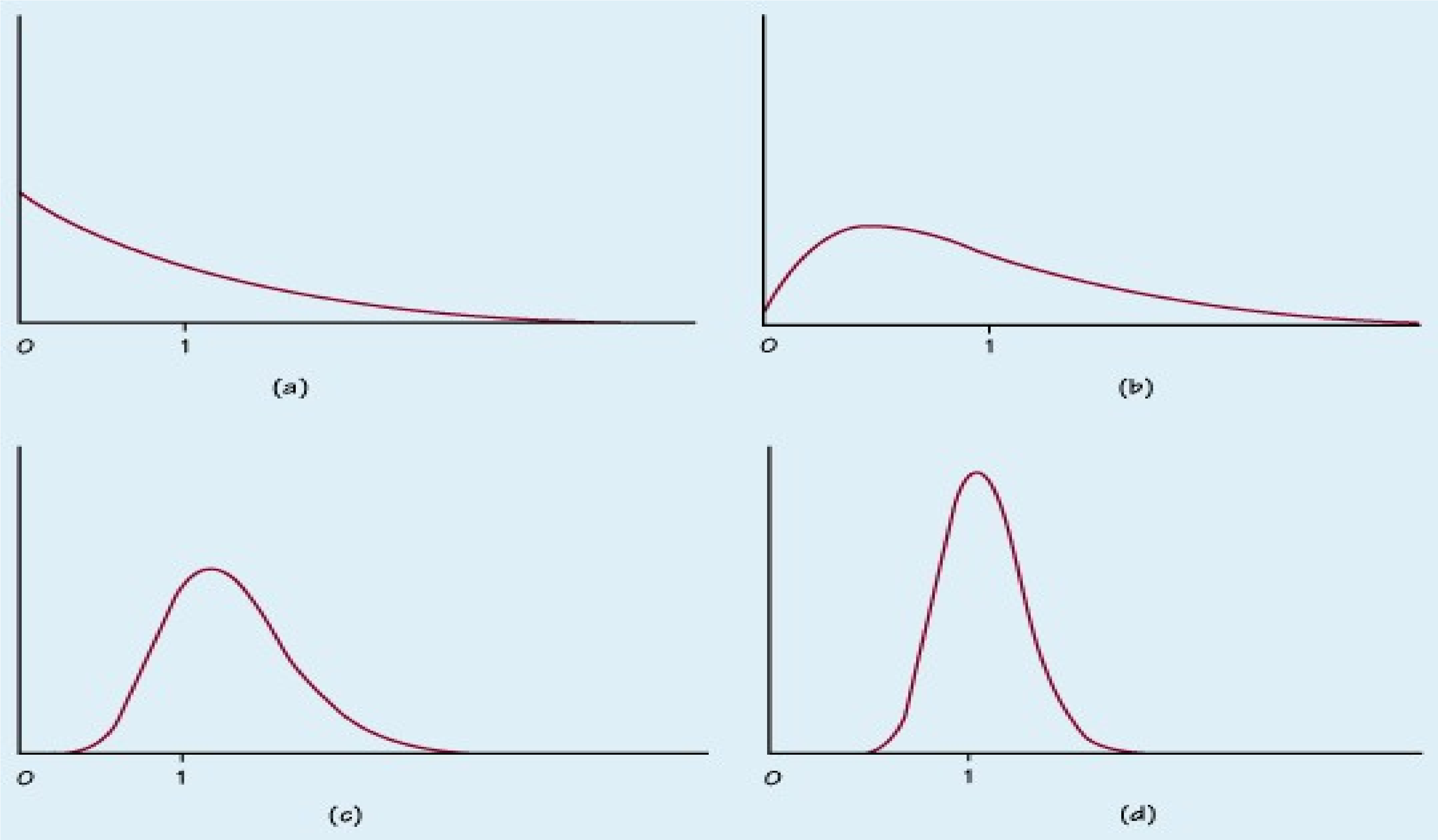
$\sigma^2_{\bar{x}} = .1167 \ (= \frac{\sigma^2_x}{25})$



Notice that $\sigma_{\bar{x}}$ is smaller than σ_x. The larger the sample size the smaller $\sigma_{\bar{x}}$. Therefore, $\bar{X}$ tends to fall closer to μ, as the sample size increases.

