

# INTERVAL ESTIMATION

## 1. Confidence Intervals For The Mean of a Normal Distribution:

Consider the problem of estimating the mean yearly salaries of Northwest hospital administrators. Suppose the yearly salaries of administrators in the Northwest are  $\mathcal{N}(\mu = ?, \sigma = 12)$  in thousands of dollars.

In order to get some idea of  $\mu$  we must take a sample and obtain either a point estimate or an interval estimate of  $\mu$ .

What would be the point estimate? Suppose we take a sample of  $n=25$  administrators and  $\bar{\mathbf{X}} = 82$ . Our point estimate of  $\mu$  is \$82,000. Now, what if we want an interval estimate?

### **Confidence Interval Estimator**

A **confidence interval estimator** for a population parameter  $\theta$  is a rule for determining (based on sample information) a range, or interval that is likely to include the parameter. The corresponding estimate is called a **confidence interval estimate**

### **Confidence Interval and Confidence Level**

Let  $\theta$  be an unknown parameter. Suppose that on the basis of sample information, random variables  $A$  and  $B$  are found such that  $P(A < \theta < B) = 1 - \alpha$ , where  $\alpha$  is any number between 0 and 1. If the specific sample values of  $A$  and  $B$  are  $a$  and  $b$ , then the interval from  $a$  to  $b$  is called a  $100(1 - \alpha)\%$  **confidence interval** of  $\theta$ . The quantity  $1 - \alpha$  is called the **confidence level** of the interval.

If the population were repeatedly sampled a very large number of times, the true value of the parameter  $\theta$  would be contained in  $100(1 - \alpha)\%$  of intervals calculated this way. The confidence interval calculated in this manner is written as  $a < \theta < b$  with  $100(1 - \alpha)\%$  confidence.

With the definitions of confidence intervals in hand, let's return to our hospital administrator problem. Suppose we now wish to find an interval estimate for the true mean salary of the Northwest hospital administrators.. That is, based on our data, we want 2 values such that we can say, we are very certain that  $\mu$  falls between these 2 values. Of course, these values depend on your data.

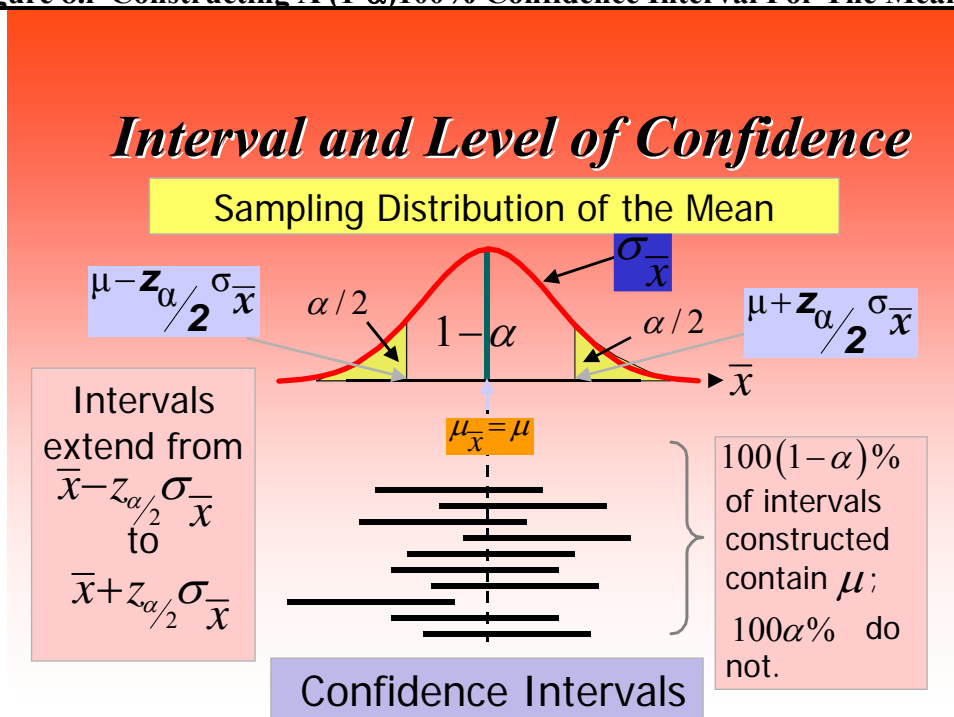
Let's say for now, pretty certain means 95% certain. Logically, we might consider an interval of the form  $\bar{\mathbf{X}} - c < \mu < \bar{\mathbf{X}} + c$  such that  $P(\bar{\mathbf{X}} - c < \mu < \bar{\mathbf{X}} + c) = .95$ . So  $A = \bar{\mathbf{X}} - c$  and  $B = \bar{\mathbf{X}} + c$  and  $1 - \alpha = .95$ .

