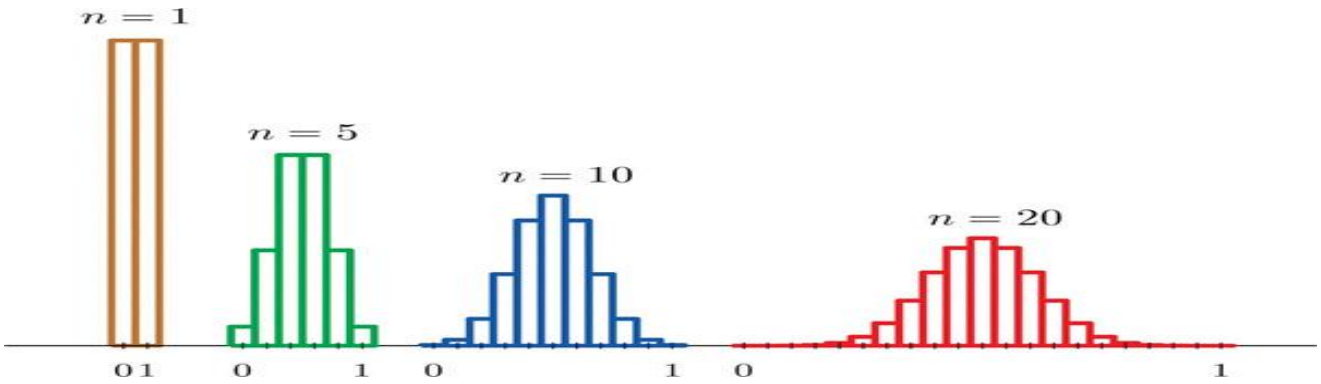
Introductory Statistics: A Problem-Solving Approach

by Stephen Kokoska

Chapter 7

Sampling Distributions



Parameters and Statistics

- A **parameter** is a numerical descriptive measure of a population. We usually cannot measure a parameter; it is an unknown constant that we would like to estimate.
- A **statistic** is any quantity computed from values in a sample. There are infinitely many quantities we could compute using the data in a sample.
- **Examples:** Identify the **boldface** number as a parameter or a statistic.

(a) In a recent survey of young adults, **66%** of millennials said they have "always believed the world is round." **Statistic**

(b) A spokesperson for Google reported that the proportion of all people working for the company who are women is **0.31**. **Parameter**

(c) The U.S. Department of Transportation recently reported that the mean age of all highway bridges in the United States is **42** years. **Parameter**

(d) The manager of a large hotel located near Disney World indicated that 20 selected guests had a mean length of stay equal to **5.6** days. **Statistic**

(e) In Canada's World Survey 2018, **21%** of those who responded said the environment was the most important world issue. **Statistic**

Sampling Variability

- Any statistic is a random variable, because it differs from sample to sample.
- We cannot predict the value of a statistic with absolute certainty.
- To make a reliable inference based on a specific statistic, we need to know the properties of the distribution of the statistic.

Definition

The **sampling distribution** of a statistic is the probability distribution of the statistic taken from all possible random samples of a specific sample size (*n*).

Random Samples

- In almost all observational studies, it is assumed that the data are obtained from a simple random sample. Usually, the sampling is done without replacement.
- A (**simple**) **random sample** (SRS) of size *n* is a sample selected in such a way that every possible sample of size *n* has the same chance of being selected.
- If the population is of finite size N and the sample size is n, then the number of possible simple random samples is $\binom{N}{n}$.

Sampling Variability

- It seems reasonable to use $\bar{x}$ to make inferences about μ.
- **Sampling variability** makes it difficult to know how far a specific $\bar{x}$ is from μ
- To make a reliable estimate, we need to know the exact probability distribution of $\bar{X}$.

