AP Statistics Notes – Unit Seven: Sampling Distributions

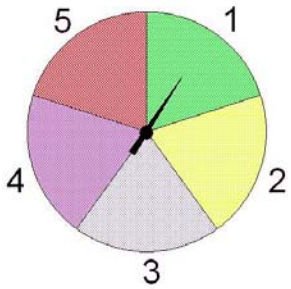
Syllabus Objectives: 2.3 – The student will distinguish between populations and samples. 2.4 – The student will distinguish between parameters and statistics. 4.2 – The student will discuss the properties of point estimators, including unbiasedness and variability. 3.21 – The student will simulate the sampling distribution of a random variable.

The inferential methods we will learn in the coming units will be based on using information from a sample to reach a conclusion about the population. In order to use this information, we must develop an understanding of how sampling information varies from sample to sample. In this unit, we will explore the behavior of sample statistics in repeated sampling and learn one of the most important theorems in Statistics – The Central Limit Theorem.

- **Statistics and Sampling Variability**
  - **Parameter** – A characteristic that is related to a population. A parameter is a number that describes the population. It is a fixed number, but in practice we do not know its value because we cannot examine the entire population.
  - The usual way to gain information about a parameter is to select a sample from the population. However, we must note that the sample information we gather may differ somewhat from the population characteristic we are trying to measure. Further, the sample information may differ from sample to sample. This sample-to-sample variability poses a problem when we try to generalize our findings to the population. In order to do so, we must gain an understanding of this variability.
  - **Statistic** – A quantity computed from the values in a sample. Values of statistics, such as sample means, sample medians, sample standard deviations or the proportion of individuals in a sample that possess a particular property are our primary sources of information about various population characteristics. We can view a sample statistic as a random variable. That is, we have no way of predicting *exactly* what statistic value we will get from a sample, but, given a population parameter, we know how those values will behave in repeated sampling.
  - The observed value of a statistic depends on the particular sample selected from the population; typically, it varies from sample to sample. This variability is called **sampling variability**.

- **Sampling Distributions**
  - **We need to understand why sampling variability is not fatal.** What would happen if we took many samples?
    1. Take a large number of samples from the sample population.
    2. Calculate the sample mean $\bar{x}$ or sample proportion $\hat{p}$ for each sample.
    3. Make a histogram of the values of $\bar{x}$ or $\hat{p}$.
    4. Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers or other deviations.
  - Of course, it is too expensive to take many samples from a population, but we can imitate this using simulation.
  - **Sampling distribution** – The distribution that would be formed by considering the value of a sample statistic for every possible different sample of a given size from a population. The sampling distribution is the ideal pattern that would emerge if we looked at all possible samples of the same size from our population.

- **Simulated example of a sampling distribution:** Consider a population that consists of the numbers 1, 2, 3, 4 and 5 generated in a manner that the probability of each of those values is 0.2 no matter what the previous selections were. This population could be described as the outcome associated with a spinner such as given below. The distribution is next to it.



x	p(x)
1	0.2
2	0.2
3	0.2
4	0.2
5	0.2

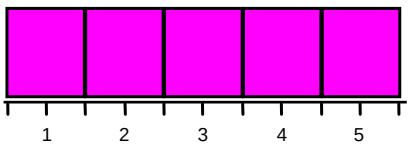
If the sampling distribution for the means of samples of size two is analyzed, it looks like the first table below. Every possible sample of two was taken from the population and the sample mean $\bar{x}$ was calculated for each sample. Then, the distribution of the 25 sample means is summarized in the second table.

Sample	
1, 1	1
1, 2	1.5
1, 3	2
1, 4	2.5
1, 5	3
2, 1	1.5
2, 2	2
2, 3	2.5
2, 4	3
2, 5	3.5
3, 1	2
3, 2	2.5
3, 3	3