The question becomes clear, we want to add and subtract some number to and from our point estimator that gives us the proper amount of certainty. That number should depend on the **variability in the population, how certain we want to be, and the sample size**. Logically and we will see mathematically, that number  $c$  is  $z$  standard errors, where  $z$  is determined by how certain we want to be.

As long as we are sampling from a normal population or by Central Limit Theorem, if  $n$  is large,  $\bar{X} \sim \mathcal{N}(\mu, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n})$  and so  $c = z \frac{\sigma}{\sqrt{n}}$  and  $z=1.96$ .

If generally, we wish to be certain with a probability of  $1-\alpha$  the picture would look like Figure 8.1.

**Figure 8.1 Constructing A  $(1-\alpha)100\%$  Confidence Interval For The Mean**



**Notation** Let  $Z_{\alpha/2}$  be the number for which  $P(Z > Z_{\alpha/2}) = \alpha/2$  where the random variable  $Z$  follows a standard normal distribution.

So, depending on how certain we want to be, the values of  $z$  change. Note for 95% certain, the  $z$ -value is 1.96 as we said it would be.

**Table 8.2 Appropriate Z-Values For Confidence Intervals For Means**

$\alpha$	0.01	0.02	0.05	0.10
$Z_{\alpha/2}$	2.58	2.33	1.96	1.645
Confidence Level	99%	98%	95%	90%

Our problem can now be completely generalized and a procedure defined.

**Confidence Intervals for the Mean of a Population that is Normally Distributed: Population Variance Known**

Consider a random sample of  $n$  observations from a normal distribution with mean  $\mu$  and known variance  $\sigma^2$ . If the sample mean is  $\bar{X}$ , then a  $100(1-\alpha)\%$  confidence interval for the population mean with known variance is given by

$$\bar{X} - \frac{Z_{\alpha/2}\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{Z_{\alpha/2}\sigma}{\sqrt{n}} \quad (8.1)$$

or equivalently,  $\bar{X} \pm B$  where the **margin of error** (also called the sampling error, the bound, or the interval half width) is given by

$$B = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (8.2)$$

If the population were repeatedly sampled a very large number of times, the true value of the parameter  $\mu$  would be contained in 95% of intervals calculated this way. So, if we were fortunate enough to get a good sample (one of the 95%), the true mean is really between the values we computed. Remember, the probability is on the sample you collect. 95% of the time, you will get a sample such that the true mean will be between the values you computed. However, you never know and that is why your answer comes with this 95% certainty attached to it.

Example - Suppose the time it takes a worker to complete a task has a mean  $\mu=?$  and standard deviation  $\sigma= 2$  minutes. If we want a 99% confidence interval for  $\mu$  based on  $n=100$  observations, what do we use?

Since  $n = 100$  is "large", by CLT  $\bar{X}$  approx~Normal so, our confidence interval will be

$$\bar{x} \pm (2.58) \frac{2}{\sqrt{100}} \equiv \bar{x} \pm .516$$

**In every problem, we first set up our confidence interval. Next, we collect the data and compute the sample mean. Suppose for our problem, the sample mean is  $\bar{x} = 20$  minutes. Then, we are 99% certain  $19.4848 < \mu < 20.5152$**

Suppose that the company manager had claimed that workers take 22 minutes, on the average to complete the task. What does this confidence interval say about this claim? Since we are 99% that the true mean is between 19.4848 and 20.5152 minutes, we can say that the data provides evidence to say the claim is incorrect.

