Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ٢

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

Confidence Interval for Means μ

Given a random sample $X_1, X_2, \ldots, X_n$ from normal distribution $\boldsymbol{N(\mu, \sigma^2)}$, we shall now consider the closeness of $\overline{X}$, the unbiased estimator of **μ** , to the **unknown** mean **μ** . to do this , we use the error structure (distribution) of $\overline{X}$, namely , that $\overline{X}$ is $\boldsymbol{N(\mu, \frac{\sigma^2}{n})}$ to construct what is called a **confidence interval** for the unknown parameter **μ** when the variance $\boldsymbol{\sigma^2}$ is **known** . For the probability $1-\alpha$ we can find a number $\boldsymbol{z_{\alpha/2}}$ from table V in Appendix E such that

$$P\left(-\,z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \; z_{\alpha/2}\right) = 1-\alpha$$

For example , if $1-\alpha = 0.95$, then $z_{\alpha/2} = \; z_{0.05} = 1.645$ .Now recalling that $\boldsymbol{\sigma > 0}$, we see that the following inequalities are equivalent :

$$-\,\boldsymbol{z_{\alpha/2}} \le \frac{\overline{\boldsymbol{X}}-\boldsymbol{\mu}}{\boldsymbol{\sigma}/\sqrt{\boldsymbol{n}}} \le \; \boldsymbol{z_{\alpha/2}}$$

$$-\,\boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right) \le \overline{\boldsymbol{X}} - \boldsymbol{\mu} \le \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)$$

$$-\overline{\boldsymbol{X}} - \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right) \le -\boldsymbol{\mu} \le -\overline{\boldsymbol{X}} + \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)$$

$$\overline{\boldsymbol{X}} + \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right) \ge \; \boldsymbol{\mu} \; \ge \overline{\boldsymbol{X}} - \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)$$

Thus , since the probability of the first of these is $1-\alpha$, the probability of the last must also be $1-\alpha$, because the latter is true if an only if the former is true . that is , we have

$$\boldsymbol{P}\left(\overline{\boldsymbol{X}} - \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right) \le \boldsymbol{\mu} \le \overline{\boldsymbol{X}} + \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)\right) = \boldsymbol{1-\alpha}$$

So the probability that the random interval

$$\left[\,\overline{\boldsymbol{X}} - \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right) , \; \overline{\boldsymbol{X}} + \; \boldsymbol{z_{\alpha/2}}\left(\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)\right]$$

Includes the **unknown** mean **μ** is $1-\alpha$

Once the sample is observed and the sample mean computed to equal $\overline{X}$, the interval [$\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$, $\overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$] becomes **known** . since the probability that the random interval covers **μ** before the sample is drawn is equal to $1-\alpha$, we now call the computed interval , $\overline{X} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ (for brevity) , a $100(1-\alpha)\%$ ***confidence interval for*** *the* ***unknown*** *mean* ***μ*** . Foe example $\overline{X} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ is a 95% **confidence interval** for **μ** . The number $100(1-\alpha)\%$, or equivalently , $1-\alpha$ is called the ***confidence coefficient***.

Example : let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer. Assume that the distribution of X is $N(\mu, 1296)$. if a random sample of $n = 27$ bulbs is tested until they burn out , yielding a sample mean of $\overline{X} = 1478$ hours , then a 95% **confidence interval for μ** is

$$\left[\overline{X} - z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right), \overline{X} + z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

$$= \left[1478 - 1.96\left(\frac{36}{\sqrt{27}}\right),\ 1478 + 1.96\left(\frac{36}{\sqrt{27}}\right)\right]$$

$$= [\,1478 - 13.58\,,\ 1478 + 13.58\,]$$

$$= [\,1464.42\,,\ 1491.58\,]$$

The next example will help to give a better intuitive feeling for the interpretation of a **confidence interval** .

Example: Let $\overline{X}$ be the observed sample mean of five observations of a random sample from the normal distribution $N(\mu, 16)$. A 90% **confidence interval** for the **unknown** mean **μ** is

$$\left[\overline{X} - 1.645\sqrt{\frac{16}{5}}\ , \overline{X} + 1.645\sqrt{\frac{16}{5}}\right]$$

Example : Let $X_1, X_2, \dots, X_{32}$ be a random sample of size 32 from a normal distribution $\mathbf{N(\mu, \sigma).^2}$If $\overline{X} = 19.07$ and $S^2 = 10.60$, then what is the 95 % **confidence interval** for the
population mean **μ** ?

