



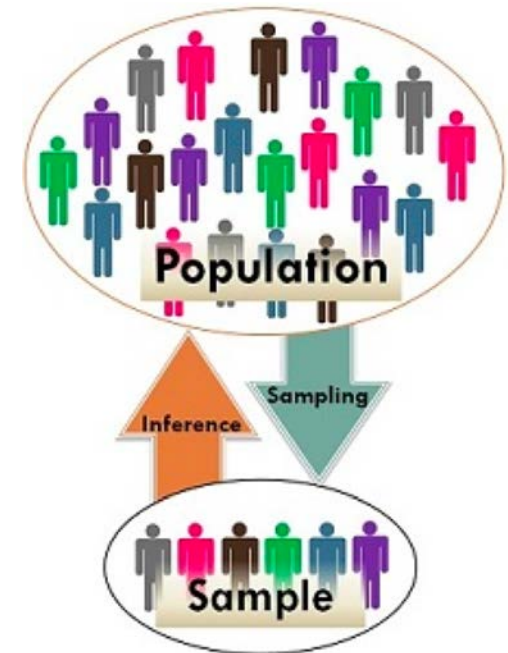
# Sampling and Estimation

# Learning Objectives

- Explain the importance of unbiased estimators
- Describe the difference between sampling error and nonsampling error
- Explain how the average, standard deviation, and distribution of means of samples changes as sample size increases
- Define sampling distribution of mean and calculate standard error of mean
- Explain practical importance of central limit theorem and how confidence intervals change as level of confidence changes
- Explain difference between point estimate and interval estimate
- Be able to compute confidence intervals for population means and proportions and to use confidence intervals to draw conclusions about population parameters
- Compute prediction interval and explain how it differs from confidence interval
- Compute sample sizes needed to ensure a confidence interval for means and proportions with a specified margin of error

# Statistical Sampling

- **Sampling:** foundation of statistical analysis.
- **Estimators:** measures used to estimate unknown population parameters
- **Point estimate: single number** derived from a sample that is used to estimate the value of a population parameters, egs:
  - **Mean:**  $\bar{x}$  is a point estimate of  $\mu$
  - **Standard deviation:**  $s$  is a point estimate of  $\sigma$



# Unbiased Estimators

**Population Mean:**  $\mu = \frac{\sum_{i=1}^N x_i}{N}$

**Sample Mean:**  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

**Population SD:**  $\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$

**Sample SD:**  $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$

- Sample mean is a **good point estimate** for population mean
- Sample variance has **different denominator** as population variance. Why?
- **Unbiased** estimator: expected value of the estimator **equals** population parameter.