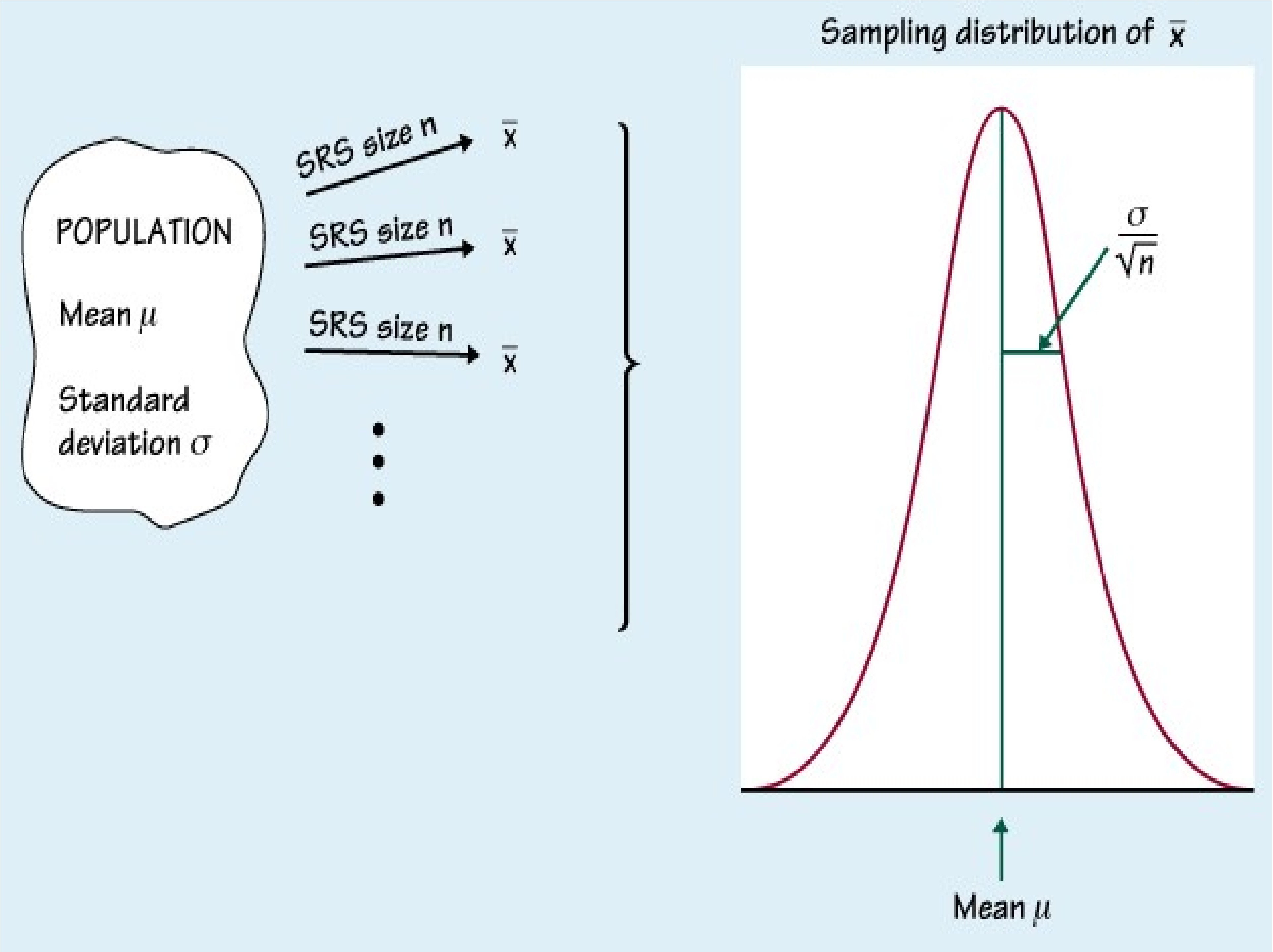


- The sampling distribution of $\overline{X}$ for samples of size 10 compared with the distribution of a single observation.