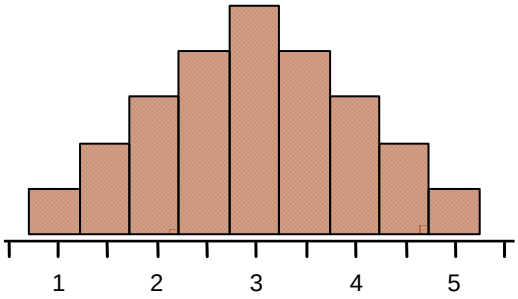
Sample	
3, 4	3.5
3, 5	4
4, 1	2.5
4, 2	3
4, 3	3.5
4, 4	4
4, 5	4.5
5, 1	3
5, 2	3.5
5, 3	4
5, 4	4.5
5, 5	5

	frequency	p(x)
1	1	0.04
1.5	2	0.08
2	3	0.12
2.5	4	0.16
3	5	0.20
3.5	4	0.16
4	3	0.12
4.5	2	0.08
5	1	0.04
	25	

The original population distribution and the sampling distribution of means of samples with $n = 2$ are summarized by the histograms below.

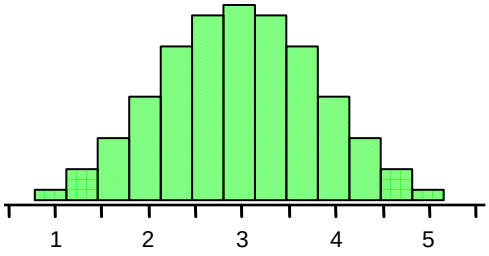


Original distribution

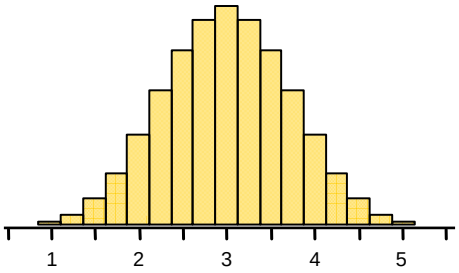


Sampling distribution
n = 2

Sampling distributions for *n* = 3 and *n* = 4 were also calculated and are illustrated below in the histograms.

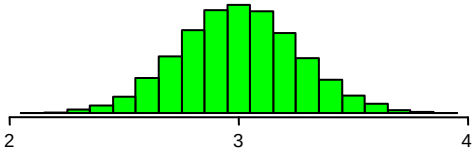


Sampling distribution n = 3

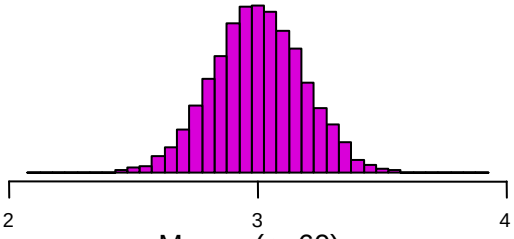


Sampling distribution n = 4

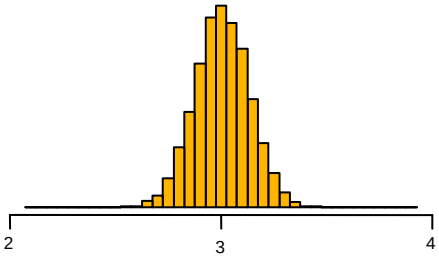
To illustrate the general behavior of samples of any fixed size n, 10,000 samples each of size 30, 60 and 120 were generated from this same uniform distribution and the means were calculated. Probability histograms were created for each of these simulated sampling distributions. Notice that all three of these look to be essentially normally distributed. Further, note that the variability decreases as the sample size increases.



Means (n=30)



Means (n=60)



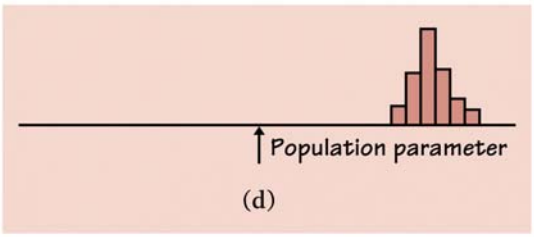
Means (n=120)

- **Describing sampling distributions**
  1. The overall shape (symmetric, skewed, uniform, bell-shaped, approximately normal, etc.)
  2. Are there any outliers or other important deviations from the overall patterns?
  3. Describe the center of the distribution.
  4. Describe the spread (standard deviation).
  5. Haphazard sampling does not give such regular and predictable results. However, when randomization is used, statistics computed from the data have a definite pattern of behavior over many repetitions, even though the result of a single repetition is uncertain.