Solution : since $n = 32 \geq 30$, $z_{\alpha/2} = 1.96$ for 95% **confidence interval** ($\alpha/2 = 0.025$)

Hence , the **confidence interval for μ** at 95% **confidence level** is

$$19.07 - 1.96\sqrt{\frac{10.60}{32}} < \mu < 19.07 + 1.96\sqrt{\frac{10.60}{32}}$$

*Thus 95% **confidence interval** : 17.94* $< \mu < 20.20$

If the random sample arises from a normal distribution , we use the fact that

$$\boldsymbol{T = \frac{\overline{X}-\mu}{S/\sqrt{n}}}$$

has a t- distribution with $r = n - 1$ ***degrees of freedom*** , *where* S^2 *is the usual unbiased estimator of* σ^2. *Select* $t_{\alpha/2(n-1)}$ *so that* $\boldsymbol{P\left[T \geq t_{\alpha/2(n-1)}\right] = \alpha/2}$

$$\boldsymbol{1 - \alpha = P\left[-t_{\alpha/2(n-1)} \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2(n-1)}\right]}$$

$$\mathbf{= P\left[-t_{\alpha/2(n-1)} \frac{S}{\sqrt{n}} \leq \overline{X} - \mu \leq t_{\alpha/2(n-1)} \frac{S}{\sqrt{n}}\right]}$$

$$\mathbf{= P\left[-\overline{X} - t_{\alpha/2(n-1)} \frac{S}{\sqrt{n}} \leq -\mu \leq -\overline{X} + t_{\alpha/2(n-1)} \frac{S}{\sqrt{n}}\right]}$$

$$= \mathbf{P}\left[\overline{\mathbf{X}} - \mathbf{t}_{\alpha/2(n-1)}\,\frac{\mathbf{S}}{\sqrt{\mathbf{n}}} \leq \boldsymbol{\mu} \leq \overline{\mathbf{X}} + \mathbf{t}_{\alpha/2(n-1)}\,\frac{\mathbf{S}}{\sqrt{\mathbf{n}}}\right]$$

Thus , the observations of a random sample provide $\overline{X}$ and S^2, and

$$\left[\overline{\mathbf{X}} - \mathbf{t}_{\alpha/2(n-1)}\,\frac{\mathbf{S}}{\sqrt{\mathbf{n}}}\,,\overline{\mathbf{X}} + \mathbf{t}_{\alpha/2(n-1)}\,\frac{\mathbf{S}}{\sqrt{\mathbf{n}}}\right]$$

is a $100(1-\alpha)\%$ **confidence interval for μ**

Example : Let X equal the amount of butterfat in pounds produced by a typical cow during a 305-day milk production period between her first and second calves . Assume that the distribution of X is $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. To estimate μ , a farmer measured the butterfat production for $n = 20$ cows and obtained the following data

481 537 513 583 453 510 570 500 457 555

618 327 350 643 499 421 505 637 599 392

For these data , $\overline{X} = 507.50$ and $S = 89.75$. Thus , a point estimate of μ is $\overline{X} = 507.50$, since $t_{0.05}(19) = 1.729$. a 90% **confidence interval for μ** is

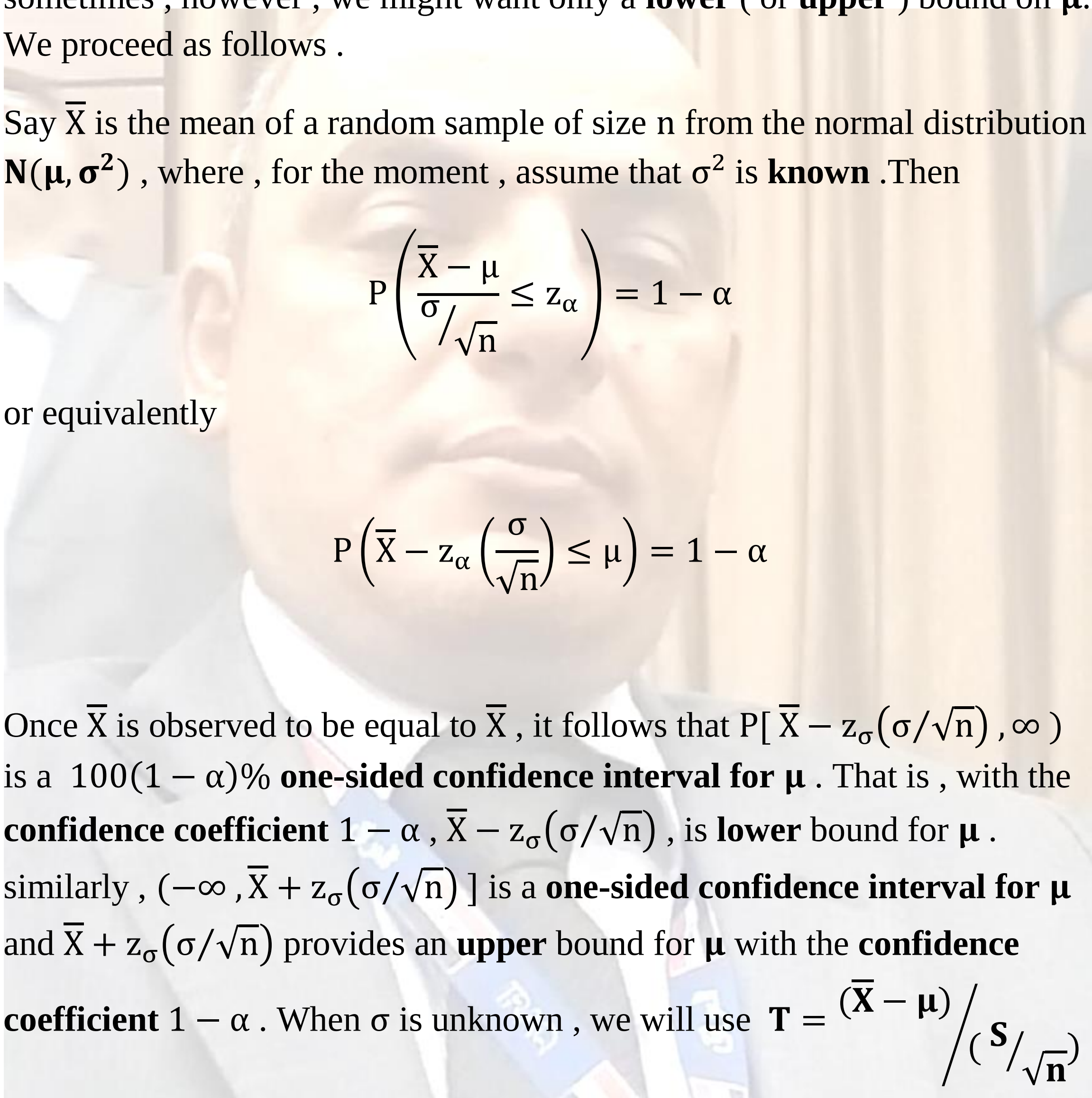
$507.50 \pm 1.729(\frac{89.75}{\sqrt{20}})$ or 507.50 ± 34.70