The distributions of $\bar{X}$ for

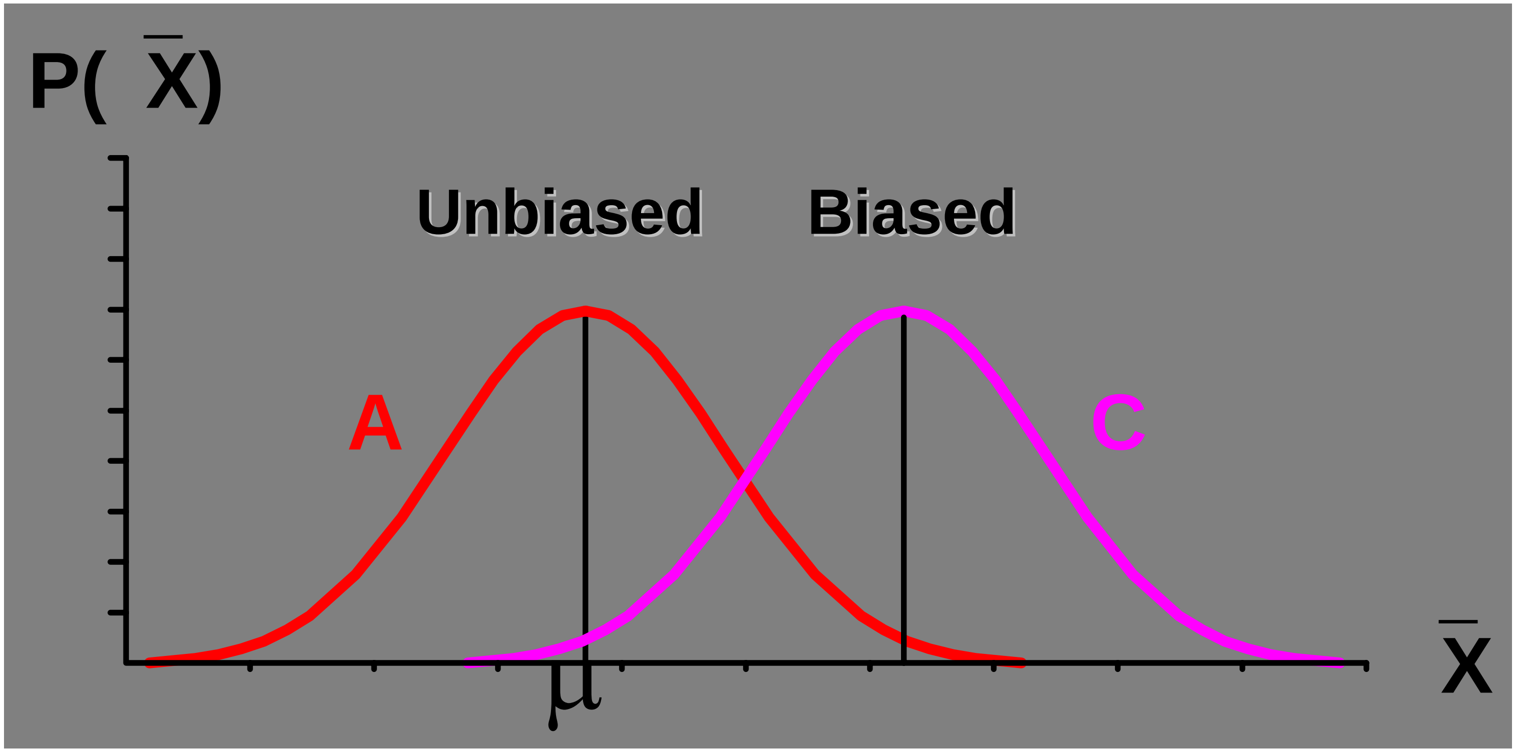
(a). 1 obs. (b). 2 obs. (c). 10 obs. (d). 25 obs.