# Sampling Error

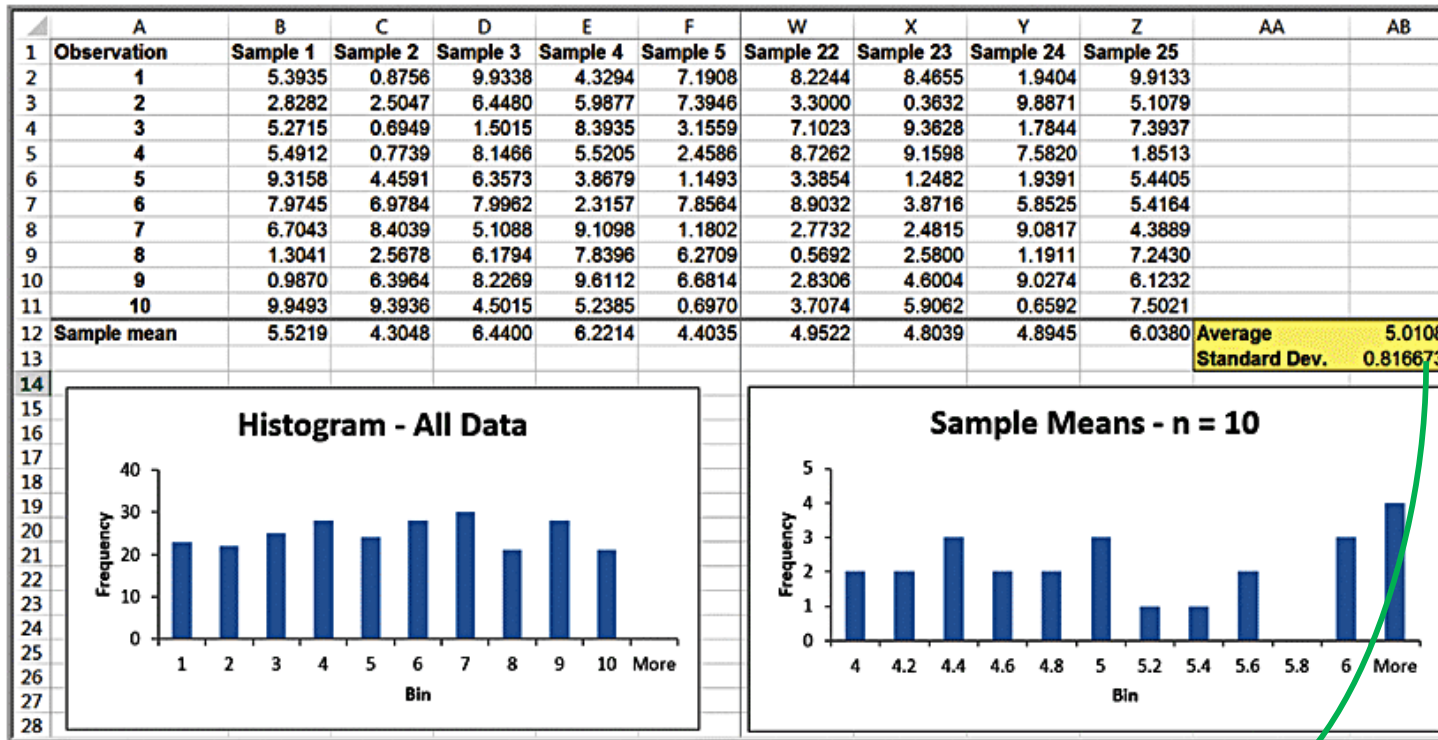
- **Sampling (statistical) error** occurs because samples are only a subset of the total population
  - Sampling error depends on the **size** of the sample relative to the population.
- **Nonsampling error** occurs when the sample does not adequately represent the target population.
  - Nonsampling error usually results from a poor sample design or choosing the wrong population frame. (e.g., convenience sample)

# A Sampling Experiment Example

- 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.



# A Sampling Experiment Example - results



Note: Average of all sample means is quite close to true population mean of 5.0.

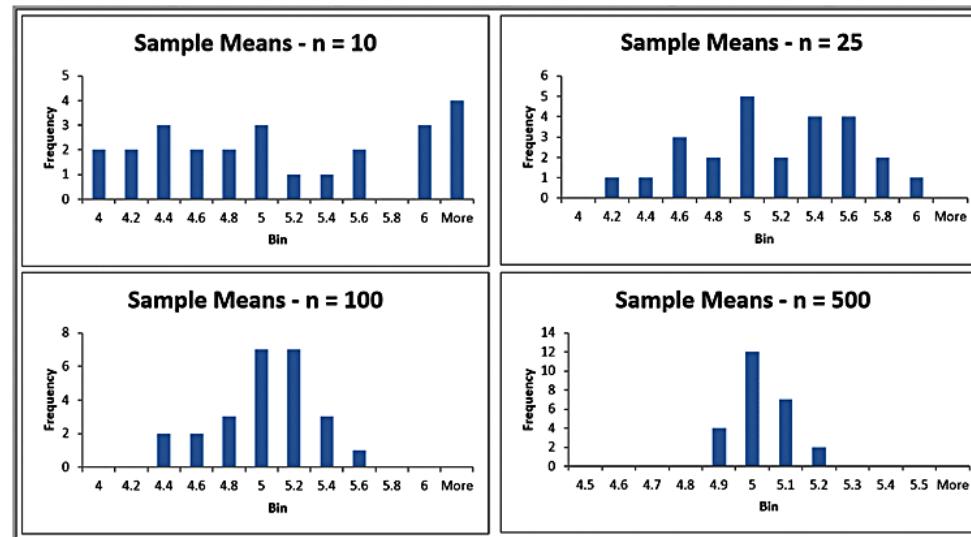
# A Sampling Experience: Other Sample Sizes

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

As **sample size increases**,

- average of sample **means** are all **still close to expected value of 5**;
- however, **standard deviation** of sample means becomes **smaller**, meaning that the means of samples are clustered closer together around the true expected value.
- 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                              |





# Estimating Sampling Error Using 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 sample size increases, sampling error decreases.

# Sampling Distributions

- Sampling distribution of the mean:
  - Distribution of means of all possible samples of a fixed size  $n$  from some population.
- Standard error of the mean:
  - Standard deviation of sampling distribution of the mean

$$\text{Standard Error of the Mean} = \sigma / \sqrt{n}$$

- As  $n$  increases, standard error decreases.