Example: Approximate Distribution of the Sample Mean

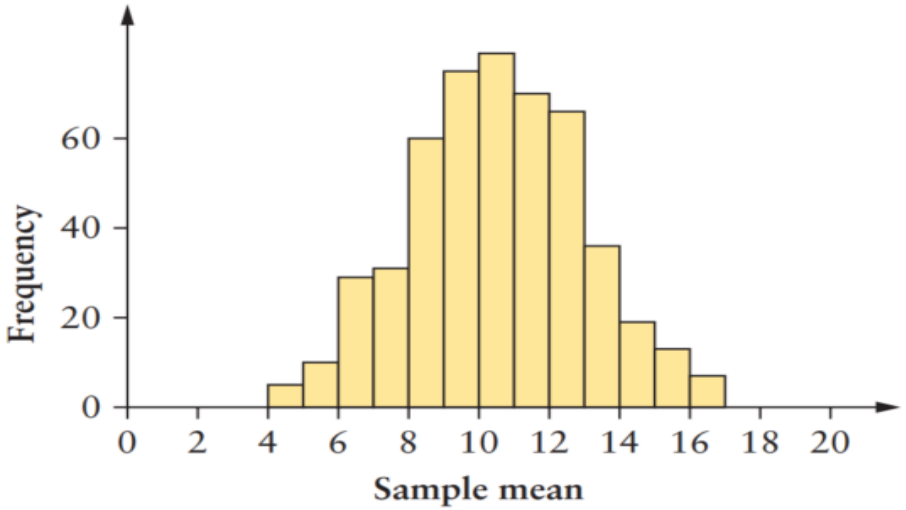
Consider a population consisting of the numbers 1, 2, 3, . . . 20. The population mean is

$\mu = (1 + 2 + 3 + \ldots + 20)/20 = 10.5$

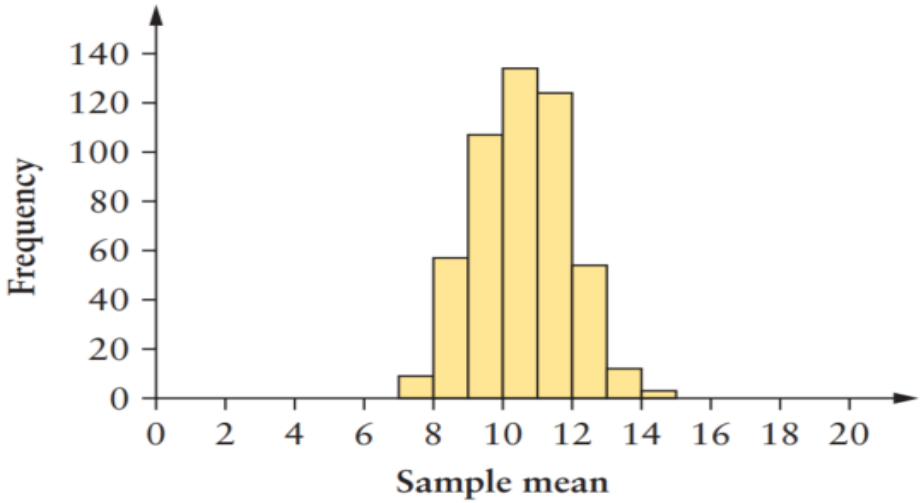
Use frequency histograms to approximate the distribution of the sample mean (shown in the next slide):

- Consider a random sample of *n* observations selected with replacement.
- For *n = 5*, five numbers are selected at random from the population, and the sample mean is computed. This procedure is repeated 500 times.
- A histogram of the resulting 500 sample means is shown on the following slide, as is the histogram for 500 samples of size *n = 10*.

Sampling Variability



Histogram of sample means for $n = 5$.