Sampling distribution of $\bar{x}$
POPULATION
Mean μ
Standard deviation σ
SRS size n
SRS size n
SRS size n
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\frac{\sigma}{\sqrt{n}}$
Mean μ

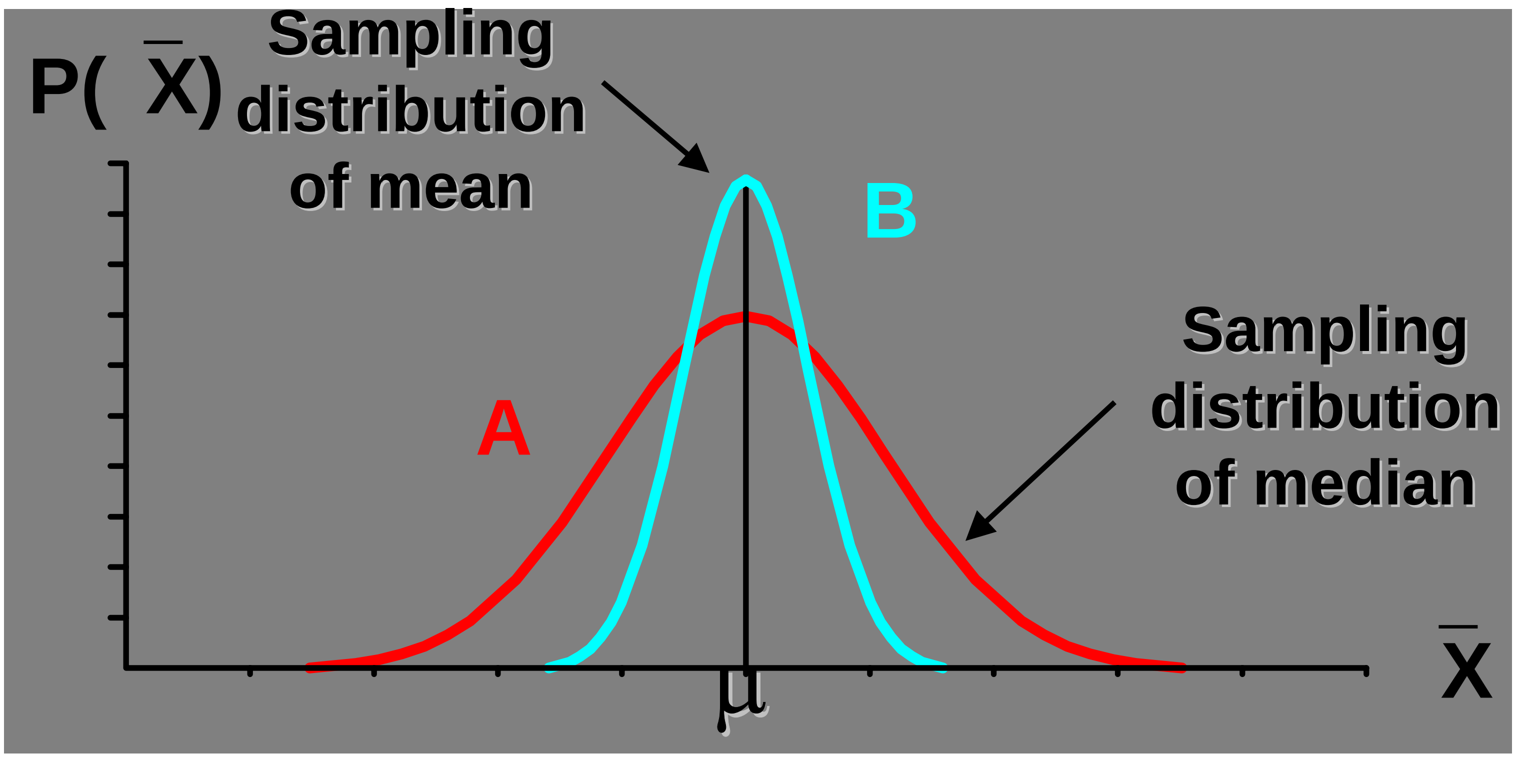
Properties of Sampling Distribution of Mean

- 1. Unbiasedness
  - Mean of Sampling Distribution Equals Population Mean
- 2. Efficiency (minimum variance)
  - Sample Mean Comes Closer to Population Mean Than Any Other Unbiased Estimator
- 3. Consistency
  - As Sample Size Increases, Variation of Sample Mean from Population Mean Decreases

