8 Interval Estimation

- Why do we need intervals for population parameter estimates when we have a single point estimate?

  - A point estimator cannot be expected to provide the exact value of the population parameter: due to the presence of sample uncertainty

  - Purpose: of an interval estimate is to provide information about how close the point estimate, provided by the sample, is to the value of the population parameter.

- An interval estimate is often computed by adding and subtracting a value called the margin of error to the point estimate

- General form of interval estimate is:

$$\text{Point estimate} \pm \text{margin of error}$$

$$\hat{\theta} \pm \text{margin of error}$$

- Note that our understanding of sampling distributions play key role in comprehending intervals.

8.1 Interval estimate for population mean μ, when population standard deviation σ is known

- To develop an interval estimate of a population mean μ and compute the margin of error by using
  - either population standard deviation, denoted by σ,
  - or we use the sample standard deviation denoted by s.

- We will look at cases where two cases, one where
  - population standard deviation, σ is known: when large amounts of historical data is available.
  - population standard deviation, σ is unknown

- Example of Lloyd's Department Store: each week it selects a simple random sample of 100 customers in order to learn about the amount spent per shopping trip.
  - x : amount spent per shopping trip
  - $\bar{x}$: point estimate of μ which is the average amount spent per shopping trip for all of Lloyd's customers

- Let us assume the following:
  - From a previous study with the Lloyd stores, $\sigma = \$20$ is the population standard deviation.
  - From this σ, we know that the population follows a normal distribution

- Example of Lloyd's Department Store: each week it selects a simple random sample of 100 customers in order to learn about the amount spent per shopping trip.

- Sample attributes are as follows:

  - Sample size, $n = 100$ and sample average is $\bar{x} = \$82$.

    - Inference: This implies that an average customer in the sample collected spends 82$ at the store.

    - Does this imply that this is the average amount spent at the store by all customers who come to the store?

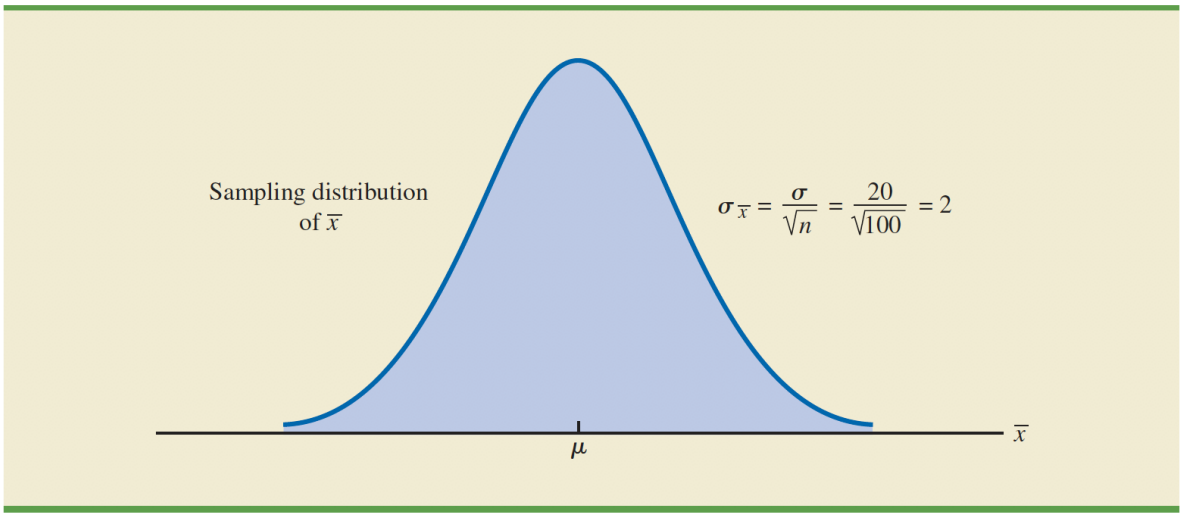
  - How do we compute margin of error across samples?

- Sampling distribution of $\bar{x}$ also follows a normal distribution (see footnote 1, page 348) and has a standard error as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = 2$$

Note: (a) sampling distribution shows how values of $\bar{x}$ are distributed around the population mean, μ the sampling distribution of $\bar{x}$ provides information about the possible differences between $\bar{x}$ and μ.

FIGURE 8.1 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN AMOUNT SPENT FROM SIMPLE RANDOM SAMPLES OF 100 CUSTOMERS