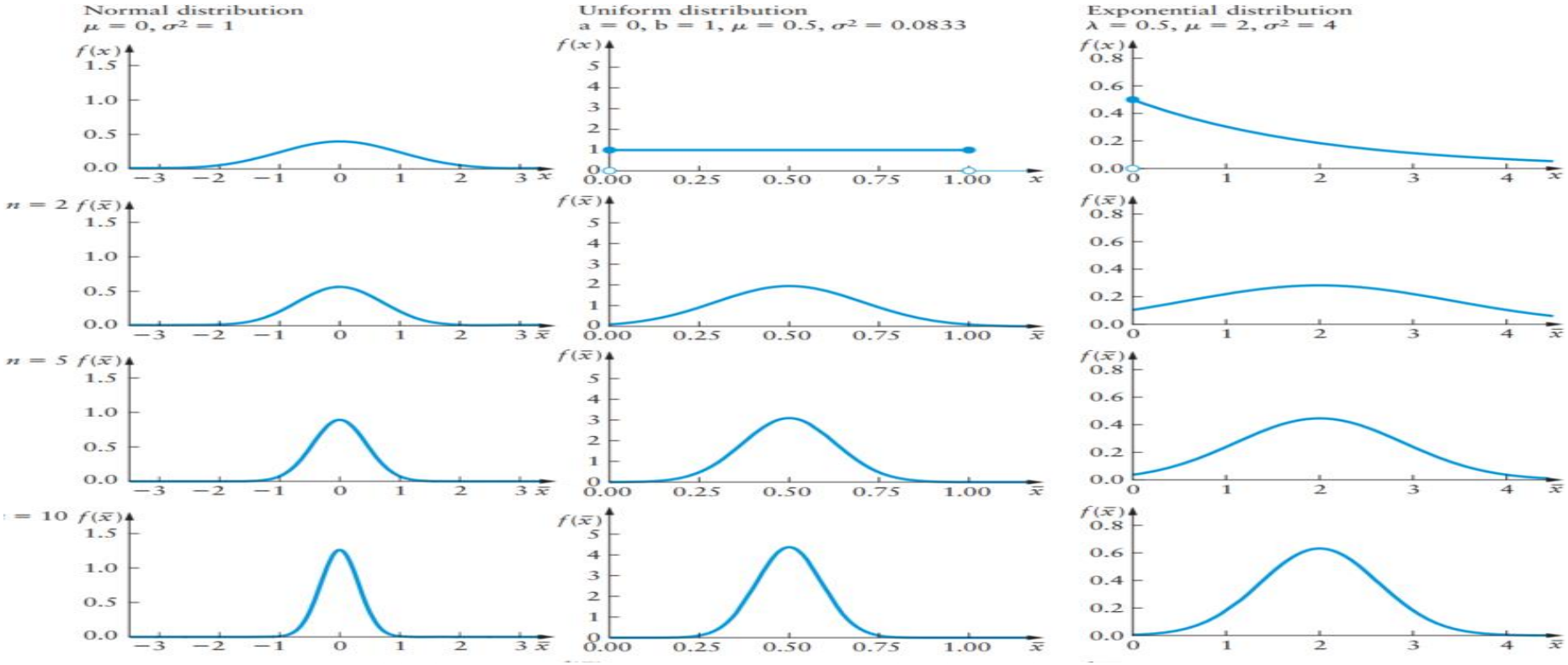


Histogram of sample means for $n = 10$.

- These histograms suggest some very surprising results.
  - ✓ Each distribution appears to be centered near the population mean 10.5.
  - ✓ In addition, the shape of each distribution is approximately normal!
  - ✓ Also, notice that the variability of the sampling distribution decreases as *n* increases. The sampling distribution of $\overline{x}$ when *n = 10* is more compact than when *n = 5*.
- This example implies that the distribution of the sample mean is approximately normal, with mean equal to the underlying, original population mean and variance related to the sample size *n*.

Distributions Connections

1. If the underlying population is normal, the distribution of the sample mean appears to be normal, regardless of the sample size.
2. Even if the underlying population is not normal, the distribution of the sample mean becomes more normal as *n* increases.
3. The sampling distribution of the mean is centered at the mean of the underlying population.
4. As the sample size *n* increases, the variance of the distribution of the sample mean decreases.