Unbiasedness

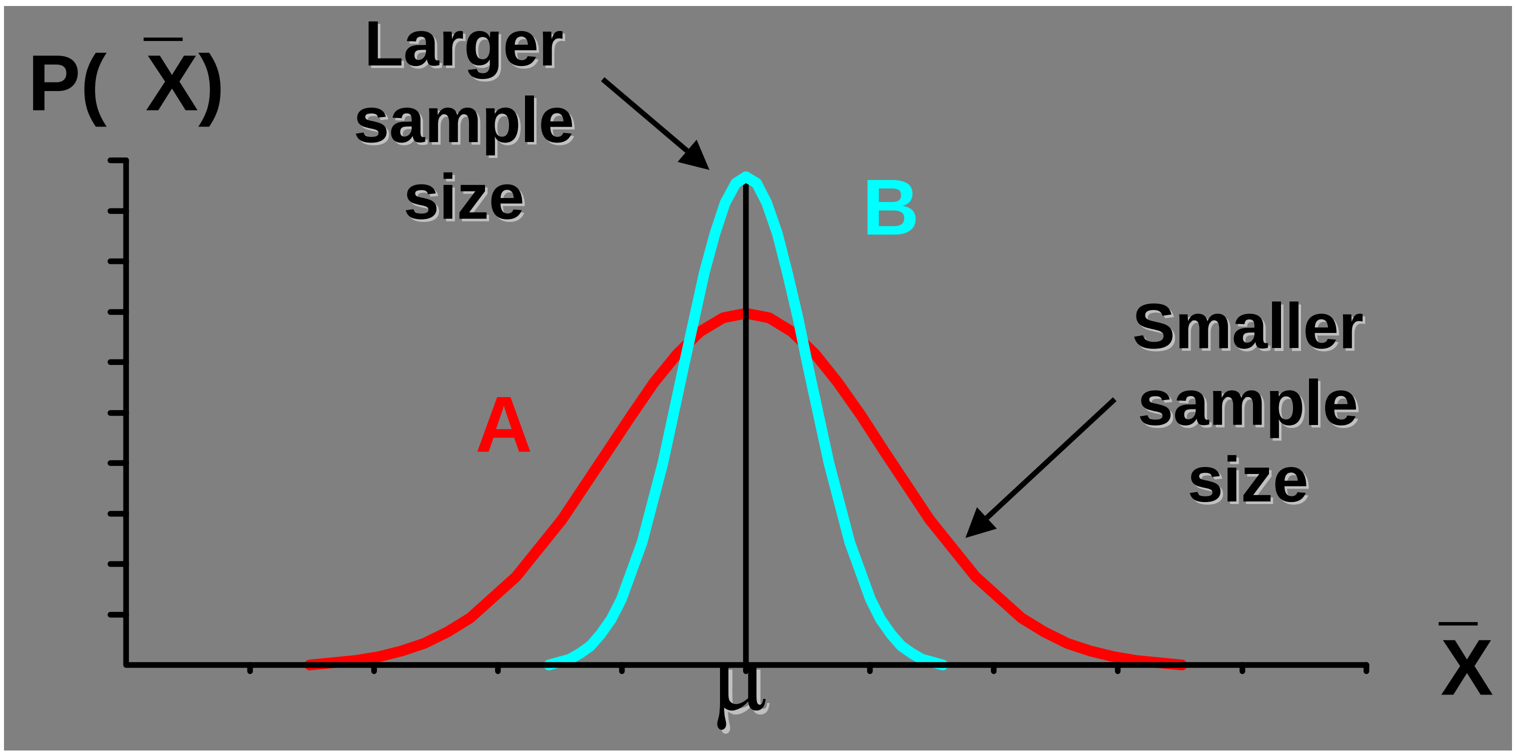


Efficiency



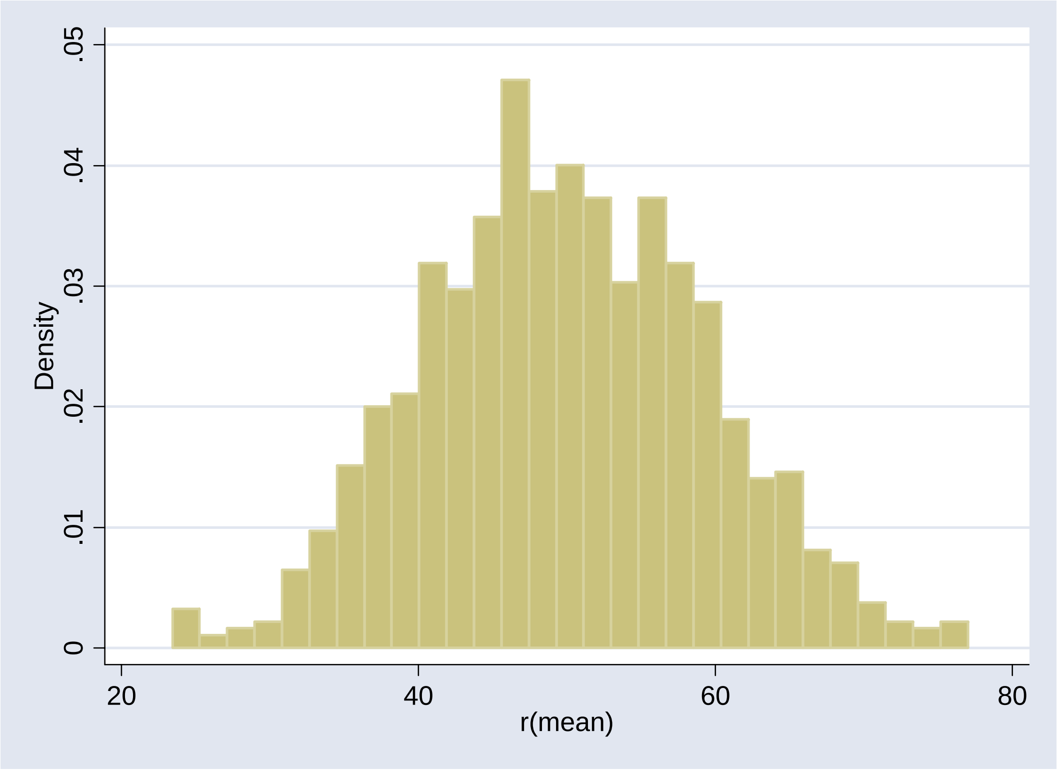
P(X̄)
Sampling distribution of mean