Suppose now that the company manager had claimed that workers take 19.75 minutes, on the average to complete the task. What does this confidence interval say about this claim? Since we are 99% that the true mean is between 19.4848 and 20.5152 minutes, we can say the manager's claim cannot be proven wrong. His claim is a possible value for the average.

If we want to be 95% certain, we have  $\bar{x} \pm (1.96) \frac{2}{\sqrt{100}} \equiv \bar{x} \pm .392$

and we are 95%  $19.608 < \mu < 20.392$

What does this tell us about confidence intervals?

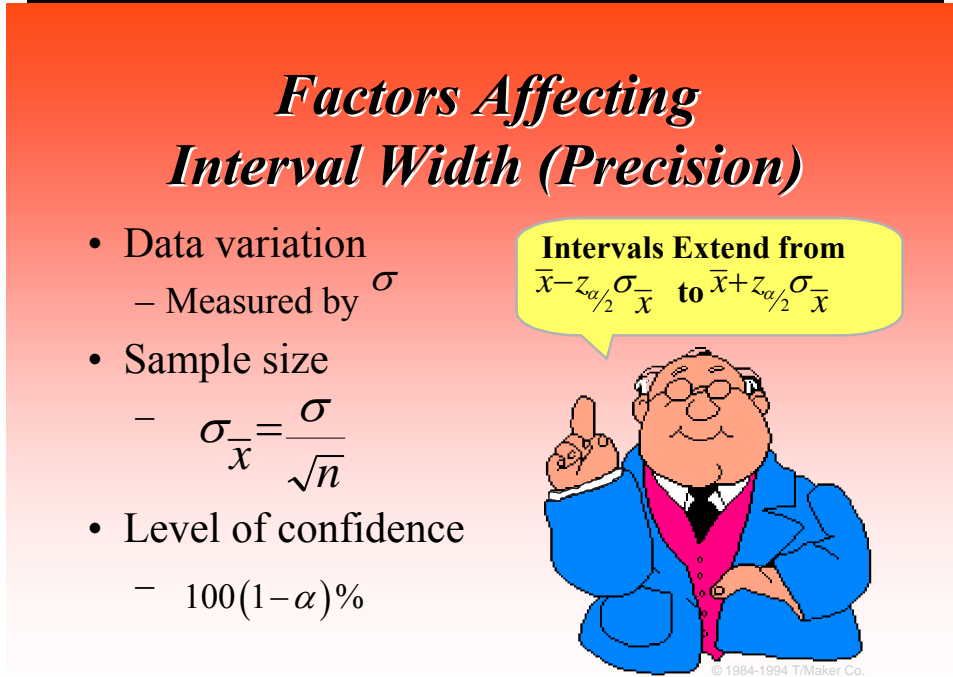
*If we want to be more certain, we will have a bigger confidence interval and hence we will know less about our parameter.*

*If we want a better confidence interval we can take a larger sample that will give us a narrower confidence interval with the same amount of probability.*

*Lastly, if we are sampling from populations with less variability, then the confidence interval width will be smaller and we will have a better confidence interval.*