# Central Limit Theorem

1. If sample size is large enough, **sampling distribution of the mean**:
  - is approximately **normally distributed** *regardless* of population distribution
  - has a mean equal to population mean
2. If **population is normally distributed**, **sampling distribution is also normally distributed** for **any** sample size.

Central limit theorem allows us to calculate probabilities for normal distributions to draw conclusions about sample means.

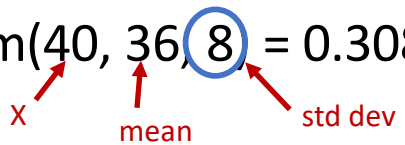
When calculating probabilities, determine whether it is related to an individual observation or mean of a sample (std dev is the std error  $\sigma/\sqrt{n}$  ).

# Using 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:


$$P(x > 40) = 1 - \text{pnorm}(40, 36, 8) = 0.3085$$



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

Use the standard error of the mean:

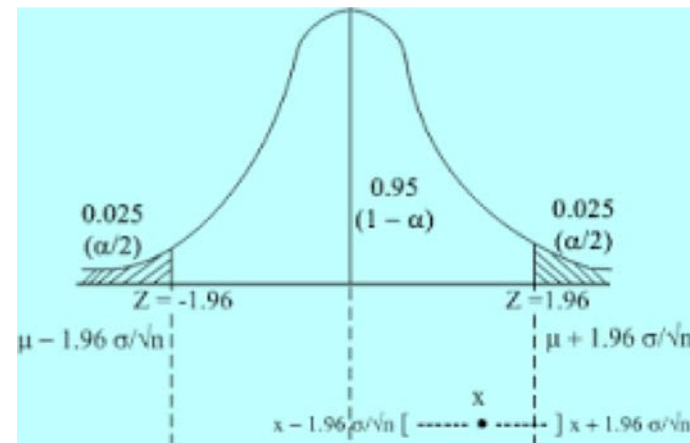
$$P(x > 40) = 1 - \text{pnorm}(40, 36, 2) = 0.0228$$


$$\text{std error} = \left( \frac{\sigma}{\sqrt{n}} \right) = \left( \frac{8}{\sqrt{16}} \right)$$

# Interval Estimates

- 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.
- $100(1 - \alpha)\%$  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 mean or median.

- Example: in a normal distribution, mean  $\pm 1$  sd describes an approximate 68% probability interval around the mean



# 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%]

**How confident are you that the candidate will win?**

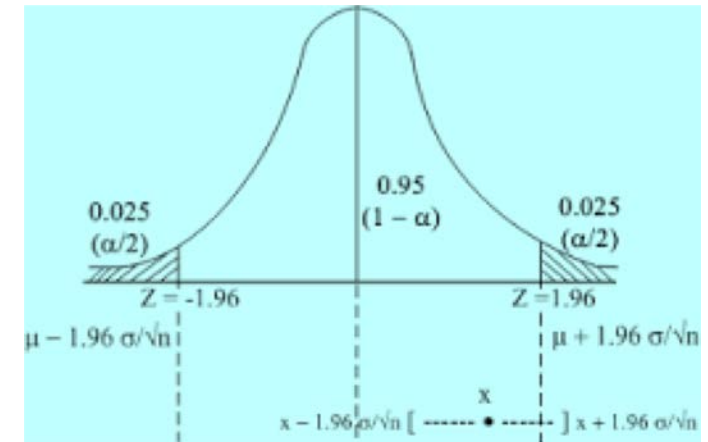
- Suppose the poll reported a 52% level of support with a  $\pm 4\%$  margin of error
  - We would be less confident in predicting a win for the candidate since the interval estimate is [48%, 56%]

**How confident are you that the candidate will win?**



# Confidence Intervals

- **confidence interval**: 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  $\alpha$  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

