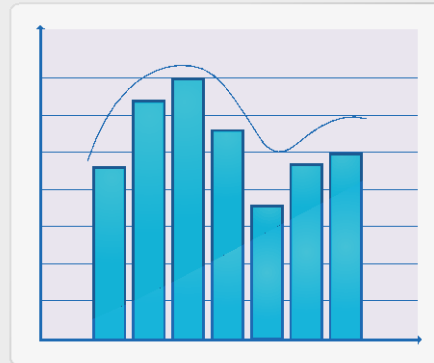
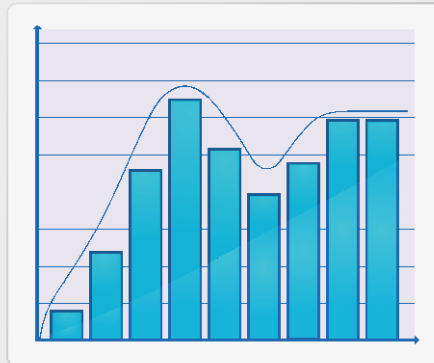
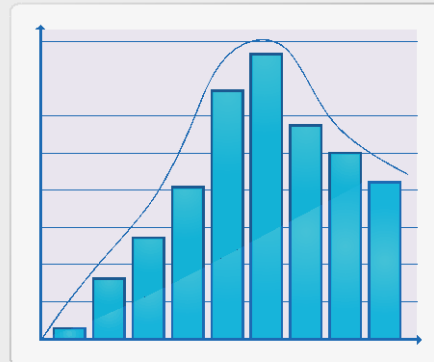
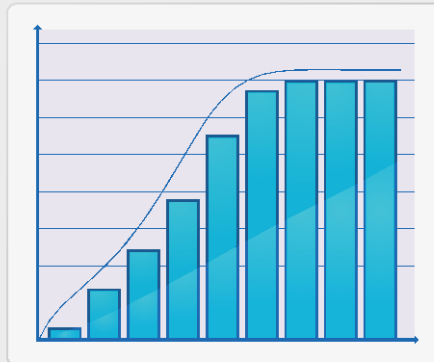


Chapter 6

Sampling and Estimation



Statistical Sampling

- ▶ Sampling is the foundation of statistical analysis.
- ▶ **Sampling plan** - a description of the approach that is used to obtain samples from a population prior to any data collection activity.
- ▶ A sampling plan states:
 - its objectives
 - target population
 - population frame (the list from which the sample is selected)
 - operational procedures for collecting data
 - statistical tools for data analysis

Example 6.1: A Sampling Plan for a Market Research Study

- ▶ A company wants to understand how golfers might respond to a membership program that provides discounts at golf courses.
 - Objective - estimate the proportion of golfers who would join the program
 - Target population - golfers over 25 years old
 - Population frame - golfers who purchased equipment at particular stores
 - Operational procedures - e-mail link to survey or direct-mail questionnaire
 - Statistical tools - PivotTables to summarize data by demographic groups and estimate likelihood of joining the program

Sampling Methods

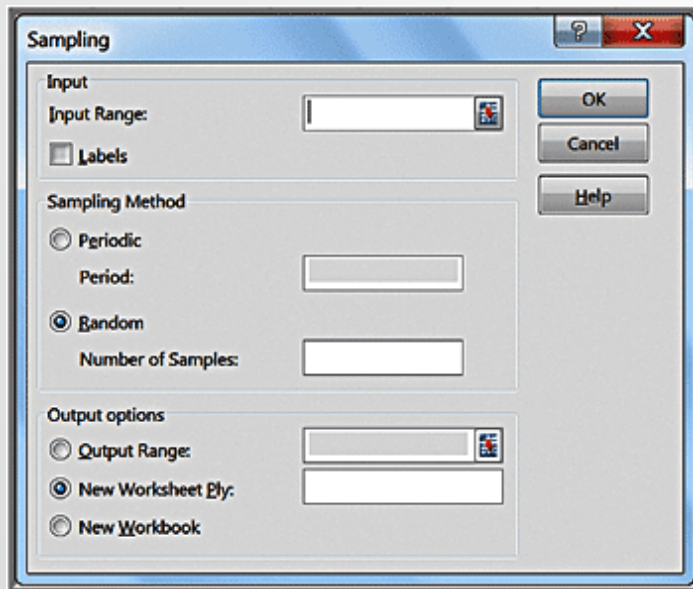
- ▶ Subjective Methods
 - ▶ **Judgment sampling** – expert judgment is used to select the sample
 - ▶ **Convenience sampling** – samples are selected based on the ease with which the data can be collected
- ▶ Probabilistic Sampling
 - ▶ **Simple random sampling** involves selecting items from a population so that every subset of a given size has an equal chance of being selected