These factors are summarized in figure 8.2

**Figure 8.2 Factors Affecting the Precision of a Confidence Interval**



## 2. Confidence Intervals for the Mean of a Normal Distribution: Population Variance Unknown

### Student's t Distribution

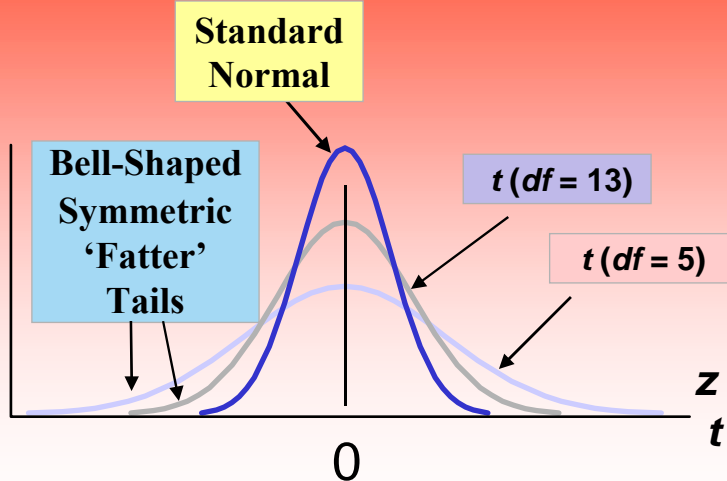
When the variance of a population is unknown, we must seek a different method because we cannot use  $s$  in our formula for the confidence interval.

To do this, we use the fact that if  $X_1, X_2, \dots, X_n$  is a random sample  $\sim N(\mu, \sigma^2)$  then the statistic  $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$  follows a t distribution with  $n-1$  degrees of freedom.

A t distribution looks very similar to a standard normal except the tails are fatter (that is it has more variability). The variability of the t distribution depends on the degrees of freedom. The variability is inversely proportionate to the degrees of freedom. Hence, the bigger the sample size, the bigger the degrees of freedom, the smaller the variability of the t distribution. This should make sense, because if we sample more, we know more, and hence the variability should be reduced. We can see this in Figure 8.3

**Figure 8.3 The Student t distribution**

## *Student's t Distribution*



### *Student's t Distribution*

Given a random sample of  $n$  observations, with mean  $\bar{X}$  and standard deviation  $s$ , from a normally distributed population with mean  $\mu$ , the random variable  $t$  follows the Student's  $t$  distribution with  $(n - 1)$  degrees of freedom and is given by  $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

Based on the above  $t$ -statistic and Figure 8.3, it is not surprising, that the confidence interval for the mean when the standard deviation is unknown is the same as the confidence interval for the mean when the standard deviation is known, except that *instead of  $z$  in the procedure with have  $t$  and instead of  $\sigma$  we have  $s$ .*

### **Population Variance Unknown**

#### ***Confidence Intervals for the Mean of a Normal Population: Population Variance Unknown***

Suppose there is a random sample of  $n$  observations from a *normal distribution* with mean  $\mu$  and unknown variance. If the sample mean and standard deviation are, respectively,  $\bar{X}$  and  $s$ , then a **100(1- $\alpha$ )% confidence interval for the population mean, variance unknown**, is given by

$$\bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \quad (8.3)$$

or equivalently,

$$\bar{X} \pm B$$

where the **margin of error**, the sampling error, or the bound,  $B$ , is given by

$$B = t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \quad (8.4)$$

and  $t_{n-1, \alpha/2}$  is the number for which  $P(t_{n-1} > t_{n-1, \alpha/2}) = \alpha/2$   
 The random variable  $t_{n-1}$  has a Student's  $t$  distribution with  $\nu = (n - 1)$  degrees of freedom.

### Notation

A random variable having the Student's  $t$  distribution with  $\nu$  degrees of freedom will be denoted  $t_\nu$ . The  $t_{\nu, \alpha/2}$  is defined as the number for which  $P(t_\nu > t_{\nu, \alpha/2}) = \alpha/2$

## 3. Confidence Intervals for the Population Proportion (Large Samples)

When we are dealing with qualitative data, we want to estimate the true proportion of successes,  $\pi$  in a population. Most often we are taking a large enough sample to apply the normal approximation to the binomial, which allows  $Z = \frac{p - \pi}{\sqrt{\pi(1 - \pi) / n}}$ . Proceeding as we did in the previous sections, and estimating  $\pi$  by  $p$  (the sample proportion) where necessary, we have the following  $(1 - \alpha)100\%$  confidence intervals for the true proportion of success.