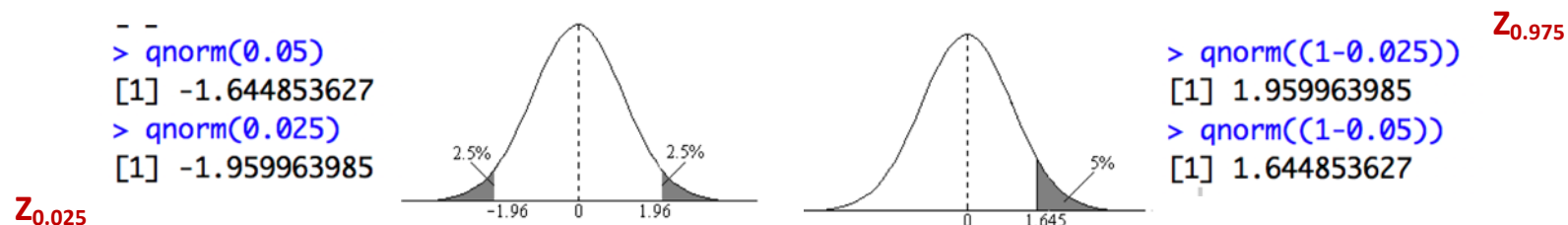
$$\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$$

► Sample mean  $\pm$  margin of error

► Margin of error is:  $z_{\alpha/2}$  (standard error)

►  $z_{\alpha/2}$ : value of standard normal random variable for an upper tail area of  $\alpha/2$  (or a lower tail area of  $1 - \alpha/2$ ).

- Example: if  $\alpha = 0.05$  (for a 95% confidence interval), then  $z_{0.975} = 1.96$ ;
- Example: if  $\alpha = 0.10$  (for a 90% confidence interval), then  $z_{0.95} = 1.645$ .



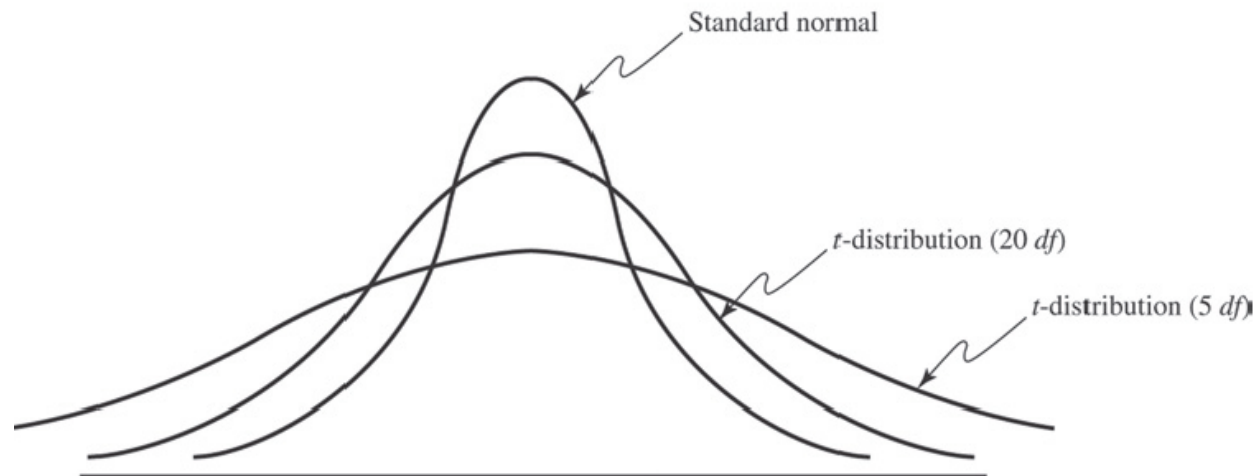
# Computing a Confidence Interval with a Known Standard Deviation

- 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

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

# The t-Distribution

- (Student's)  $t$ -Distribution
- Used for **confidence intervals** when the **population standard deviation is unknown**.
- Its only parameter is the degrees of freedom ( $df$ ) [*no. of sample values – no. of est parameters*]



# Confidence Interval for the Mean with Unknown Population Standard Deviation

► Formula:

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

►  $t_{\alpha/2, n-1}$  : value from **t-distribution** with (n-1) degrees of freedom, giving an upper tail probability of  $\alpha/2$  (or a lower tail area of  $1 - \alpha/2$ ).

- Example: if  $\alpha = 0.05$ ,  $n=30$  (for a 95% confidence interval), then  $t_{0.975, 29} = 2.05$ ;
- Example: if  $\alpha = 0.10$  (for a 90% confidence interval), then  $t_{0.95, 29} = 1.70$ .