Definition 8.1. Interval Estimate of a Population Mean with Known σ

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

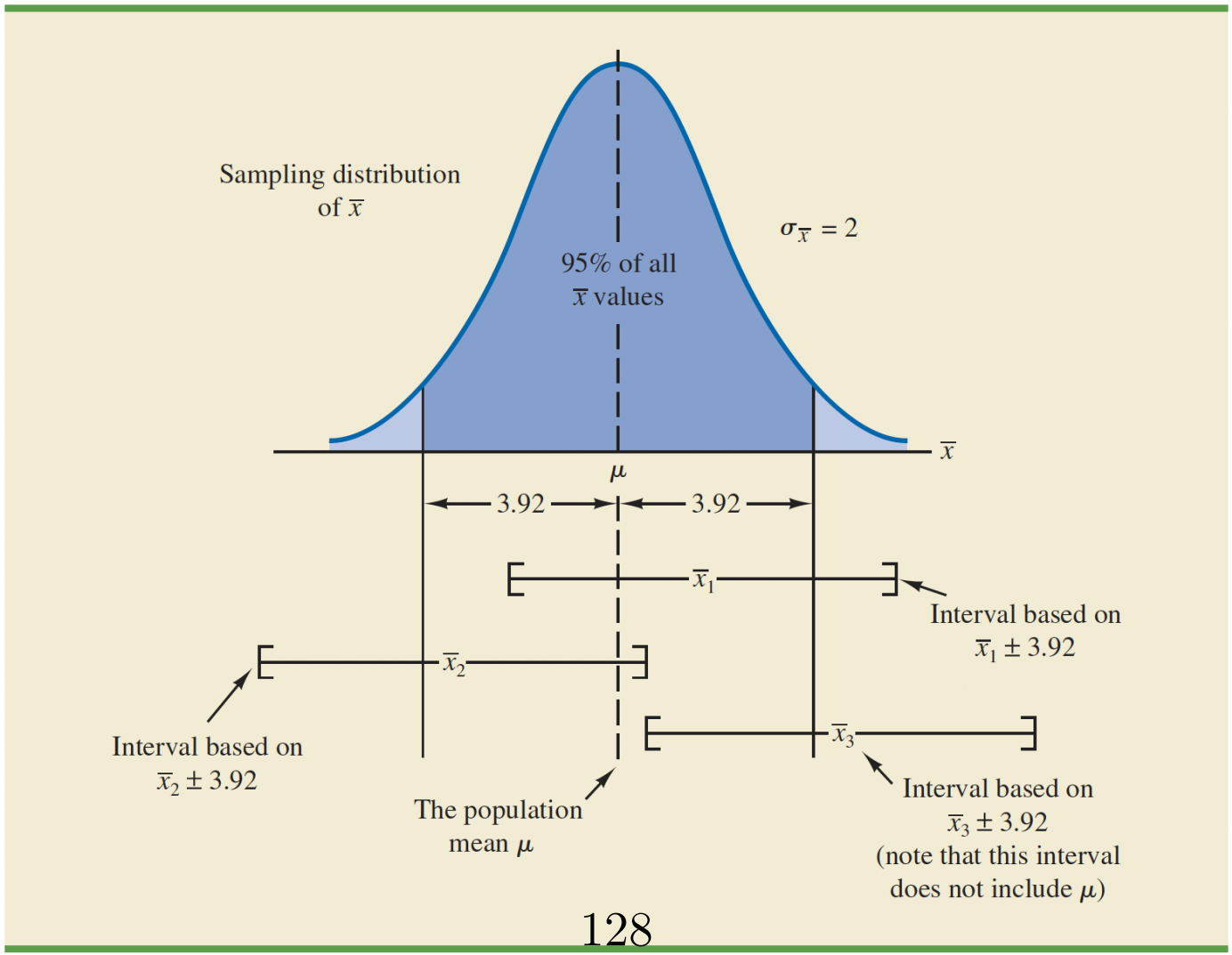
where $(1-\alpha)$ is the confidence coefficient and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution

- From our knowledge of standard normal distribution, we find that
  - 95% of the $\bar{x}$ values must be within $\pm z_{\alpha/2}\sigma_{\bar{x}}$
  - 95% of the samples constructed from any $\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} = 1.96(2) = 3.92$ will contain the unknown average μ money spent by shopper at Lloyd's
- Inference: 95% of all average amount spent at the store obtained from a sample of size 100 would lie in ± 3.92 of the unknown population mean, μ

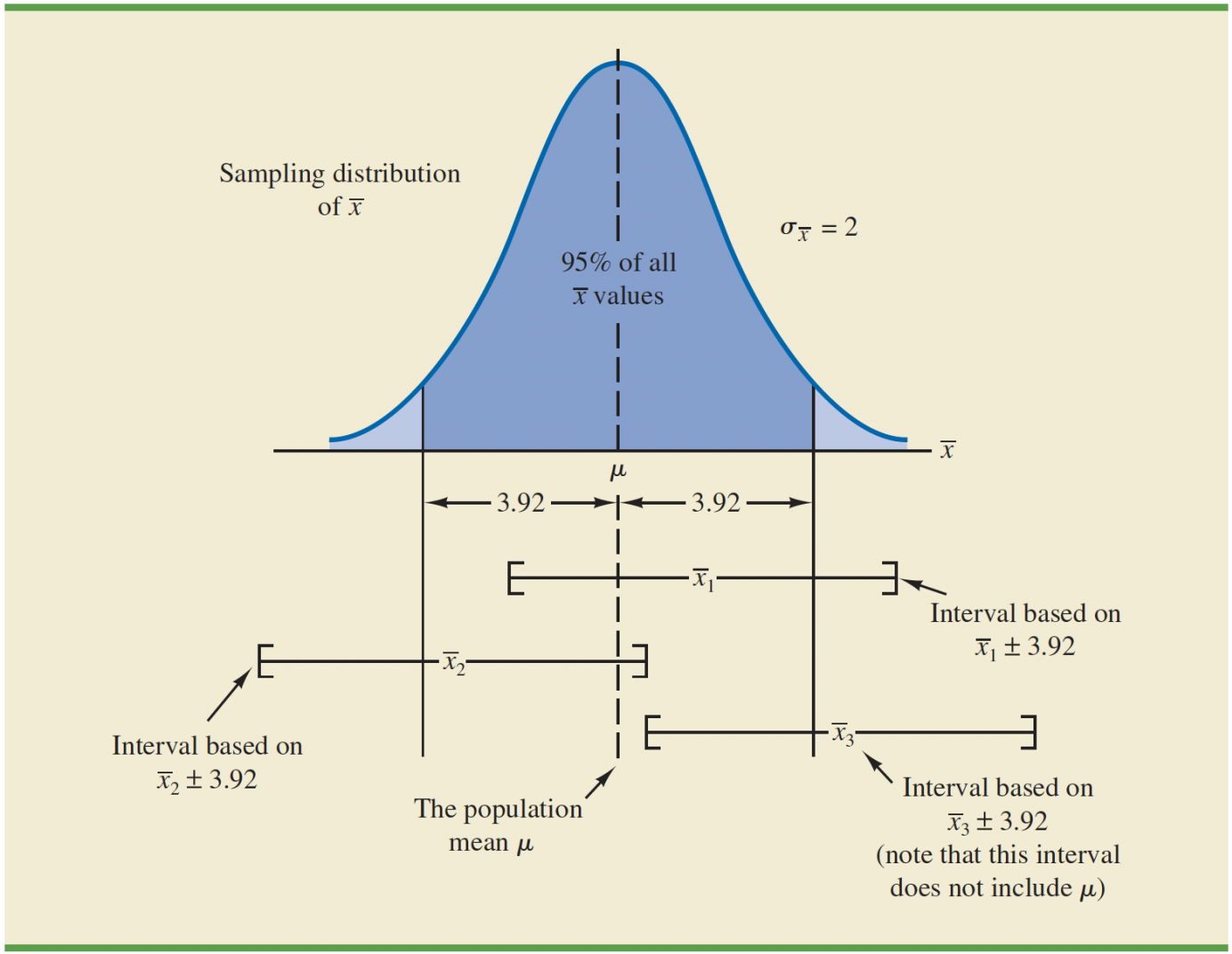
- Let us take three different samples of Lloyd's customers by collecting a sample every day
  - Comprising of $n = 100$ Lloyd customers with respective sample means as $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$
  - Three intervals would be