### Confidence Intervals for Population Proportion (Large Samples)

Let  $p$  denote the observed proportion of "successes" in a random sample of  $n$  observations from a population with a proportion  $\pi$  of successes. Then, if  $n$  is large enough that  $n\pi(1 - \pi) > 9$ , then a  $100(1 - \alpha)\%$  **confidence interval for the population proportion** is given by

$$p - Z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}} < \pi < p + Z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}} \quad (8.5)$$

or equivalently,

$$p \pm B$$

where the **margin of error**, the sampling error or the bound,  $B$ , is given by

$$B = Z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}} \quad (8.6)$$

and  $Z_{\alpha/2}$ , is the number for which a standard normal variable  $Z$  satisfies  $P(Z > Z_{\alpha/2}) = \alpha/2$

Example) Suppose a physical fitness company wishes to know the true proportion,  $\pi$ , of workers who are overweight. Find a 95% confidence interval for  $\pi$ .

**To do this, they will take a sample of  $n=250$  workers is taken and we observe**

$x = \#$  overweight to be 180. So,  $p = \text{sample proportion} = \frac{180}{250} = .72$ .

$p = .72$  is our point estimate of  $\pi$

**Based on this data, can we find a 95% confidence interval for  $p$ ?**

$$.72 \pm 1.96 \sqrt{\frac{.72(1-.72)}{250}} \equiv .72 \pm .0284 \equiv .664336 < p < .77564$$

**We are 95% that between 66.43% and 77.56% of the workers are overweight. Hence the company will institute a rigorous fitness program for its own workers.**

Lastly, when sampling from one population, we will now consider **confidence intervals for the variance of a population.**

#### 4. Confidence Intervals for the Variance of a Normal Population

Suppose  $X_1, X_2, \dots, X_n$  i.i.d.  $N(\mu, \sigma^2)$  then we already showed that  $S^2 = \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \right]$  is an

unbiased estimator for  $s^2$  and the statistic  $\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(v=n-1)}$

Using the information about the distribution of the sample variance, the confidence interval for the true variance is developed and is as follows:

##### **Confidence Intervals for the Variance of a Normal Population**

Suppose that there is a random sample of  $n$  observations from a normally distributed population with variance  $\sigma^2$ . If the observed sample variance is  $s^2$ , then a  $100(1-\alpha)\%$  **confidence interval for the population variance** is given by

$$\frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2} \quad (8.7)$$

where  $\chi_{n-1, \alpha/2}^2$  is the number for which  $P(\chi_{n-1}^2 > \chi_{n-1, \alpha/2}^2) = \frac{\alpha}{2}$  and  $\chi_{n-1, 1-\alpha/2}^2$  is the number for which  $P(\chi_{n-1}^2 < \chi_{n-1, 1-\alpha/2}^2) = \frac{\alpha}{2}$  and the random variable  $\chi_{n-1}^2$  follows a chi-square distribution with  $(n-1)$  degrees of freedom.

##### **Notation**

A random variable having the chi-square distribution with  $v = n-1$  degrees of freedom will be denoted by  $\chi_v^2$  or simply  $\chi_{n-1}^2$ . Define as  $\chi_{n-1, \alpha}^2$  the number for which  $P(\chi_{n-1}^2 > \chi_{n-1, \alpha}^2) = \alpha$

#### 5. Confidence Intervals for the Difference Between Means of Two Normal Populations

Matched paired data just reduces to the case of one variable. When looking at matched paired data, the differences are looked at, and so we would apply all our one variable procedures on the differences. For example, if we wanted to see if a diet is effective, we would like at  $X$  = the before weight and  $Y$  = weight after being on the diet. The  $(X, Y)$  data points must go together and  $D = X - Y$ . Positive  $D$  values indicate an effective diet. A confidence interval for the mean of the  $D$  population reduces to the one variable problem. Hence the following result is not surprising.