Properties of the Sample Mean

Let $\overline{X}$ be the mean of observations in a random sample of size n drawn from a population with mean μ and variance σ^2.

1. The mean of $\overline{X}$ is equal to the mean of the underlying population.

   In symbols: $\mu_{\overline{X}} = \mu$

2. The variance of $\overline{X}$ is equal to the variance of the underlying population divided by the sample size.

   In symbols: $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$

   The standard deviation, or standard error, of $\overline{X}$ is $\sigma_{\overline{X}} = \sqrt{\dfrac{\sigma^2}{n}} = \dfrac{\sigma}{\sqrt{n}}$.

3. If the underlying population is normally distributed, then the distribution of $\overline{X}$ is also exactly normal for any sample size.

   In symbols: $\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n)$

Central Limit Theorem (CLT)

Let $\overline{X}$ be the mean of observations in a random sample of size *n* drawn from a population with mean μ and variance σ^2.

As the sample size *n* increases, the sampling distribution of $\overline{X}$ will increasingly approximate a normal distribution with mean μ and variance σ^2/n, regardless of the shape of the underlying population distribution.

1. A better name for this result might be the *normal convergence theorem*. The distribution of $\overline{X}$ converges to a normal distribution.
2. If the original population is normally distributed, then the distribution of $\overline{X}$ is normal, no matter the sample size (*n*).
3. If the original population is not normal, the CLT says the distribution of $\overline{X}$ approaches a normal distribution as n increases and the approximation gets better as n gets bigger.
4. To compute a probability involving $\overline{X}$ we treat it just like any other random variable.

Example: Green Line Time

The Massachusetts Bay Transportation Authority (MBTA) Green Line from the Boston College stop to Park Street has trolleys leaving regularly throughout the day beginning at 5:01 A.M. The mean time for the trip is approximately 40 min. Suppose the travel time is normally distributed with a standard deviation $\sigma = 4$ min and a random sample of 25 trips is obtained.

(a) Find the probability that the sample mean time will be less than 38 min.

(b) Find the probability that the sample mean will be within 1 min of the population mean (40 min).

The underlying distribution is normal with $\mu = 40$, $\sigma^2 = 16$, $n = 25$. The sample mean is (exactly) normally distributed. $\bar{X} \sim \mathrm{N}(40, 16/25);\ \sigma_{\bar{X}} = \sqrt{16/25} = 4/5 = 0.8$ min

$$\mathrm{P}(\bar{X} < 38) = \mathrm{P}\left(\frac{\bar{X} - 40}{0.8} < \frac{38 - 40}{0.8}\right)$$

$$= \mathrm{P}(Z < -2.5) = 0.0062$$

The probability that the sample mean time will be less than 38 min is 0.0062.

$$\mathrm{P}(39 \le \bar{X} \le 41) = \mathrm{P}\left(\frac{39 - 40}{0.8} \le \frac{\bar{X} - 40}{0.8} \le \frac{41 - 40}{0.8}\right)$$

$$= \mathrm{P}(-1.25 \le Z \le 1.25)$$

$$= \mathrm{P}(Z \le 1.25) - \mathrm{P}(Z \le -1.25)$$

$$= 0.8944 - 0.0156 = 0.7888$$

Example: Milk Deliveries

In upstate New York, milk tanker trucks follow a daily routine, stopping at the same dairy farms every day. Farm output, however, varies because of weather, time of year, number of cows, and other factors. From years of recorded data, the mean amount of milk collected by a truck for processing is 7750 liters (L), with a standard deviation of 150 L. Suppose 36 trucks are randomly selected.

(a) Find the probability that the sample mean amount of milk picked up by the 36 trucks is more than 7800 L.