$$\bar{x}_1 \pm 3.92 \qquad \bar{x}_2 \pm 3.92 \qquad \bar{x}_3 \pm 3.92$$

Caution: You could obtain a sample which has an average spending denoted by $\bar{x}_3$. Now since $\bar{x}_3$ falls in the upper tail of the sampling distribution and is farther than 3.92 from μ, subtracting and adding 3.92 to $\bar{x}_3$ forms an interval that does not include μ. You still do not know μ.



Caution: Our claim is that 95% of all sample means will lie in the $\mu \pm 3.92$, or the blue shaded range, that the possibility of obtaining $\bar{x}_3$ will only be 5% under our assumptions.



- Lloyd Departmental Store:
  - Interval estimate: $\bar{x} = 82 \pm 3.92$ or 78.08 to 82
  - Claim: We are 95% confident that the interval includes population mean μ since 95% of the intervals constructed using *any* $\bar{x} \pm 3.92$ will contain the unknown population mean.

- For 95% confidence interval using known population standard deviation, σ,
  - We first obtain sample standard deviation, $\sigma_{\bar{x}}$
  - $\alpha = 0.05$ which implies that there is a 0.05 percent chance that our sample does not contain the population mean
  - Divide this in the two extremes i.e. $\alpha/2 = 0.025$ on lower and upper provides a z statistic for most samples as 1.96
  - Calculate the deviation from the sample mean $\bar{x} = 82$ given $\sigma = 20$

$$82 \pm 1.96 \frac{20}{\sqrt{100}}$$

- To summarize:
  - This interval has been established at the 95% confidence level.
  - The value 0.95 is referred to as the confidence coefficient $(1 - \alpha)$,
  - and the interval $[78.08, 85.92]$ is called the 95% confidence interval.

Lecture 8: 24 Feb 2021, Wednesday

TBA students present

- Interval estimation with unknown population parameter
- Introducing t-stats
- Average credit card debt example

TABLE 8.1 VALUES OF $Z_{\alpha/2}$ FOR THE MOST COMMONLY USED CONFIDENCE LEVELS

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.05	1.645
95%	.05	.025	1.960
99%	.01	.005	2.576

- Compute the 99% confidence interval for Lloyd's example when $n = 100$ and sample average is $\bar{x} = 82$

- This comes out to be

$$82 \pm 2.576 \frac{20}{\sqrt{100}} = 82 \pm 5.15 = [76.85, 87.15]$$

- Interpretation: we are 99% confident that this interval will contain the unknown population mean, μ

- Read Practical Advice: What happens if population does not follow normal distribution? What can we say about the accuracy of our intervals?

8.2 Interval estimate for population mean μ, when population standard deviation σ is unknown

- When the population standard deviation σ is unknown – as it usually is – we use the sample standard deviation s to *proxy* its population counterpart σ,

- In such cases, the margin of error and the interval estimate for the population mean are based on a probability distribution known as the t-distribution.

- Mean of the t-distribution is zero just like the standard normal distribution, but its spread and peak depends upon its degrees of freedom.