B
A
Sampling distribution of median
μ
X̄

Consistency



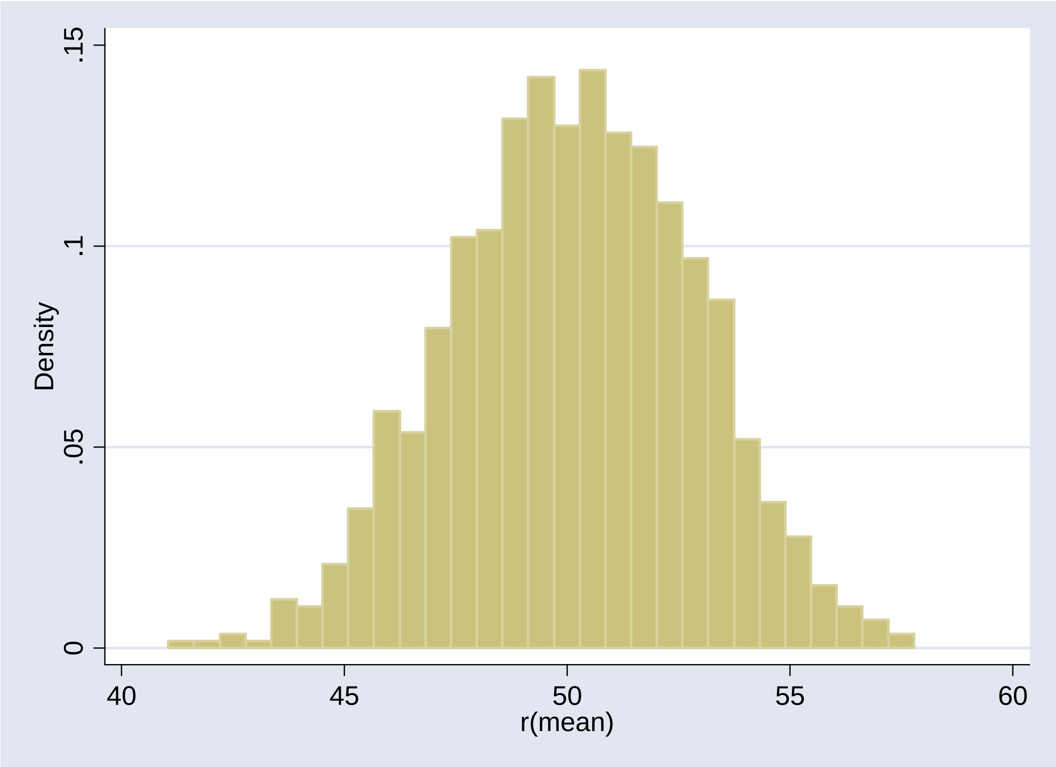
N=9 Simulation (1000 trials)