(b) Find a value *m* such that the probability that the sample mean is less than *m* is 0.1.

The exact distribution of the underlying population is not known. However, the sample size is large: $n = 36$ (≥30). Therefore, the CLT can be applied.

The underlying distribution has $\mu = 7750$, $\sigma = 150$, and $n = 36$. The distribution of $\bar{X}$ is approximately normal:

$$\bar{X} \overset{\cdot}{\sim} \mathrm{N}\left(7750, 150^2/36\right) = \mathrm{N}(7750, 625); \ \sigma_{\bar{X}} = \sqrt{625} = 25L$$

$$\mathrm{P}(\bar{X} > 7800) = \mathrm{P}\left(\frac{\bar{X} - 7750}{25} > \frac{7800 - 7750}{25}\right)$$

$$= \mathrm{P}(Z > 2)$$

$$= 1 - \mathrm{P}(Z \le 2)$$

$$= 1 - 0.9772 = 0.0228$$

The probability that the sample mean amount of milk picked up by the 36 trucks is more than 7800 L is 0.0228.

Example: Milk Deliveries

(b) Find a value *m* such that the probability that the sample mean is less than *m* is 0.1.

As the probability is already given, this is an inverse cumulative probability problem. Work backward in Table 3:

$$P(\bar{X} < m) = P\left(\frac{\bar{X} - 7750}{25} < \frac{m - 7750}{25}\right)$$

$$= P\left(Z < \frac{m - 7750}{25}\right) = 0.1$$

$$z = \frac{m - 7750}{25}$$

$$-1.28 = \frac{m - 7750}{25}$$

$$-1.28(25) = m - 7750 \Rightarrow m = 7718$$

The probability that the sample mean will be less than *m* = 7718 is (approximately) 0.1.

Distribution of the Sample Proportion

- We are often interested in drawing a conclusion about the population proportion *p* (the probability of the success). It seems reasonable to use the value of the sample proportion $\hat{P}$ to make an inference about the population proportion.
- Consider a sample of *n* individual/trails, and let X be the number of successes in the sample. The sample proportion is

$$P = \frac{X}{n} = \frac{\text{number of successes in the sample}}{\text{sample size}}$$

Sampling Distribution of *P* hat:

Let $\hat{P}$ be the sample proportion of successes in a sample of size n from a population with true proportion of success *p*:

1. The mean of $\hat{P}$ is the population proportion, $\mu_P = p$

2. The variance of P is $\sigma_P^2 = \frac{p(1-p)}{n}$ and the standard deviation is $\sigma_P = \sqrt{\frac{p(1-p)}{n}}$.

3. If n is large **and** both $np \geq 5$ and $n(1-p) \geq 5$,

then the distribution of $\hat{P}$ is approximately normal, $\hat{P} \overset{\cdot}{\sim} N(p, p(1-p)/n)$

Example: Too Many Regulations

Company executives often complain that there are too many government regulations. Suppose that 60% of all CEOs believe there are too many government regulations. A random sample of 150 CEOs is obtained, and each is asked whether he or she believes that there are too many government regulations.

(a) Find a value *r* such that the probability that the sample proportion is greater than *r* is 0.25.

(b) In recent years, big business has lobbied politicians to relax regulations to stimulate the economy. Suppose the sample proportion for the 150 CEOs is 0.56. Is there any evidence to suggest that the true proportion of CEOs who believe there are too many regulations has decreased?

For $n = 150$ and $p = 0.60$, check the nonskewness criterion.