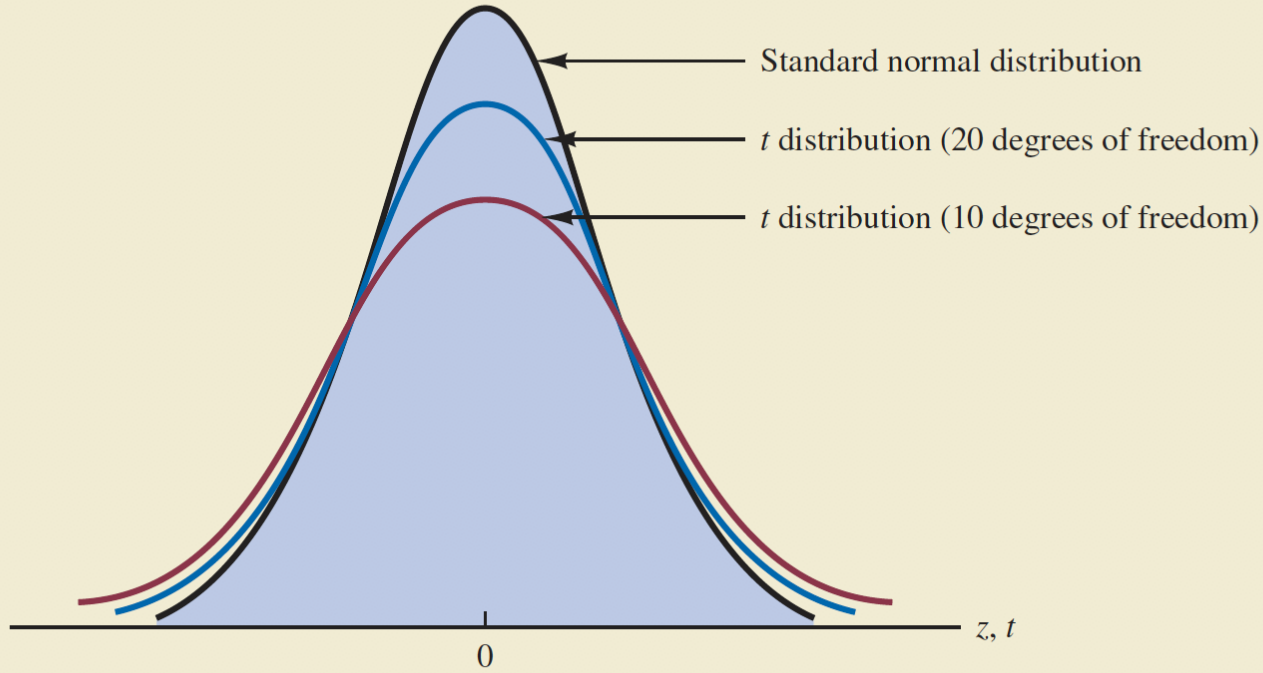
- A specific $t-$distribution depends on a parameter called degrees of freedom (df)

- Degrees of freedom, *df* refer to the number of independent pieces of information that go into the computation of the parameter of interest

  - A t-distribution with 1 *df* looks different from one with 2 *df* and so on.

  - As degree of freedom increases, t-distribution tends towards a normal distribution.

FIGURE 8.4 COMPARISON OF THE STANDARD NORMAL DISTRIBUTION WITH *t* DISTRIBUTIONS HAVING 10 AND 20 DEGREES OF FREEDOM



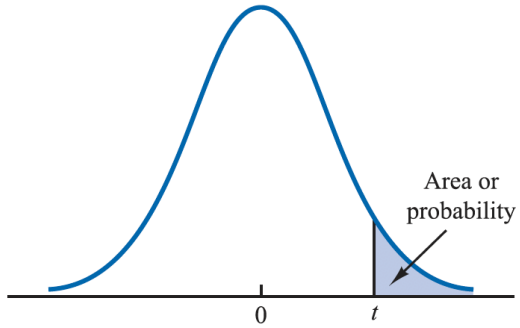
- $t_{\alpha/2}$ is the t-value with an area of $\alpha/2$ on its right i.e. on the upper tail of the distribution.

- From the table observe that as degree of freedom continues to increase, $t_{0.025}$ looks like standard normal $z_{0.025}$

- Reading the Table: For $df = 6$ and an area of 0.025 i.e. 2.5%, we obtain a t-value of 2.447. This implies that a t-value larger than 2.447 has a probability of only 2.5% or lower.

- If $df > 100$, infinite degrees of freedom or even the z-table can be used to approximate t

TABLE 8.2 SELECTED VALUES FROM THE T DISTRIBUTION TABLE*



Degrees of Freedom	Area in Upper Tail						
	.20	.10	.05	.025	.01	.05	.005
1	1.376	3.078	6.314	12.706	31.821	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	2.821	3.250
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	.848	1.296	1.671	2.000	2.390	2.390	2.660
61	.848	1.296	1.670	2.000	2.389	2.389	2.659
62	.847	1.295	1.670	1.999	2.388	2.388	2.657
63	.847	1.295	1.669	1.998	2.387	2.387	2.656
64	.847	1.295	1.669	1.998	2.386	2.386	2.655
65	.847	1.295	1.669	1.997	2.385	2.385	2.654
66	.847	1.295	1.668	1.997	2.384	2.384	2.652
67	.847	1.294	1.668	1.996	2.383	2.383	2.651
68	.847	1.294	1.668	1.995	2.382	2.382	2.650
69	.847	1.294	1.667	1.995	2.382	2.382	2.649
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	.846	1.291	1.662	1.987	2.368	2.368	2.632
91	.846	1.291	1.662	1.986	2.368	2.368	2.631

Definition 8.2. Interval Estimate of Population Mean with Unknown σ We can construct a margin for error and an interval for population mean, μ by using the t-statistic and the sample standard deviation, s in the case of unknown population standard deviation, σ

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

- $(1 - \alpha)$ is the confidence coefficient,
- and $t_{\alpha/2}$ is the t-value providing an area of $(\alpha/2)$ in the upper tail of the t distribution,
- with (n-1) degrees of freedom.