- Mean = 49.92; sigma = 9.43



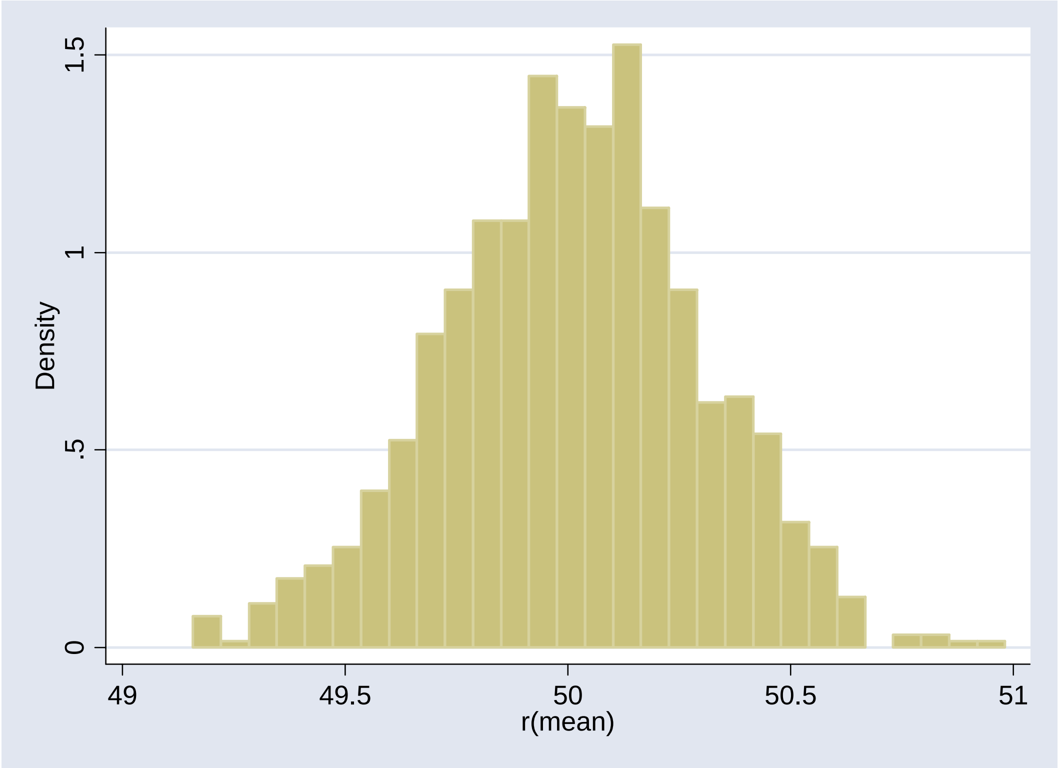
N = 100 Simulation (1000 trials)

- Mean = 50.15; sigma = 2.72



N = 10,000 Simulation (1000 trials)

- Mean = 50.01; sigma = .289



Example 6: Average Weight Loss

Weight-loss clinic interested in average weight loss for participants in its program. Weight losses assumed to be bell-shaped, so Rule applies for any sample size. ***Population*** is **all current and potential clients**, and ***measurement*** is **weight loss**.