$np = (150)(0.60) = 90 \geq 5$

$n(1 - p) = (150)(0.40) = 60 \geq 5$

Thus, both inequalities are satisfied. Then the distribution of

$\hat{P}$ is approximately normal, $\hat{P} \overset{\cdot}{\sim} N(0.6, 0.0016)$

$$\sigma_{\hat{P}}^2 = \frac{p(1-p)}{n}$$

$$\mu_{\hat{P}} = p = 0.6$$

$$= \frac{0.6(0.4)}{150} = 0.0016$$

Example: Too Many Regulations

(a) Find a value r such that the probability that the sample proportion is greater than r is 0.25.

$$\mathrm{P}(\hat{P} \geq r) = \mathrm{P}\left(\frac{\hat{P} - 0.60}{0.04} \geq \frac{r - 0.60}{0.04} \right)$$

$$= 1 - \mathrm{P}\left(Z \leq \frac{r - 0.60}{0.04} \right) = 0.25$$

$$\Rightarrow \mathrm{P}\left(Z \leq \frac{r - 0.60}{0.04} \right) = 1 - 0.25$$

$$\mathrm{P}\left(Z \leq \frac{r - 0.60}{0.04} \right) = 0.75$$

Find a value in the body of Table 3 as close to 0.75 as possible which is 0.7486.

$$z = \frac{r - 0.60}{0.04}$$

$$0.67 = \frac{r - 0.60}{0.04}$$

$$r = 0.6268$$

The probability that the sample proportion is greater than 0.6268 is 0.25.

Example: Too Many Regulations

(b) In recent years, big business has lobbied politicians to relax regulations to stimulate the economy. Suppose the sample proportion for the 150 CEOs is 0.56. Is there any evidence to suggest that the true proportion of CEOs who believe there are too many regulations has decreased?

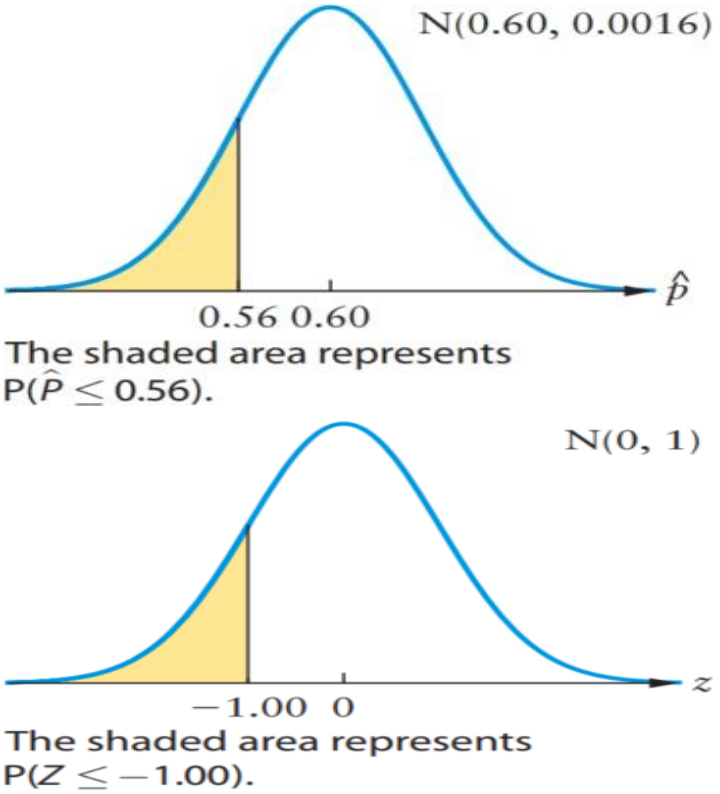
Claim : $p = 0.6 \Rightarrow \dot{P} \sim \mathrm{N}(0.60, 0.0016)$

Experiment : $\hat{p} = 0.56$

Likelihood :

$$\mathrm{P}(P < 0.56) = \mathrm{P}\left(\frac{P - 0.6}{0.04} \leq \frac{0.56 - 0.6}{0.04}\right)$$

$$= \mathrm{P}\,(Z \leq -1.00) = 0.1587$$



Conclusion: This likelihood probability is larger than 0.05. There is no evidence to suggest the claim of $p = 0.60$ is wrong.