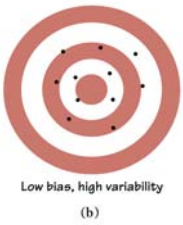
- **Bias of a statistic**
  - We have no way of knowing whether or not our statistic value is equal to the parameter we are trying to estimate. How trusty is a statistic as an estimator of a parameter? **Bias** concerns the center of the sampling distribution. Bias means the center of the sampling distribution is not equal to the true value of the parameter.
  - A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is **equal** to the true value of the parameter being estimated.

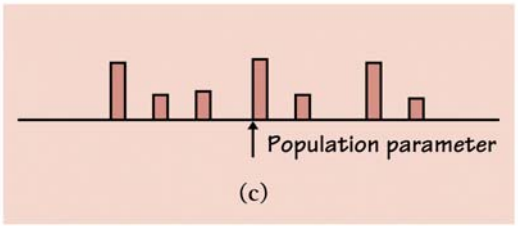
- **Variability of a statistic**
  - The variability of a statistic is described as the spread of its sampling distribution. This spread is determined by the sampling design and the size of the sample. Larger samples give smaller spread. Variability of sample results is controlled by the size of the sample, and not the size of the population. A statistic from an SRS of size 2500 from more than 280,000,000 residents is just as precise as an SRS of size 2500 from 750,000 residents. Larger samples will give us less variability, so this has to be considered when designing a sample. As long as the population is much larger than the sample (at least 10 times as large), the spread of the sampling distribution is approximately the same for any population size.
  - **Examples of bias and variability:** The bulls-eyes and histograms below illustrate bias and variability. Bias means that our aim is off (not hitting the center), whereas high variability means that repeated shots are widely scattered. Properly chosen statistics computed from random samples of sufficient size will have low bias and low variability (which is good!). Remember, randomization helps reduce bias. Stratifying or blocking and larger same sizes help reduce variability.

$(a) \rightarrow (d)$

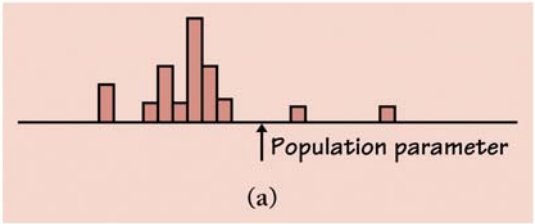


$(b) \rightarrow (c)$

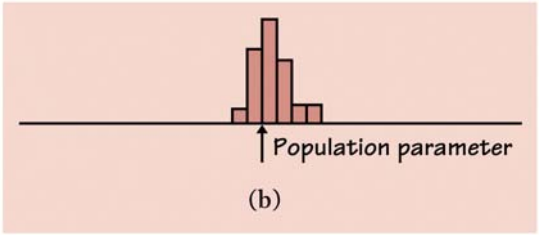




$(c) \rightarrow (a)$



$(d) \rightarrow (b)$



Syllabus Objectives: 3.15 – The student will analyze sampling distributions of a sample proportion. 3.21 – The student will simulate the sampling distribution of a random variable.

- **Sampling distribution of a sample proportion**
  - **Sample proportion –** it is the proportion of successes in the sample. $\hat{p} = \frac{\text{count of "success" in sample}}{\text{size of sample}} = \frac{X}{n}$. Since both *X* and $\hat{p}$ will vary in repeated samples, both are random variables. Also, $\hat{p}$ is an unbiased estimator of the population parameter, *p*.

  - **Sampling distribution of $\hat{p}$** – Choose an SRS of size *n* from a large population with population proportion *p* having some characteristic of interest. Let $\hat{p}$ be the proportion of the sample having that characteristic and denote the mean value of $\hat{p}$ by $\mu_{\hat{p}}$ and the standard deviation by $\sigma_{\hat{p}}$. Then, the following rules hold:
    - The **mean** of the sampling distribution is exactly *p*. $\mu_{\hat{p}} = p$
    - The **standard deviation** of the sampling distribution is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.
      - ◊ Use the recipe for the standard deviation of $\hat{p}$ only when the population is at least 10 times as large as the sample. This will be referred to as Rule of Thumb 1.
    - When ***n*** is large and ***p*** is not too near 0 or 1, the sampling distribution of $\hat{p}$ is approximately normal. To assure this, check the following rule of thumb:
      - ◊ We will use the normal approximation to the sampling distribution of $\hat{p}$ for values of *n* and *p* that satisfy both $np \geq 10$ and $n(1-p) \geq 10$. This will be referred to as Rule of Thumb 2.