Example 6.2: Simple Random Sampling with Excel

- ▶ *Sales Transactions* database

Data > Data Analysis > Sampling

- ▶ *Periodic* selects every n^{th} number
- ▶ *Random* selects a simple random sample



	A
1	Sample of Customer IDs
2	10009
3	10092
4	10102
5	10118
6	10167
7	10176
8	10256
9	10261
10	10266
11	10293
12	10320
13	10336
14	10355
15	10355
16	10377
17	10393
18	10413
19	10438
20	10438
21	10455

Sampling is done *with replacement* so duplicates may occur.

Additional Probabilistic Sampling Methods

- ▶ **Systematic (periodic) sampling** – a sampling plan that selects every n^{th} item from the population.
- ▶ **Stratified sampling** – applies to populations that are divided into natural subsets (called **strata**) and allocates the appropriate proportion of samples to each stratum.
- ▶ **Cluster sampling** - based on dividing a population into subgroups (clusters), sampling a set of clusters, and (usually) conducting a complete census within the clusters sampled
- ▶ **Sampling from a continuous process**
 - Select a time at random; then select the next n items produced after that time.
 - Select n times at random; then select the next item produced after each of these times.

Estimating Population Parameters

- ▶ **Estimation** involves assessing the value of an unknown population parameter using sample data
- ▶ **Estimators** are the measures used to estimate population parameters
 - E.g., sample mean, sample variance, sample proportion
- ▶ A **point estimate** is a single number derived from sample data that is used to estimate the value of a population parameter.
- ▶ If the expected value of an estimator equals the population parameter it is intended to estimate, the estimator is said to be **unbiased**.

Sampling Error

- ▶ **Sampling (statistical) error** occurs because samples are only a subset of the total population
 - Sampling error is inherent in any sampling process, and although it can be minimized, it cannot be totally avoided.
- ▶ **Nonsampling error** occurs when the sample does not represent the target population adequately .
 - Nonsampling error usually results from a poor sample design or inadequate data reliability.

Example 6.3: A Sampling Experiment

- ▶ A population is uniformly distributed between 0 and 10.
 - Mean = $(0 + 10)/2 = 5$
 - Variance = $(10 - 0)^2/12 = 8.333$
- ▶ Experiment:
 - Generate 25 samples of size 10 from this population.
 - Compute the mean of each sample.
 - Prepare a histogram of the 250 observations,
 - Prepare a histogram of the 25 sample means.
 - Repeat for larger sample sizes and draw comparative conclusions.

Example 6.3: Experiment Results

	A	B	C	D	E	F	W	X	Y	Z	AA	AB
1	Observation	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 22	Sample 23	Sample 24	Sample 25		
2	1	5.3935	0.8756	9.9338	4.3294	7.1908	8.2244	8.4655	1.9404	9.9133		
3	2	2.8282	2.5047	6.4480	5.9877	7.3946	3.3000	0.3632	9.8871	5.1079		
4	3	5.2715	0.6949	1.5015	8.3935	3.1559	7.1023	9.3628	1.7844	7.3937		
5	4	5.4912	0.7739	8.1466	5.5205	2.4586	8.7262	9.1598	7.5820	1.8513		
6	5	9.3158	4.4591	6.3573	3.8679	1.1493	3.3854	1.2482	1.9391	5.4405		
7	6	7.9745	6.9784	7.9962	2.3157	7.8564	8.9032	3.8716	5.8525	5.4164		
8	7	6.7043	8.4039	5.1088	9.1098	1.1802	2.7732	2.4815	9.0817	4.3889		
9	8	1.3041	2.5678	6.1794	7.8396	6.2709	0.5692	2.5800	1.1911	7.2430		
10	9	0.9870	6.3964	8.2269	9.6112	6.6814	2.8306	4.6004	9.0274	6.1232		
11	10	9.9493	9.3936	4.5015	5.2385	0.6970	3.7074	5.9062	0.6592	7.5021		
12	Sample mean	5.5219	4.3048	6.4400	6.2214	4.4035	4.9522	4.8039	4.8945	6.0380	Average	5.0108
13											Standard Dev.	0.816673