8.2.1 Degrees of Freedom

- Why is (n-1) the degrees of freedom?

- Recall standard deviation formula:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

- Degree of freedom refer to the number of independent pieces of information that go into the computation of $\sum (x_i - \bar{x})^2$. And what are they? They are the sequence of $(x_1 - \bar{x})$, $(x_1 - \bar{x})$ utpo $(x_n - \bar{x})$

- We know that $\sum(x_i - \bar{x}) = 0$ for any data set which implies that of the n sequence, only $n - 1$ are independent pieces of information.

- Idea: We can obtain the n^{th} observation back from the formula $\sum(x_i - \bar{x}) = 0$ and the remaining (n-1) sum of squared deviations

8.2.2 Example: Average Credit Card Debt

- ▷ Example of Average Credit Card Debt: study designed to estimate the mean credit card debt for a population of US households
  - ○ Sample size, n = 70 with population standard deviation σ unknown
  - ○ Sample estimates are $\bar{x} = \$9312$ and $s = \$4007$

TABLE 8.3 CREDIT CARD BALANCES FOR A SAMPLE OF 70 HOUSEHOLDS

9430	14661	7159	9071	9691	11032
7535	12195	8137	3603	11448	6525
4078	10544	9467	16804	8279	5239
5604	13659	12595	13479	5649	6195
5179	7061	7917	14044	11298	12584
4416	6245	11346	6817	4353	15415
10676	13021	12806	6845	3467	15917
1627	9719	4972	10493	6191	12591
10112	2200	11356	615	12851	9743
6567	10746	7117	13627	5337	10324
13627	12744	9465	12557	8372	
18719	5742	19263	6232	7445	

- Construct an interval estimate on which you can be 95% confident that it would capture the population mean, μ:

- Since we do not have information about population standard deviation, we use the $t_{\alpha/2} = t_{0.025}$ with $n - 1 = 69$ degrees of freedom

- From the table we obtain $t_{0.025} = 1.995$

- Example of Average Credit Card Debt: we can obtain an interval from [\$8357, \$10267] computed using the margin of error from a t with 69 degrees of freedom and 95% confidence

$$9312 \pm 1.995 \frac{4007}{\sqrt{70}}$$

- Inference: we are 95% confident that the mean credit card balance for the population of all households is between \$8357 and \$10,267.

- When you suspect a skewed population then the intervals using t-statistic will only be an approximation to the actual intervals.
  - What should the sample size be? At least 30 observation for a sample to have smaller approximation errors.
  - Sample is highly skewed or has outliers, then size should be 50.
  - Certainly having a larger sample makes us more sure, probabilistically

8.2.3 Using a Small Sample

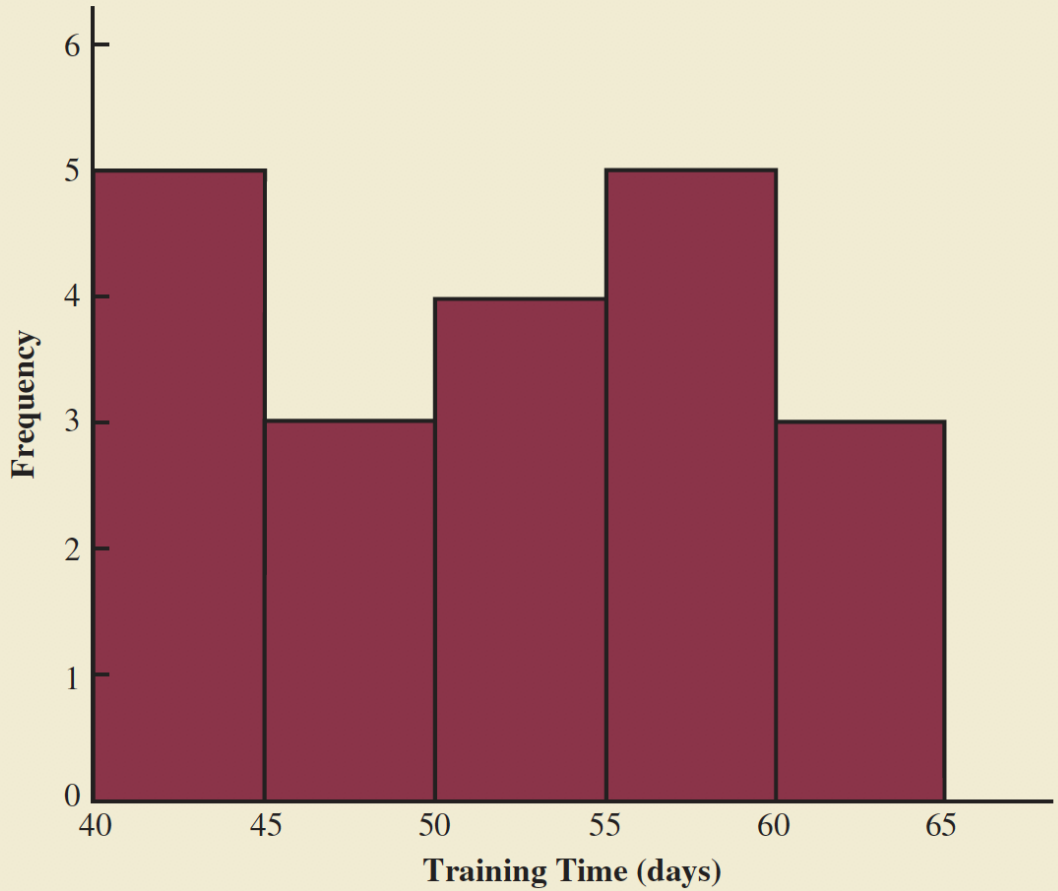
- ▷ Example of Scheer Industries: which is considering a new computer-assisted program to train maintenance employees to do machine repairs.