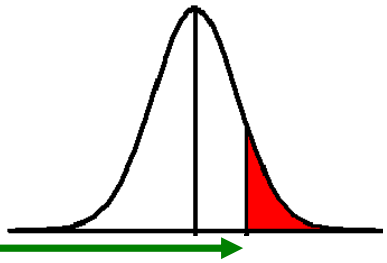
    Thus, the sampling distribution of $\hat{p}$ is always centered at the value of the population success proportion, *p* , and the extent to which the distribution spreads out about *p* decreases as the sample size *n* increases.

  - **Sampling distribution of $\hat{p}$ Example 1** – If the true proportion of defectives produced by a certain manufacturing process is 0.08 and a sample of 400 is chosen, what is the probability that the proportion of defectives in the sample is greater than 0.10?
    - **Solution:** We must first check our rule of thumbs. We will assume that the population > $10 \cdot n > 10 \cdot 400 > 4000$, which means Rule of Thumb 1 is satisfied and we may use the formula to find the standard deviation. Also, since $np = 400(0.08) = 32 \geq 10$ and $n(1-p) = 400(0.92) = 368 \geq 10$, Rule of Thumb 2 is satisfied and it is reasonable to use the normal approximation. Since this is a normal distribution, we will find the mean and standard deviation using the formulas above and then use z-scores to find the probability.
    - $u_{\hat{p}} = p = 0.08$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(0.92)}{400}} = 0.013565$

- We are interested in the probability that our $\hat{p}$ is greater than 0.10, or $P(\hat{p} > 0.1)$. Since we have $N(0.08, 0.013565)$,

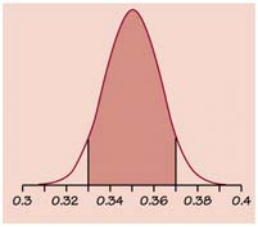
$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.10 - 0.08}{0.013565} = 1.47$$

$$P(\hat{p} > 0.1) = P(z > 1.47)$$
$$= 1 - 0.9292 = 0.0708$$



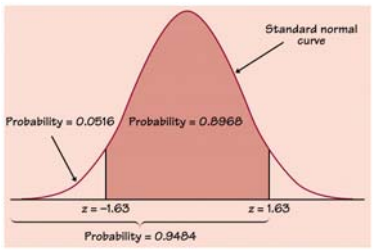
- **Sampling distribution of $\hat{p}$ Example 2** – A polling organization asks an SRS of 1500 first-year college students whether they applied for admission to any other college. In fact, 35% of all first-year students applied to colleges besides the one they are attending. What is the probability that the random sample of 1500 students will give a result within 2 percentage points of this true value?

  - **Solution:** We must first check our rule of thumbs. We have an SRS of size *n* =1500 drawn from a population in which the proportion *p* = 0.35 applied to other colleges. By the first "rule of thumb", the population must contain at least 10(1500) = 15,000 people for us to use the standard deviation formula. There are over 1.7 million first-year college students, so we are okay. We can use a normal approximation because $np = 1500(0.35) = 525 \geq 10$ and $n(1-p) = 1500(0.65) = 975 \geq 10$, and our "second rule of thumb" is satisfied.
  - $u_{\hat{p}} = p = 0.35$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.35(0.65)}{1500}} = 0.0123$
  - We want to find the probability that $\hat{p}$ falls within 2 percentage points, or 0.02 of 0.35 and this is a normal distribution calculation.
  - $N(0.35, 0.0123)$ and find $P(0.33 \leq \hat{p} \leq 0.37)$



$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.33 - 0.35}{0.0123} = -1.63$$

  - $$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.37 - 0.35}{0.0123} = 1.63$$
  - $P(0.33 \leq \hat{p} \leq 0.37) = P(-1.63 \leq z \leq 1.63) = 0.9484 - 0.0516 = 0.8968$