Histogram - All Data

This histogram shows the frequency distribution of all data points across 10 bins. The x-axis is labeled 'Bin' and ranges from 1 to 10, with a 'More' category. The y-axis is labeled 'Frequency' and ranges from 0 to 40. The bars show a roughly bell-shaped distribution centered around bin 5.

Bin	Frequency
1	23
2	22
3	25
4	28
5	24
6	28
7	30
8	21
9	28
10	21

Sample Means - n = 10

This histogram shows the frequency distribution of the 10 sample means. The x-axis is labeled 'Bin' and ranges from 4 to 6, with a 'More' category. The y-axis is labeled 'Frequency' and ranges from 0 to 5. The distribution is centered around bin 5, which has the highest frequency of 4.

Bin	Frequency
4	2
4.2	2
4.4	3
4.6	2
4.8	2
5	3
5.2	1
5.4	1
5.6	2
6	3
More	4

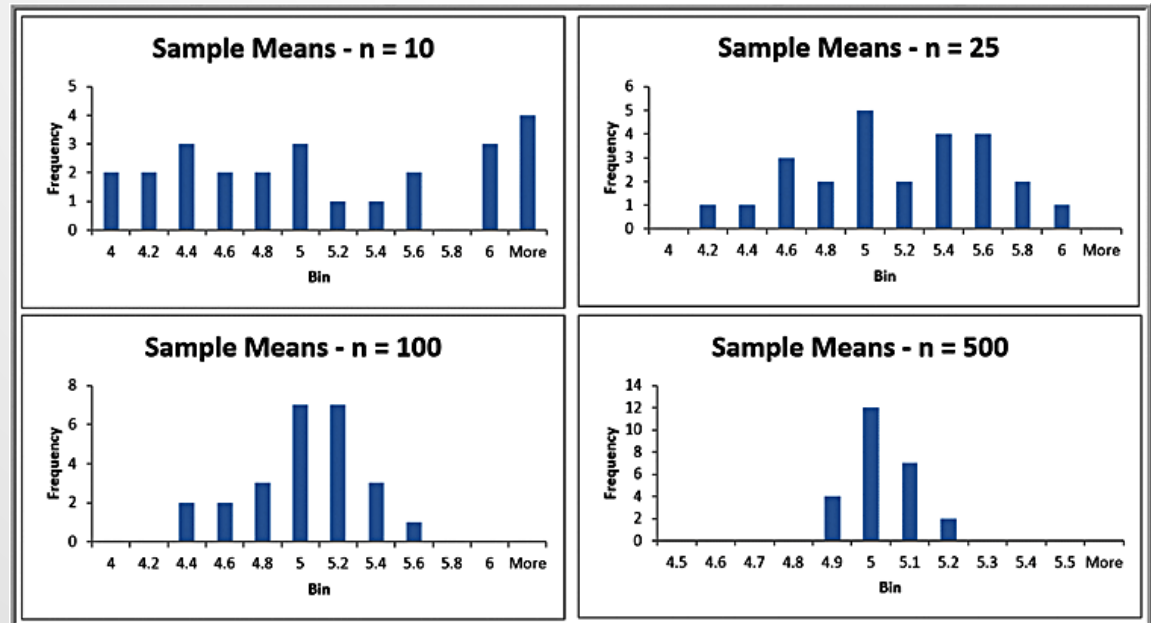
Note that the average of all the sample means is quite close the true population mean of 5.0.

Example 6.3: Other Sample Sizes

- Repeat the sampling experiment for samples of size 25, 100, and 500

As the sample size increases, the average of the sample means are all still close to the expected value of 5; however, the standard deviation of the sample means becomes smaller, meaning that the means of samples are clustered closer together around the true expected value. The distributions become normal.