### **Confidence Intervals For Two Means: Matched Pairs**

Suppose that there is a random sample of  $n$  matched pairs of observations from normal distributions with means  $\mu_X$  and  $\mu_Y$ . That is, let  $x_1, x_2, \dots, x_n$  denotes the values of the observations from the population with mean  $\mu_X$ ; and  $y_1, y_2, \dots, y_n$  the matched sampled values from the population with the mean  $\mu_Y$ . Let  $\bar{d}$  and  $s_d$  denote the observed sample mean and standard deviation for the  $n$  differences  $d_i = x_i - y_i$ . If the population distribution of the differences is assumed to be normal, then a  $100(1-\alpha)\%$  **confidence interval for the difference between means** ( $\mu_d = \mu_X - \mu_Y$ ) is given by

$$\bar{d} - t_{n-1, \alpha/2} \frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1, \alpha/2} \frac{s_d}{\sqrt{n}} \quad (8.8)$$

or equivalently,

$$\bar{d} \pm B$$

where the **margin of error**, the sampling error or the bound,  $B$ , is given by

$$B = t_{n-1, \alpha/2} \frac{s_d}{\sqrt{n}} \quad (8.9)$$

and  $t_{n-1, \alpha/2}$  is the number for which  $P(t_{n-1} > t_{n-1, \alpha/2}) = \alpha/2$

The random variable  $t_{n-1}$ , has a Student's  $t$  distribution with  $(n - 1)$  degrees of freedom.

### **Two Means, Independent Samples, Known Population Variances**

This situation is a straightforward generalization of the one population problem. To compare the means, we look for a confidence interval for  $(\mu_X - \mu_Y)$  and since the variances are known, it reduces to a  $z$  confidence interval.

### **Confidence Intervals for Difference Between Means: Independent Samples (Normal Distributions and Known Population Variances)**

Suppose that there are two **independent random samples** of  $n_X$  and  $n_Y$  observations from normally distributed populations with means  $\mu_X$  and  $\mu_Y$  and variances  $\sigma_X^2$  and  $\sigma_Y^2$ . If the observed sample means are  $\bar{X}$  and  $\bar{Y}$ , then a **100(1- $\alpha$ )%** confidence interval for  $(\mu_X - \mu_Y)$  is given by

$$(\bar{X} - \bar{Y}) - Z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} < \mu_X - \mu_Y < (\bar{X} - \bar{Y}) + Z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \quad (8.10)$$

or equivalently,

$$(\bar{X} - \bar{Y}) \pm B$$

where the **margin of error** is given by  $B = Z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$  (8.11)

### **Two Means, Independent Samples, Unknown Population Variances Assumed to be Equal**

Most often when we want to compare the means of two populations, we are in a situation where the population variances are unknown. Based on what we did in the one

population case, you might expect us to just replace the population variances in formula 8.10 with the sample variances and use a t with the appropriate degrees of freedom instead of z. Actually, that is what is done when we believe the unknown variances are unequal.

If we believe the population variances are **unknown but equal** (this situation comes up the most in practice), then we want estimates of each of the population variances that reflect this relationship. Hence, we replace each of the population variances in formula 8.10 with an estimate of the variance based on pooling together all the information we get from both our samples. This estimate is called **the pooled variance estimate** and is simply a **weighted average of each of the sample variances**.

A confidence interval for  $(\mu_x - \mu_y)$  for this case, then is as follows:

***Confidence Intervals Two Means: Unknown Population Variances that are Assumed to be Equal***

**Suppose that there are two independent random samples with  $n_x$  and  $n_y$  observations from  $n$  normally distributed populations with means  $\mu_x$  and  $\mu_y$  and a common, but unknown population variance. If the observed sample means are  $\bar{X}$  and  $\bar{Y}$ , and the observed sample variances are  $s_x^2$  and  $s_y^2$ , then a  $100(1-\alpha)\%$  confidence interval for  $(\mu_x - \mu_y)$  is given by**

$$(\bar{X} - \bar{Y}) - t_{n_x+n_y-2, \alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} < \mu_x - \mu_y < (\bar{X} - \bar{Y}) + t_{n_x+n_y-2, \alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} \quad (8.12)$$

or equivalently,  $(\bar{X} - \bar{Y}) \pm B$

where the margin of error is  $B = t_{n_x+n_y-2, \alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$  (8.13)

and the pooled sample variance,  $s_p^2$ , is given by  $s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$  (8.14)