**T Dist, n=30**

```
> qt((1-0.025), df=29)
[1] 2.045229642
> qt((1-0.05), df=29)
[1] 1.699127027
```

**T Dist, n=100**

```
> qt((1-0.025), df=99)
[1] 1.984216952
> qt((1-0.05), df=99)
[1] 1.660391156
```

**T Dist, n=1000**

```
> qt((1-0.025), df=999)
[1] 1.962341461
> qt((1-0.05), df=999)
[1] 1.646380345
```

**Normal Dist**

```
> qnorm((1-0.025))
[1] 1.959963985
> qnorm((1-0.05))
[1] 1.644853627
```

# Computing a Confidence Interval with Unknown Standard Deviation

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

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

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

```
View(Credit_Approval_Decisions)
df1<-NA
df1<-Credit_Approval_Decisions
home.data<-subset(df1,Homeowner=`Y`)
mRB<-mean(home.data$`Revolving Balance`)
sdRB<-sd(home.data$`Revolving Balance`)
seRB<- sdRB/sqrt(nrow(home.data))
uCIRB<-mRB+(qt(0.975,df=26)*seRB)
lCIRB<-mRB-(qt(0.975,df=26)*seRB)
```

|    | 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  $\pi$  (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

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

# Computing a Confidence Interval for a Proportion

- Datafile *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:

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

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

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

```
View(Insurance_Survey)
df2<-NA
df2<-Insurance_Survey
xdata<-subset(df2,`Premium/Deductible**`=="Y")
phat<-nrow(xdata)/nrow(df2)
LCIphat<-phat+(qnorm(0.025)*sqrt(phat*(1-phat)/nrow(df2)))
UCIphat<-phat-(qnorm(0.025)*sqrt(phat*(1-phat)/nrow(df2)))
```

## variable definition

Satisfaction: Measured from 1-5 with 5 being highly satisfied.

Premium/Deductible: Would you be willing to pay a lower premium for a higher deductible?

# Using C.I. for decision making

- In the earlier eg on slide 17, required volume for the bottle-filling process is 800 and the sample mean is 796 mls. We obtained a confidence interval for population mean of [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 |   |   |   |   |

If mean = 792, CI = [786.12, 797.88]

→ quite sure machine will not perform to standard and be able to fill up to 800.

# Using a Confidence Interval to Predict Election Results

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?

- 95% confidence interval for the proportion is [0.505, 0.559]  
→ population proportion of voter favoring candidate highly likely to exceed 50%
- What if the sample proportion is 0.515, and the confidence interval for the population proportion is [0.488, 0.543]?  
→ true population proportion could be less than 50%, so you cannot predict the winner

# Prediction Intervals

- Prediction interval: a range for predicting value of a new observation from same population.
- While confidence interval is associated with sampling distribution of a statistic, a prediction interval is associated with the distribution of random variable itself.
- $100(1 - \alpha)\%$  prediction interval for a new observation is

---

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

# Computing a Prediction Interval

- Compute a 95% prediction interval for the revolving balances of customers (*Credit Approval Decisions*)
- Sample mean = \$12,630.37;  $s = \$5393.38$ ; standard error = \$1037.96;  $t_{0.025, 26} = 2.056$ .

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

---

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



# Confidence Intervals and Sample Size

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

- Recall

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

- Sample size for the **mean**: 
$$n \geq (z_{\alpha/2})^2 \frac{\sigma^2}{E^2}$$

- sample size of **proportion**: 
$$E \geq z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}}$$

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

- Use the sample proportion from a preliminary sample as an estimate of  $\pi$  or set  $\pi = 0.5$  for a conservative estimate to guarantee the required precision (maximizes qty of  $\pi(1-\pi)$ ).

# Sample Size Determination for the Mean

[790.12,801.88]

- In liquid detergent example, **sampling error** was  $\pm 5.88$  mls, sd=15
- 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  | <b>Confidence Interval for Population Mean, Standard Deviation Known</b> |          |   |   |   |   |
| 2  |  |          |   |   |   |   |
| 3  | <b>Alpha</b>   | 0.05     |   |   |   |   |
| 4  | <b>Standard deviation</b>  | 15       |   |   |   |   |
| 5  | <b>Sample size</b>   | 97       |   |   |   |   |
| 6  | <b>Sample average</b>  | 796      |   |   |   |   |
| 7  |  |          |   |   |   |   |
| 8  | <b>Confidence Interval</b>   | 95%      |   |   |   |   |
| 9  | <b>Error</b>   | 2.985063 |   |   |   |   |
| 10 | <b>Lower</b>   | 793.0149 |   |   |   |   |
| 11 | <b>Upper</b>   | 798.9851 |   |   |   |   |

# Sample-Size Determination for a Proportion

- Voting example: determine number of voters to poll to ensure a **sampling error of at most  $\pm 2\%$** . With no information, use  $\pi = 0.5$  (proportion who poll):

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