Sample Size	Average of 25 Sample Means	Standard Deviation of 25 Sample Means
10	5.0108	0.816673
25	5.0779	0.451351
100	4.9173	0.301941
500	4.9754	0.078993



Example 6.4: Estimating Sampling Error Using the Empirical Rules

- ▶ Using the empirical rule for 3 standard deviations away from the mean, ~99.7% of sample means should be between:

[2.55, 7.45] for $n = 10$

[3.65, 6.35] for $n = 25$

[4.09, 5.91] for $n = 100$

[4.76, 5.24] for $n = 500$

Sample Size	Average of 25 Sample Means	Standard Deviation of 25 Sample Means
10	5.0108	0.816673
25	5.0779	0.451351
100	4.9173	0.301941
500	4.9754	0.078993

- ▶ As the sample size increases, the sampling error decreases.

Sampling Distributions

- ▶ The **sampling distribution of the mean** is the distribution of the means of all possible samples of a fixed size n from some population.
- ▶ The standard deviation of the sampling distribution of the mean is called the **standard error of the mean**:

$$\text{Standard Error of the Mean} = \sigma / \sqrt{n} \quad (6.1)$$

- ▶ As n increases, the standard error decreases.
 - Larger sample sizes have less sampling error.

Example 6.5: Computing the Standard Error of the Mean

- ▶ For the uniformly distributed population, we found $\sigma^2 = 8.333$ and, therefore, $\sigma = 2.89$

$$\text{Standard Error of the Mean} = \sigma / \sqrt{n} \quad (6.1)$$

Sample Size, n	Standard Error of the Mean
10	0.914
25	0.577
100	0.289
500	0.129

Central Limit Theorem

1. If the sample size is large enough, then the sampling distribution of the mean is:

- approximately normally distributed *regardless* of the distribution of the population
- has a mean equal to the population mean

2. If the population is normally distributed, then the sampling distribution is also normally distributed for *any* sample size.

- The central limit theorem allows us to use the theory we learned about calculating probabilities for normal distributions to draw conclusions about sample means.

Applying the Sampling Distribution of the Mean

- ▶ The key to applying sampling distribution of the mean correctly is to understand whether the probability that you wish to compute relates to an individual observation or to the mean of a sample.
 - If it relates to the mean of a sample, then you must use the sampling distribution of the mean, whose standard deviation is the standard error, not the standard deviation of the population.

Example 6.6: Using the Standard Error in Probability Calculations

- ▶ The purchase order amounts for books on a publisher's Web site is normally distributed with a mean of \$36 and a standard deviation of \$8.
- ▶ Find the probability that:
 - a) someone's purchase amount exceeds \$40.

Use the population standard deviation:

$$P(x > 40) = 1 - \text{NORM.DIST}(40, 36, 8, \text{TRUE}) = 0.3085$$

- b) the mean purchase amount for 16 customers exceeds \$40.

Use the standard error of the mean:

$$P(x > 40) = 1 - \text{NORM.DIST}(40, 36, 2, \text{TRUE}) = 0.0228$$

Interval Estimates

- ▶ An **interval estimate** provides a range for a population characteristic based on a sample.
 - Intervals specify a range of plausible values for the characteristic of interest and a way of assessing “how plausible” they are.
- ▶ In general, a $100(1 - \alpha)\%$ **probability interval** is any interval $[A, B]$ such that the probability of falling between A and B is $1 - \alpha$.
 - Probability intervals are often centered on the mean or median.
 - Example: in a normal distribution, the mean plus or minus 1 standard deviation describes an approximate 68% probability interval around the mean.

Example 6.7: Interval Estimates in the News

- ▶ A Gallup poll might report that 56% of voters support a certain candidate with a margin of error of $\pm 3\%$.
 - We would have a lot of confidence that the candidate would win since the interval estimate is [53%, 59%]
- ▶ Suppose the poll reported a 52% level of support with a $\pm 4\%$ margin of error.
 - We would be less confident in predicting a win for the candidate since the interval estimate is [48%, 56%].

Confidence Intervals

- ▶ A **confidence interval** is a range of values between which the value of the population parameter is believed to be, along with a probability that the interval correctly estimates the true (unknown) population parameter.
 - This probability is called the **level of confidence**, denoted by $1 - \alpha$, where α is a number between 0 and 1.
 - The level of confidence is usually expressed as a percent; common values are 90%, 95%, or 99%.
- ▶ For a 95% confidence interval, if we chose 100 different samples, leading to 100 different interval estimates, we would expect that 95% of them would contain the true population mean.