$t_{n_x+n_y-2, \alpha/2}$  is the number for which  $P(t_{n_x+n_y-2} > t_{n_x+n_y-2, \alpha/2}) = \frac{\alpha}{2}$  The random variable, T, is **approximately a Student's t distribution with  $n_x + n_y - 2$  degrees of freedom**

and T is given by, 
$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

**6. Confidence Intervals for the Difference Between Two Population Proportions (Large Samples)**

To find a confidence interval for the difference of 2 proportions is a straightforward generalization of the one population case and its applications are just as similar.

***Confidence Intervals for the Difference Between Population Proportions (Large Samples)***

Let  $p_X$ , denote the observed proportion of successes in a random sample of  $n_X$  observations from a population with proportion  $\pi_X$  successes, and let  $p_Y$ , denote the proportion of successes observed in an independent random sample from a population with proportion  $\pi_Y$  successes. Then, if the sample sizes are large (generally at least forty observations in each sample), a 100(1- $\alpha$ )% **confidence interval for the difference between population proportions, ( $\pi_X - \pi_Y$ )**

is given by  $(p_X - p_Y) \pm B$  where the **margin of error** is  $B = Z_{\alpha/2} \sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}$

**7. Sample Size Determination**

**Sample Size Determination for Mean of a Normally Distribution Population with Known Population Variances**

When finding confidence intervals for means, we said that in order to get a better confidence interval (one with a smaller width), one way is to increase the sample size. Hence, we can decide, before we actually gather our data, how large a sample we would want in order to keep our margin of error  $B$  small.

If we solve for  $n$  in the formula for the margin of error  $B$ , we can determine the sample size. When the variance is known and we want to calculate the sample size, the solution is as follow

***Sample Size for the Mean of a Normally Distributed Population with Known Population Variance***

Suppose that a random sample from a normally distributed population with known variance  $\sigma^2$  is selected. Then a 100(1- $\alpha$ )% confidence interval for the population mean extends a distance  $B$  (sometimes called the bound, sampling error, or the margin of error) on each side of the sample

mean, if the **sample size,  $n$** , is  $n = \frac{Z^2_{\alpha/2} \sigma^2}{B^2}$

Example) Let's return to the problem of the Northwest Hospital Administrators whose salaries were normally distribution with known standard deviation of 12 (\$000). If we wish to find a 95% confidence interval for the true mean salary that has a margin of error of \$3,000 (hence a width of \$6,000), how large a sample must we take?

$n = \frac{Z^2_{\alpha/2} \sigma^2}{B^2} = \frac{(1.96)^2 (12)^2}{3^2} = 61.4656$  **A sample size is always a whole number, the rule is to**

**always round up. Hence the answer is  $n=62$ .**

**If the variance were unknown we would put in an estimate for the variance in the formula. It is always best to overestimate the variance, because you can only do better by sampling more.**

Often we read newspaper articles or hear on the radio about surveys or political polls that are taken which say that the margin of error of the survey or political poll is  $\pm 4$  percentage

points. This is actually equivalent to the situation of estimating a proportion by a sample proportion and making the margin of error .04. If we go to the formula for the confidence

interval for  $\pi$  and solve for  $n$ , it turns out that  $n = \frac{z_{\alpha/2}^2 \pi(1-\pi)}{B^2}$ . However, we do not know the

true proportion  $\pi$ , but we do know  $0 < \pi < 1$ . So, the most the sample size can be is  $\frac{0.25(Z_{\alpha/2})^2}{B^2}$

since the maximum value of  $\pi(1-\pi)$  is .25. This leads to the sample size formula for  $p$ , as follows:

***Sample Size for Population Proportion***

Suppose that a random sample is selected from a population. Then a  $100(1-\alpha)\%$  confidence interval for the population proportion, extending a distance of at most  $B^*$  on each side of the

sample proportion, can be guaranteed if the sample size is  $n = \frac{0.25(Z_{\alpha/2})^2}{B^2}$

---