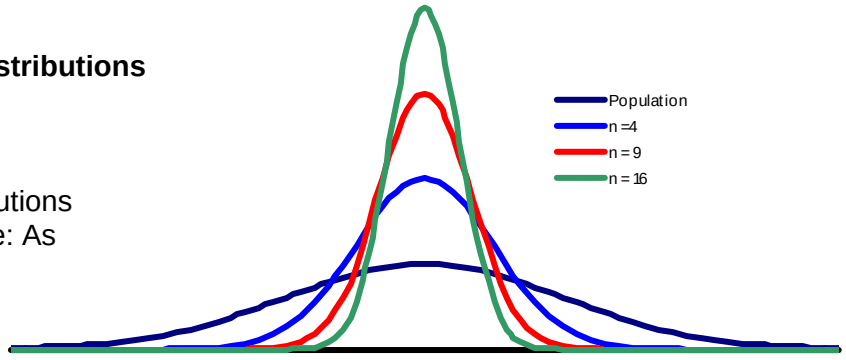
Syllabus Objectives: 3.16 – The student will analyze sampling distributions of a sample mean. 3.17 – The student will describe the properties of the central limit theorem. 3.18 – The student will solve problems using the central limit theorem.

- **Sampling distribution of a sample mean**
  - **Sample mean –** it is the arithmetic average of the sample. $\bar{x} = \frac{\sum x}{n}$.

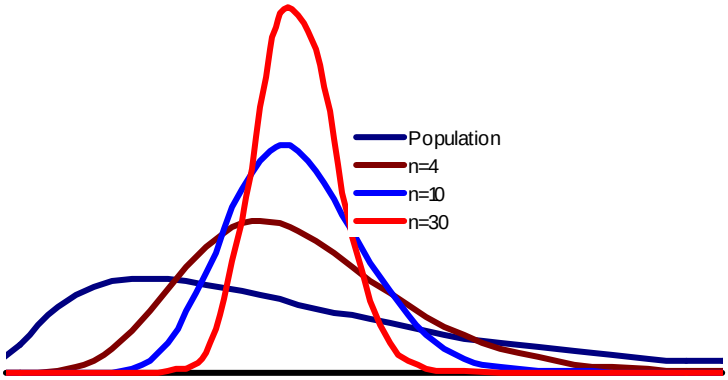
    Because sample means are just averages of observations, they are among the most common statistics. Two facts contribute to the popularity of sample means in statistical inference: averages are less variable and are more normal than individual observations. Since both *X* and $\bar{x}$ will vary in repeated samples, both are random variables. Also, $\bar{x}$ is an unbiased estimator of the population parameter, μ.

  - **Sampling distribution of** $\bar{x}$ – Suppose that $\bar{x}$ is the mean of an SRS of size *n* drawn from a large population with mean μ and standard deviation σ. Let us denote the mean value of $\bar{x}$ by $\mu_{\bar{x}}$ and the standard deviation by $\sigma_{\bar{x}}$. Then, the following rules hold:
    - The **mean** of the sampling distribution is exactly μ. $\mu_{\bar{x}} = \mu$
    - The **standard deviation** of the sampling distribution is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
      - ◊ Use the recipe for the standard deviation of $\bar{x}$ only when the population is at least 10 times as large as the sample. This will be referred to as Rule of Thumb 1.
    - When the population distribution is normal, the sampling distribution of $\bar{x}$ is also normal for any sample size *n*. However, in most situations, the shape of the population distribution is unknown and we need the following rule.
    - When *n* is sufficiently large, the sampling distribution of $\bar{x}$ is approximately normally distributed, even when the population distribution is not itself normal. This is known as the **central limit theorem**.
      - ◊ More about the central limit theorem (CLT): What is sufficiently large? The Central Limit Theorem can safely be applied when *n* exceeds 30. Some books go as high as 40, some as low as 20, but 30 is a nice conservative number. If *n* > 30, then the standardized variable $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ has approximately a standard normal (z) distribution.

- **Illustrations of the sampling distributions**
  - A normal population is shown at the right. No matter what the sample size, the sampling distributions are approx. normal. Note: As sample size increases, variability decreases.