Confidence Interval for the Mean with Known Population Standard Deviation

$$\bar{x} \pm z_{\alpha/2}(\sigma / \sqrt{n}) \quad (6.2)$$

- ▶ Sample mean \pm margin of error
- ▶ Margin of error is: $\pm z_{\alpha/2}$ (standard error)
 - ▶ $z_{\alpha/2}$ is the value of the standard normal random variable for an upper tail area of $\alpha/2$ (or a lower tail area of $1 - \alpha/2$).
 - $z_{\alpha/2}$ is computed as =NORM.S.INV($1 - \alpha/2$)
 - Example: if $\alpha = 0.05$ (for a 95% confidence interval), then NORM.S.INV(0.975) = 1.96;
 - Example: if $\alpha = 0.10$ (for a 90% confidence interval), then NORM.S.INV(0.95) = 1.645,
- ▶ The margin of error can also be computed by =CONFIDENCE.NORM(*alpha, standard_deviation, size*).

Example 6.8: Computing a Confidence Interval with a Known Standard Deviation

- ▶ A production process fills bottles of liquid detergent. The standard deviation in filling volumes is constant at 15 mls. A sample of 25 bottles revealed a mean filling volume of 796 mls.
- ▶ A 95% confidence interval estimate of the mean filling volume for the population is

$$\begin{aligned} & \bar{x} \pm z_{\alpha/2} (\sigma/\sqrt{n}) \\ & = 796 \pm 1.96(15/\sqrt{25}) = 796 \pm 5.88, \text{ or } [790.12, 801.88] \end{aligned}$$

Excel Workbook for Confidence Intervals

- ▶ The worksheet *Population Mean Sigma Known* in the Excel workbook *Confidence Intervals* computes this interval using the CONFIDENCE.NORM function

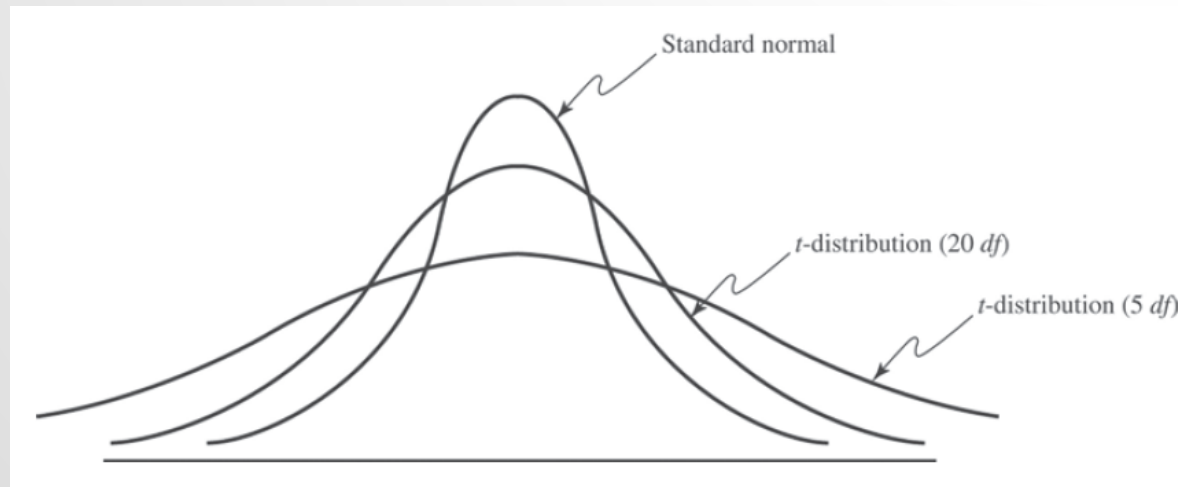
	A	B	C	D	E	F
1	Confidence Interval for Population Mean, Standard Deviation Known					
2						
3	Alpha	0.05				
4	Standard deviation	15				
5	Sample size	25				
6	Sample average	796				
7						
8	Confidence Interval	95%				
9	Error	5.879892				
10	Lower	790.1201				
11	Upper	801.8799				

Confidence Interval Properties

- ▶ As the level of confidence, $1 - \alpha$, decreases, $z_{\alpha/2}$ decreases, and the confidence interval becomes narrower.
 - For example, a 90% confidence interval will be narrower than a 95% confidence interval. Similarly, a 99% confidence interval will be wider than a 95% confidence interval.
- ▶ Essentially, you must trade off a higher level of accuracy with the risk that the confidence interval does not contain the true mean.
 - To reduce the risk, you should consider increasing the sample size.