Example 7: Average Age at Death

Researcher is interested in average age at which left-handed adults die, assuming they have lived to be at least 50. Ages at death not bell-shaped, so need at least 30 such ages at death. ***Population*** is **all left-handed people** who live to be at least 50 years old. The ***measurement*** is **age at death**.

Defining the Rule for Sample Means

If numerous samples or repetitions of the same size are taken, the frequency curve of means from various samples will be **approximately bell-shaped**.

Mean will be same as mean for the population.

Standard deviation will be:

$$\frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$

Example 9: Using Rule for Sample Means

Weight-loss example, population mean and standard deviation were 8 pounds and 5 pounds, respectively, and we were taking random samples of size 25.

Potential sample means represented by a **bell-shaped** curve with **mean of 8 pounds** and **standard deviation**:

$$\frac{5}{\sqrt{25}} = \mathbf{1\ pound}$$

For our sample of 25 people:

- **68% chance** sample mean is between 7 and 9 pounds
- **95% chance** sample mean is between 6 and 10 pounds
- **almost certain** sample mean is between 5 and 11 pounds

Increasing the Size of the Sample

Weight-loss example: suppose a sample of 100 people instead of 25 was taken.

Potential sample means still represented by a **bell-shaped** curve with **mean of 8 pounds** but **standard deviation**:

$$\frac{5}{\sqrt{100}} = \mathbf{0.5\ pounds}$$

For our sample of 100 people:

- **68% chance** sample mean is between 7.5 and 8.5 pounds
- **95% chance** sample mean is between 7 and 9 pounds
- **almost certain** sample mean is between 6.5 and 9.5 pounds

Larger samples tend to result in more accurate estimates of population values than do smaller samples.

19.4 What to Expect in Other Situations

- So far two common situations – (1) want to know what **proportion of a population** fall into one category of a ***categorical*** variable, (2) want to know the **mean of a population** for a ***measurement*** variable.
- Many other situations and similar rules apply to most other situations

Two Basic Statistical Techniques

- **Confidence Intervals**
  Interval of values the researcher is fairly sure covers the true value for the population.

- **Hypothesis Testing**
  Uses sample data to attempt to reject the hypothesis that nothing interesting is happening—that is, to reject the notion that chance alone can explain the sample results.

Case Study 19.1: Do Americans Really Vote When They Say They Do?

Reported in *Time* magazine (Nov 28, 1994)**:**

- Telephone **poll of 800** adults (2 days after election) – **56% reported they had voted**.
- Committee for Study of American Electorate stated only **39% of American adults had voted**.

Could it be the results of poll simply reflected a sample that, by chance, voted with greater frequency than general population?

Case Study 19.1: Do Americans Really Vote When They Say They Do?

Suppose only **39% of American adults voted**. We can expect sample proportions to be represented by a **bell-shaped** curve with **mean 0.39** and **standard deviation**:

$$\sqrt{\frac{(0.39)(1-0.39)}{800}} = \mathbf{0.017}$$

For our sample of 800 adults, we can be **almost certain** to see a sample proportion between 33.9% and 44.1%. The reported 56% is far above 44.1%.

The **standard score** for 56% is: (0.56 – 0.39)/0.017 = **10**. Virtually impossible to see a standard score of 10 or more.

For Those Who Like Formulas

$$\hat{p} \sim N(p, \frac{p(1-p)}{n}) \quad \text{and} \quad \overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Notation for Population and Sample Proportions

Sample size = n

Population proportion = p

Sample proportion = $\hat{p}$, which is read "p-hat" because the p appears to have a little hat on it.

The Rule for Sample Proportions

If numerous samples or repetitions of size n are taken, the frequency curve of the $\hat{p}$'s from the various samples will be approximately bell-shaped. The mean of those $\hat{p}$'s will be p. The standard deviation will be

$$\sqrt{\frac{p(1-p)}{n}}$$

Notation for Population and Sample Means and Standard Deviations

Population mean = μ (read "mu"), population standard deviation = σ (read "sigma")

Sample mean = $\overline{X}$, sample standard deviation = s

The Rule for Sample Means

If numerous samples of size n are taken, the frequency curve of the $\overline{X}$'s from the various samples is approximately bell-shaped with mean μ and standard deviation $\sigma/\sqrt{n}$.