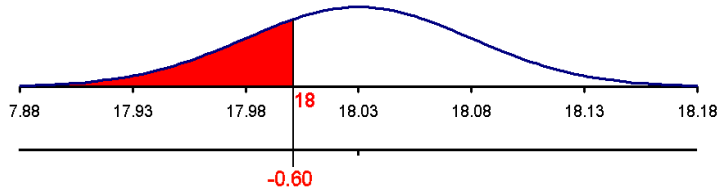
- A skewed distribution is shown at the right. For small sample sizes, the sampling distribution is still skewed, however, as stated in the CLT, as the sample size increases, the sampling distribution becomes approximately normal.



Thus, the sampling distribution of $\bar{x}$ is always centered at the true value of the population mean, μ, and the extent to which the distribution spreads out about μ decreases as the sample size *n* increases.

- **Sampling distribution of $\bar{x}$ Example 1 -** A food company sells "18 ounce" boxes of cereal. Let x denote the actual amount of cereal in a box of cereal. Suppose that x is normally distributed with $\mu = 18.03$ ounces and $\sigma = 0.05$. What proportion of boxes will contain less than 18 ounces?

  - **Solution:** We must first check our rules. Rule of Thumb 1 is satisfied because we can assume the population is greater than ten times the sample size and clearly, $n = 1$, so population > 10. Also, it is stated in the problem that our population is normally distributed, so our sampling distribution with be approximately normal.
  - $\mu_{\bar{x}} = \mu = 18.03$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.05}{\sqrt{1}} = 0.05$.
  - $$P(x < 18) = P\left(z < \frac{18 - 18.03}{0.05}\right)$$
    $$= P(z < -0.60) = 0.2743$$



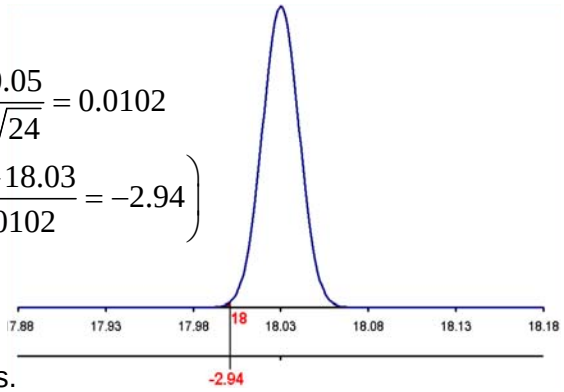
There is a 27.5% chance that the box will contain less than 18 ounces.

- Part 2: A case consists of 24 boxes of cereal. What is the probability that the mean amount of cereal (per box in a case) is less than 18 ounces?

- **Solution:** The difference is now our sample size is 24 and we are interested in an average. We now have to assume that the population is greater than 240. This assumption is not difficult to make. Even though our sample size is less than 30 and the Central Limit Theorem (CLT) does not apply, our original population is normally distributed, so our sampling distribution will be approximately normal.

- $\mu_{\bar{x}} = \mu = 18.03$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0.05}{\sqrt{24}} = 0.0102$
- $P(\bar{x} < 18) = P\left(z < \dfrac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \dfrac{18 - 18.03}{0.0102} = -2.94\right)$
- $P(z < -2.94) = 0.0016$

  There is only a .16% chance that a sample of 24 boxes would have a mean less than 18 ounces.



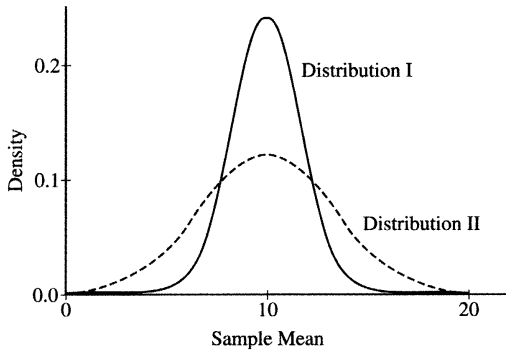
- **Sampling distribution of $\bar{x}$ Example 2 -** A hot dog manufacturer asserts that one of its brands of hot dogs has an average fat content of $\mu = 18$ g per hot dog. Consumers of this brand would probably not be disturbed if the mean is less than 18 but would be unhappy if it exceeds 18. Let x denote the fat content of a randomly selected hot dog, and suppose that σ, the standard deviation of the x distribution, is 1. An independent testing organization is asked to analyze a random sample of 36 hot dogs. Let $\bar{x}$ be the average fat content for this sample. Suppose that the sample resulted in a mean of $\bar{x} = 18.4$ g. Does this result suggest that the manufacturer's claim is incorrect?