The t-Distribution

- ▶ The t-distribution is a family of probability distributions with a shape similar to the standard normal distribution. Different t-distributions are distinguished by an additional parameter, **degrees of freedom (df)**.
 - As the number of degrees of freedom increases, the t-distribution converges to the standard normal distribution



Confidence Interval for the Mean with Unknown Population Standard Deviation

$$\bar{x} \pm t_{\alpha/2, n-1}(s/\sqrt{n}) \quad (6.3)$$

where $t_{\alpha/2}$ is the value of the t -distribution with $df = n - 1$ for an upper tail area of $\alpha/2$.

- ▶ t values are found in Table 2 of Appendix A or with the Excel function `T.INV(1 - $\alpha/2$, $n - 1$)`.
- ▶ The Excel function
`=CONFIDENCE.T(alpha, standard_deviation, size)`
can be used to compute the margin of error

Example 6.9: Computing a Confidence Interval with Unknown Standard Deviation

- ▶ Excel file *Credit Approval Decisions*. Find a 95% confidence interval estimate of the mean revolving balance of homeowner applicants (first, sort the data by homeowner).
- ▶ Sample mean = \$12,630.37; $s = \$5393.38$; standard error = \$1037.96; $t_{0.025, 26} = 2.056$.

$$\bar{x} \pm t_{\alpha/2, n-1}(s/\sqrt{n}) \quad (6.3)$$

$$12,630.37 \pm 2.056(5393.38/\sqrt{27})$$

	A	B	C	D	E
1	Confidence Interval for Population Mean, Standard Deviation Unknown				
2					
3	Alpha	0.05			
4	Sample standard deviation	5393.38			
5	Sample size	27			
6	Sample average	12630.37			
7					
8	Confidence Interval	95%			
9		t-value	2.056		
10		Error	2133.55		
11		Lower	10496.82		
12		Upper	14763.92		

Confidence Interval for a Proportion

- ▶ An unbiased estimator of a population proportion π (this is not the number $\pi = 3.14159 \dots$) is the statistic $\hat{p} = x / n$ (the sample proportion), where x is the number in the sample having the desired characteristic and n is the sample size.
- ▶ A $100(1 - \alpha)\%$ confidence interval for the proportion is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (6.4)$$

Example 6.10: Computing a Confidence Interval for a Proportion

- ▶ Excel file *Insurance Survey*. We are interested in the proportion of individuals who would be willing to pay a lower premium for a higher deductible for their health insurance.
 - Sample proportion = $6/24 = 0.25$.
- ▶ Confidence interval:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (6.4)$$

$$0.25 \pm 1.96 \sqrt{\frac{0.25(0.75)}{24}} = 0.25 \pm 0.173, \text{ or } [0.077, 0.423]$$

	A	B
1	Confidence Interval for a Proportion	
2		
3	Alpha	0.05
4	Sample proportion	0.25
5	Sample size	24
6		
7	Confidence Interval	95%
8	z-value	1.96
9	Standard error	0.088388
10	Lower	0.076762
11	Upper	0.423238

Example 6.11: Drawing a Conclusion about a Population Mean Using a Confidence Interval

- ▶ In Example 6.8, the required volume for the bottle-filling process is 800 and the sample mean is 796 mls. We obtained a confidence interval [790.12, 801.88]. Should machine adjustments be made?

Although the sample mean is less than 800, the sample does not provide sufficient evidence to draw that conclusion that the population mean is less than 800 because 800 is contained within the confidence interval.