- ▷ Director requests you to estimate the population mean time required for maintenance employees to complete the computer-assisted training program

  - ○ You select a sample of 20 recording the training time in days for Scheer employees.

52	59	54	42
44	50	42	48
55	54	60	55
44	62	62	57
45	46	43	56

FIGURE 8.7 HISTOGRAM OF TRAINING TIMES FOR THE SCHEER INDUSTRIES SAMPLE



Does not look normal. But no outliers as such so we could construct an interval estimate based on t-statistic.

- Obtain a 95% confidence interval when $t_{0.025} = 2.093$ under 19 degrees of freedom

- Step 1: Calculate mean or the point estimate of the population mean which is the average number of days required for Scheer employees to complete the computer assisted training.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1030}{20} = 51.5\text{days}$$

- Step 2: Calculate the sample standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{889}{20-1}} = 6.84\text{days}$$

- Step 3: Check the t-table for n-1 degree of freedom for the sample standard deviation. We obtain $t_{0.025} = 2.093$

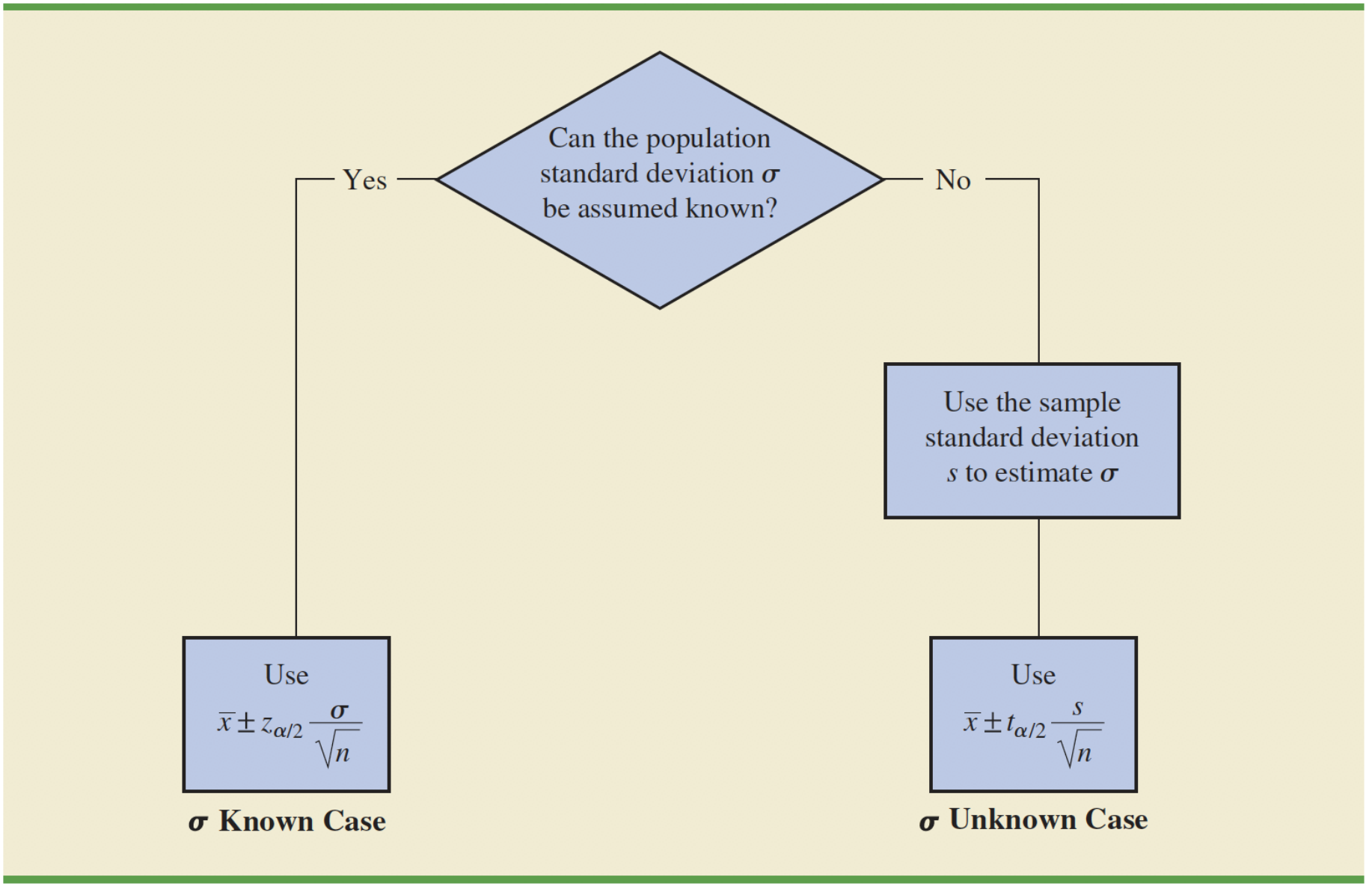
- Step 4: Calculate the margin of error

$$t_{\alpha/2}\frac{s}{\sqrt{n}} = 2.093\frac{6.84}{\sqrt{20}} = 3.2$$

- Step 5: Interested interval is obtained as

$$51.5 \pm 3.2$$

- Step 6 is Inference: we can now say that we are 95% confidence that an employee can complete the training program in between $[48.3, 54.7]$ days.