  - **Solution:** Again, we must check that the first "rule of thumb" is satisfied. We can assume that the population of hot dogs is greater than ten times the sample size or pop > $10 \cdot 36 > 360$. Since this rule is satisfied, we can find the standard deviation. It is not stated in the problem that the population is normally distributed, so we have to look at the sample size. The sample size of 36 is large enough to rely on the Central Limit Theorem and to regard the $\bar{x}$ sampling distribution as approximately normal.
  - $\mu_{\bar{x}} = \mu = 18$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{1}{\sqrt{36}} = 0.1667$.
  - If the company's claim is correct, how likely is it that we would see a sample mean at least as large as 18.4 when the population mean is really 18? We will find $P(\bar{x} \geq 18.4)$.
  - $P(\bar{x} \geq 18.4) = P\left(z \geq \dfrac{18.4 - 18}{0.1667} \geq 2.40\right)$. This is the area under the *z* curve to the right of 2.40 ≈ 0.0082. Values of $\bar{x}$ at least as large as 18.4 will be observed only approximately 0.82% of the time when a random sample of size 36 is taken from a population with mean 18 and standard deviation 1. The value $\bar{x}$ = 18.4 exceeds 18 by enough to cast substantial doubt on the manufacturer's claim.

- **Sampling distribution questions found on previous AP exams**
  - **Example 1:** The graphs of the sampling distributions, I and II, of the sample mean of the same random variable for samples of two different sizes are shown below. Which of the following statements must be true about the sample sizes?

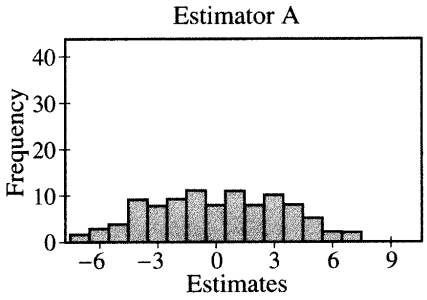
    (A) The sample size of I is less than the sample size of II.
    (B) The sample size of I is greater than the sample size of II.

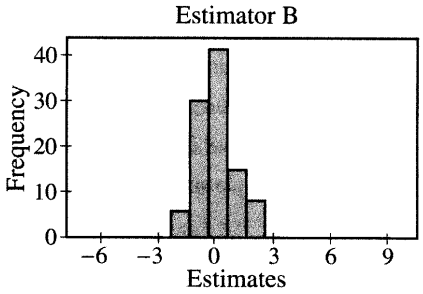
(C) The sample size of I is equal to the sample size of II.
(D) The sample size does not affect the sampling distribution.
(E) The sample sizes cannot be compared based on these graphs.

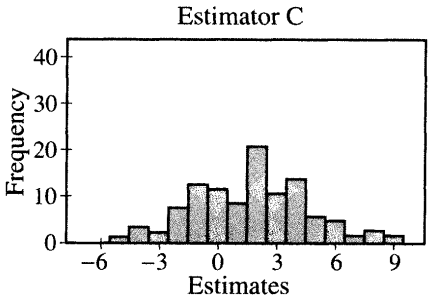


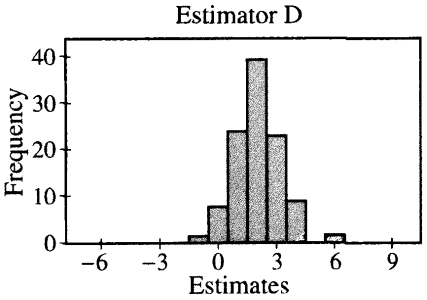
- **Solution:** The answer is B. As the sample size increases, the variability decreases. Smaller variability is shown in the density curve of Distribution I.

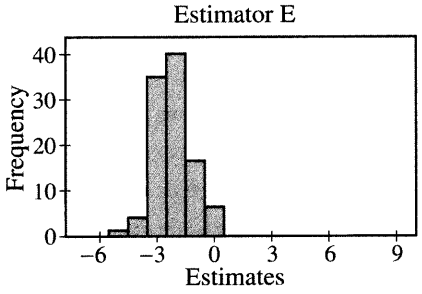
- **Example 2:** Five estimators for a parameter are being evaluated. The true value of the parameter is 0. Simulations of 100 random samples, each of size *n*, are drawn from the population. For each simulated sample, the five estimates are computed. The histograms below display the simulated sampling distributions for the five estimators. Which simulated sampling distribution is associated with the best estimator for this parameter?