	A	B	C	D	E	F
1	Confidence Interval for Population Mean, Standard Deviation Known					
2						
3	Alpha	0.05				
4	Standard deviation	15				
5	Sample size	25				
6	Sample average	796				
7						
8	Confidence Interval	95%				
9	Error	5.879892				
10	Lower	790.1201				
11	Upper	801.8799				

Example 6.12: Using a Confidence Interval to Predict Election Returns

- ▶ An exit poll of 1,300 voters found that 692 voted for a particular candidate in a two-person race. This represents a proportion of 53.23% of the sample. Could we conclude that the candidate will likely win the election?
- ▶ A 95% confidence interval for the proportion is $[0.505, 0.559]$. This suggests that the population proportion of voters who favor this candidate is highly likely to exceed 50%, so it is safe to predict the winner.
- ▶ If the sample proportion is 0.515, the confidence interval for the population proportion is $[0.488, 0.543]$. Even though the sample proportion is larger than 50%, the sampling error is large, and the confidence interval suggests that it is reasonably likely that the true population proportion could be less than 50%, so you cannot predict the winner.

Prediction Intervals

- ▶ A **prediction interval** is one that provides a range for predicting the value of a new observation from the same population.
 - A confidence interval is associated with the sampling distribution of a statistic, but a prediction interval is associated with the distribution of the random variable itself.
- ▶ A $100(1 - \alpha)\%$ prediction interval for a new observation is

$$\bar{x} \pm t_{\alpha/2, n-1} \left(s \sqrt{1 + \frac{1}{n}} \right) \quad (6.5)$$

Example 6.13: Computing a Prediction Interval

- ▶ Compute a 95% prediction interval for the revolving balances of customers (*Credit Approval Decisions*)

$$\bar{x} \pm t_{\alpha/2, n-1} \left(s \sqrt{1 + \frac{1}{n}} \right) \quad (6.5)$$

$$\begin{aligned} & \$12,630.37 \pm 2.056(\$5,393.38) \sqrt{1 + \frac{1}{27}}, \text{ or} \\ & [\$338.10, \$23,922.64] \end{aligned}$$

Confidence Intervals and Sample Size

- ▶ We can determine the appropriate sample size needed to estimate the population parameter within a specified level of precision ($\pm E$).
- ▶ Sample size for the mean:

$$n \geq (z_{\alpha/2})^2 \frac{\sigma^2}{E^2} \quad (6.6)$$

- ▶ Sample size for the proportion:

$$n \geq (z_{\alpha/2})^2 \frac{\pi(1 - \pi)}{E^2} \quad (6.7)$$

- Use the sample proportion from a preliminary sample as an estimate of π or set $p = 0.5$ for a conservative estimate to guarantee the required precision.


Example 6.14: Sample Size Determination for the Mean

- ▶ In Example 6.8, the sampling error was ± 5.88 mls.
- ▶ What sample size is needed to reduce the margin of error to at most 3 mls?

$$n \geq (z_{\alpha/2})^2 \frac{(\sigma^2)}{E^2}$$
$$= (1.96)^2 \frac{(15^2)}{3^2} = 96.04$$

Round up to
97 samples.

	A	B	C	D	E	F
1	Confidence Interval for Population Mean, Standard Deviation Known					
2						
3	Alpha	0.05				
4	Standard deviation	15				
5	Sample size	97				
6	Sample average	796				
7						
8	Confidence Interval	95%				
9	Error	2.985063				
10	Lower	793.0149				
11	Upper	798.9851				



Example 6.15: Sample-Size Determination for a Proportion

- ▶ For the voting example we discussed, suppose that we wish to determine the number of voters to poll to ensure a sampling error of at most $\pm 2\%$. With no information, use $\pi = 0.5$:

$$\begin{aligned} n &\geq (z_{\alpha/2})^2 \frac{\pi(1 - \pi)}{E^2} \\ &= (1.96)^2 \frac{(0.5)(1 - 0.5)}{0.02^2} = 2,401 \end{aligned}$$

Google Analytics Project 2

Email me (albert.kalim@asbury.edu)

your answers by Sunday, 6/5, 11:59 p.m. ET (10 points total)

Log in to your Google Analytics dashboard [here](#) and click Sign in to Analytics.

After you logged in, focus on the Acquisition tools on the left-hand side. **Tell me as much as you can any information about the acquisition for the month of April 2022.**

You will be graded based on your familiarity with the tools and how you read/interpret the data.