Can the population standard deviation σ be assumed known?
Yes
No
Use the sample standard deviation s to estimate σ
Use $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Use $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$
σ Known Case
σ Unknown Case

8.3 Determining the Sample Size

- How large should my sample be so that I minimize my margin of error?
  - What are the costs of a small sample size?
  - What are the benefits of a large sample size?

- Recall our margin for error under the desired confidence coefficient $(1-\alpha)$ with known population standard deviation σ case:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Lets call this E, the desired margin for error. So for a desired margin of error, we can obtain the sample size once we have σ and $z_{\alpha/2}$.

Definition 8.3. Sample size according to Desired Margin of Error We obtain the optimal sample size for an interval estimate based on our desired margin for error by rearranging the desired margin for error, E:

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

- Optimal Sample Size: Problem is to choose n so that we make the E according to our acceptable limits. What can we do in the case of unknown population standard deviation σ

  - We use planning values to substitute σ, by using s from previous samples studies
  - We do a quick sample collection – a pilot sample – and use its s to proxy our unknown σ
  - Use our judgment about σ according to the variable in question.

- Example: Renting a Mid-Size Car – A previous study that investigated the cost of renting automobiles in the United States found a mean cost of approximately \$55 per day for renting a midsize automobile.

- New Study required to examine this cost in 2020. Study directs you to keep your margin for error to a narrow \$2 from the population mean, and maintain at least a 95% confidence. How large should your sample be if a previous sample for 2018 had a sample standard deviation of \$9.65.

- Optimal sample Size: can be calculated from the margin for error as:

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2} = \frac{1.96^2 9.65^2}{2^2} = 89.43$$

8.4 Population Proportion

- General form of interval estimate of a population proportion:

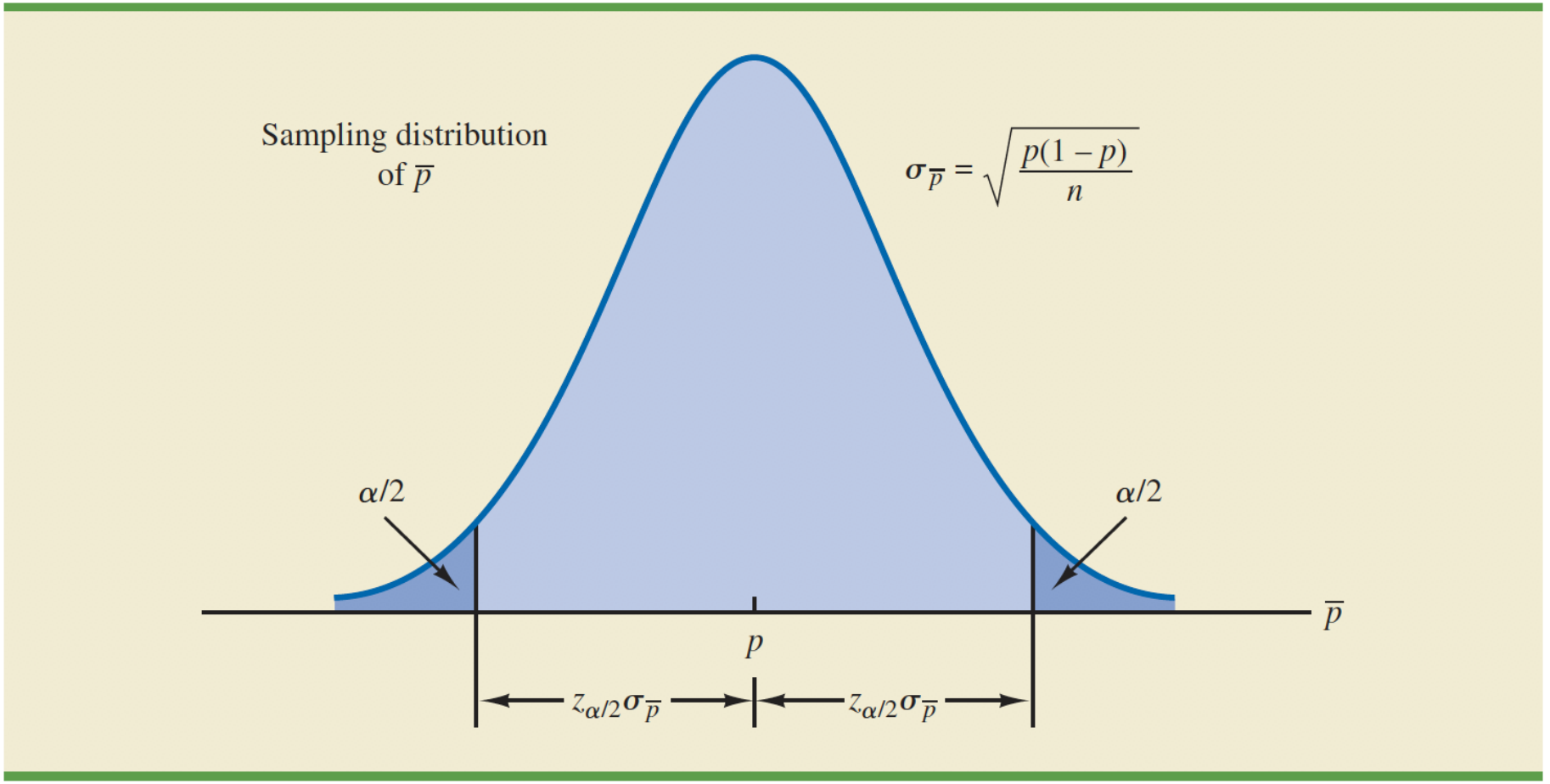
$$\bar{p} \pm \text{margin for error}$$

- Recall: sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever

$$np \geq 5 \qquad n(1-p) \geq 5$$

Understand: Mean of sampling distribution of $\bar{p}$ is the actual unknown population proportion, p, with standard deviation denoted by $\sigma_{\bar{p}}$

FIGURE 8.9 NORMAL APPROXIMATION OF THE SAMPLING DISTRIBUTION OF $\bar{p}$



- Standard error for sample proportion is

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

- If the two conditions for normal approximation are satisfied, then we use $z_{\alpha/2}\sigma_{\bar{p}}$ as the margin for error

- Unknown actual standard error implies since population proportion, p is not known so use $\bar{p}$ to arrive at margin for error as:

$$\text{Margin for error } = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Definition 8.4. Interval Estimate of a Population Proportion

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where $(1-\alpha)$ is the confidence coefficient
and $z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal distribution.

8.4.1 Example: Women Tee Times in Golf

- Example on Women Discrimination in Golf: A national survey of 900 women golfers was conducted to learn how women golfers view their treatment at golf courses in the United States.

- Results of the survey found that 396 of the women golfers were satisfied with the availability of tee times.

- Compute the point estimate of the proportion of the population (p) of women golfers who are satisfied with the availability of tee times

$$\bar{p} = \frac{396}{900} = 0.44$$

- Using this, compute a 95% confidence interval for the population proportion

- Step 1: check the conditions for normality: $np = 396$ and $n(1-p) = 504$. This implies we can approximate it with normal distribution

- Step 2: Obtain the critical z-value: $z_{\alpha/2}$for 95% confidence is 1.96

- Step 3: Calculate the margin of error using population proportions, $\bar{p}$

$$1.96\sqrt{\frac{0.44(1-0.44)}{900}} = 0.0324$$

- ▷ Step 4: obtain the confidence interval by adjusting the point estimate for the margin for error.

$$0.44 \pm 0.324$$

- ▷ Step 5 Interval and its interpretation: We can state with 95% confidence that the actual porportion of women who are satisfied with their tee times on this Golf course would be a proportion between 40.76 to 47.24.

8.4.2 Determining the Sample Size

- Flip Question: how large the sample size should be to obtain an estimate of a population proportion at a specified level of precision. Let E be the desired margin for error:

$$E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

  - Larger sample sizes provide a smaller margin of error and better precision.

- Solving for n we get a sample size estimate for the desired margin for error, E

$$n = \frac{z^2_{\alpha/2}\bar{p}(1-\bar{p})}{E^2}$$

- But n cannot be determined without $\bar{p}$.

- Thus let us assume a proportion p^*, called the planning value and calculate the sample size associated with this.

$$n = \frac{z^2_{\alpha/2} p^*(1-p^*)}{E^2}$$

- Considerations for choosing planning value:
  - Use the sample proportion from a previous sample of the same or similar units
  - Use a pilot study (baby or small study) to select a preliminary sample.
  - Use judgment or a "best guess" for the value of p^*
  - If none of the preceding alternatives apply, use a planning value of $p^* = 0.50$

8.4.3 Determining Sample Size with Planning Value

▷ Golfer Tee Times: How large should the sample be if the survey director wants to estimate the population proportion with a margin of error of 0.025 at 95% confidence?

a. Use the previous survey proportion as the planning value, $p^* = 0.44$

$$n = \frac{z^2_{\alpha/2} p^*(1-p^*)}{E^2} = \frac{(1.96)^2(0.44)(1-0.44)}{0.025^2} = 1514.5$$

- Desired sample size: sample size of 1515.
- Implication: This sample size of women's golfers satisfies a margin error of 0.025 for a proportion which were satisfied with their tee times
- Use a planning value of $p^* = 0.50$

▷ Golfer Tee Times: How large should the sample be if the survey director wants to estimate the population proportion with a margin of error of 0.025 at 95% confidence?

b. Use a neutral planning value of $p^* = 0.50$, we obtain the desired sample size as

$$n = \frac{(1.96)^2(0.50)(1-0.50)}{0.025^2} = 1536.64$$

◦ Desired Sample Size: increases to 1537 and with the margin of error requirement, such a sample size is appropriate for the sample proportion.

◦ Note: Larger the sample size, smaller is the margin of error

Optimal Sample Size with Proportions

- Cochran's formula for computing sample size for infinite population

$$n = \frac{z^2_{\alpha/2} p(1-p)}{E^2}$$

where:

  - z is the critical value for the desired confidence interval
  - p is the estimated or guess proportion of our attribute that is present in the population
  - E is the desired margin of error

- Yamane's formula for computing sample size with known population size

$$n = \frac{N}{1 + N(E^2)}$$

where:

  - N is the known size of the population
  - E is the desired margin of error