(A)


(B)


(C)


(D)


(E)


- **Solution:** The answer is B. We are looking for the distribution with the smallest bias and smallest variability. Note that Estimator B not only has a small spread (from -3 to 3) but is also centered at the true parameter – zero.

- **Example 3:** A volunteer for a mayoral candidate's campaign periodically conducts polls to estimate the proportion of people in the city who are planning to vote for this candidate in the upcoming election. Two weeks before the election, the volunteer plans to double the sample size in the polls. The main purpose of this is to

  (A) reduce nonresponse bias
  (B) reduce the effects of confounding variables
  (C) reduce bias due to the interviewer effect
  (D) decrease the variability in the population
  (E) decrease the standard deviation of the sampling distribution of the sample proportion

  - **Solution:** The answer is E. When we increase the sampling size, the standard deviation of the $\hat{p}$ distribution decreases: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

- **Example 4:** The population {2, 3, 5, 7} has mean μ = 4.25 and standard deviation σ = 1.92. When sampling with replacement, there are 16 different possible ordered samples of size 2 that can be selected from this population. The mean of each of these 16 samples is computed. For example, 1 of the 16 samples is (2, 5), which has a mean of 3.5. The distribution of the 16 sample means has its own mean $\mu_{\bar{x}}$ and its own standard deviation $\sigma_{\bar{x}}$. Which of the following statements is true?

  (A) $\mu_{\bar{x}}$ = 4.25 and $\sigma_{\bar{x}}$ = 1.92
  (B) $\mu_{\bar{x}}$ = 4.25 and $\sigma_{\bar{x}}$ > 1.92
  (C) $\mu_{\bar{x}}$ = 4.25 and $\sigma_{\bar{x}}$ < 1.92
  (D) $\mu_{\bar{x}}$ > 4.25
  (E) $\mu_{\bar{x}}$ < 4.25

  - **Solution:** The answer is C. Using the formulas for the sampling distribution of $\bar{x}$, we know that $\mu_{\bar{x}} = \mu = 4.25$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.92}{\sqrt{2}} < 1.92$.

- **Example 5:** Big Town Fisheries recently stocked a new lake in a city park with 2,000 fish of various sizes. The distribution of the lengths of these fish is approximately normal.

  (a) Big Town Fisheries claims that the mean length of the fish is 8 inches. If this claim is true, which of the following would be more likely?

- A random sample of 15 fish having a mean length that is greater than 10 inches

or

- A random sample of 50 fish having a mean length that is greater than 10 inches

Justify your answer.

- **Solution:** The random sample of *n* =15 fish is more likely to have a sample mean length greater than 10 inches. The sampling distribution of the sample mean $\overline{x}$ is normal with mean $\mu = 8$ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Thus, both sampling distributions will be centered at 8 inches, but the sampling distribution of the sample mean when *n* =15 will have more variability than the sampling distribution of the sample mean when *n* =50. The tail area $(\overline{x} > 10)$ will be larger for the distribution that is less concentrated about the mean of 8 inches, which occurs when the sample size is *n* =15.

(b) Suppose the standard deviation of the sampling distribution of the sample mean for random samples of size 50 is 0.3 inch. If the mean length of the fish is 8 inches, use the normal distribution to compute the probability that a random sample of 50 fish will have a mean length less than 7.5 inches.

- **Solution:** $P(\overline{x} < 7.5) = P\left(z < \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} < \frac{7.5 - 8}{0.3} \right) = P\left(z < -1.67 \right) = 0.0475$

(c) Suppose the distribution of fish lengths in this lake was nonnormal but had the same mean and standard deviation. Would it still be appropriate to use the normal distribution to compute the probability in part (b)? Justify your answer.

- **Solution:** Yes. The Central Limit Theorem says that the sampling distribution of the sample mean will become approximately normal as the sample size *n* increases. Since the sample size is reasonably large (*n* = 50), the calculation in part (b) will provide a good approximation to the probability of interest even though the population is nonnormal.