Or equivalently $[472.80\,, 542.20]$

If we are not able to assume that the underlining distribution is normal , but μ and α are both **unknown** , approximate **confidence interval for μ** can still be constructed with the formula

$$\mathbf{T} = \frac{\overline{\mathbf{X}} - \boldsymbol{\mu}}{\mathbf{S}/\sqrt{\mathbf{n}}}$$

Which now only has an approximate t- distribution . Generally , this approximation is quite good (i.e., it is robust) for many not normal distribution ; in particular , it works will if the underlining distribution is symmetric , unimodal , and of the continuous type . However , if the distribution is highly skewed , there is great danger in using that approximation . in such a situation , it would be safer to use certain **nonparametric methods** for finding a **confidence interval** for the **median** of the distribution , one of which is given in this lecture. There is one other aspect of **confidence interval** that should be mentioned . so far , we have created only that are called **two- sided confidence interval for the mean $\boldsymbol{\mu}$** . sometimes , however , we might want only a **lower** (or **upper**) bound on $\boldsymbol{\mu}$. We proceed as follows .



Say $\overline{\mathrm{X}}$ is the mean of a random sample of size n from the normal distribution $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^{\mathbf{2}})$, where , for the moment , assume that σ^2 is **known** .Then

$$\mathrm{P}\left(\frac{\overline{\mathrm{X}}-\mu}{\sigma/\sqrt{\mathrm{n}}} \leq z_{\alpha}\right) = 1-\alpha$$

or equivalently

$$\mathrm{P}\left(\overline{\mathrm{X}} - z_{\alpha}\left(\frac{\sigma}{\sqrt{\mathrm{n}}}\right) \leq \mu\right) = 1-\alpha$$

Once $\overline{\mathrm{X}}$ is observed to be equal to $\overline{\mathrm{X}}$, it follows that $\mathrm{P}[\ \overline{\mathrm{X}} - z_{\sigma}\left(\sigma/\sqrt{\mathrm{n}}\right), \infty\ )$ is a $100(1-\alpha)\%$ **one-sided confidence interval for $\boldsymbol{\mu}$** . That is , with the **confidence coefficient** $1-\alpha$, $\overline{\mathrm{X}} - z_{\sigma}\left(\sigma/\sqrt{\mathrm{n}}\right)$, is **lower** bound for $\boldsymbol{\mu}$. similarly , $(-\infty, \overline{\mathrm{X}} + z_{\sigma}\left(\sigma/\sqrt{\mathrm{n}}\right)\ ]$ is a **one-sided confidence interval for $\boldsymbol{\mu}$** and $\overline{\mathrm{X}} + z_{\sigma}\left(\sigma/\sqrt{\mathrm{n}}\right)$ provides an **upper** bound for $\boldsymbol{\mu}$ with the **confidence coefficient** $1-\alpha$. When σ is unknown , we will use $\mathbf{T} = (\overline{\mathbf{X}}-\boldsymbol{\mu}) \big/ \left(\mathbf{S}/\sqrt{\mathbf{n}}\right)$

to find the corresponding **lower** or **upper** bounds for $\boldsymbol{\mu}$, namely

$$\overline{\mathbf{X}} - \mathbf{t}_{\boldsymbol{\alpha}}(\mathbf{n}-\mathbf{1})(\mathbf{S}/\sqrt{\mathbf{n}}) \quad \textbf{and} \quad \overline{\mathbf{X}} + \mathbf{t}_{\boldsymbol{\alpha}}(\mathbf{n}-\mathbf{1})(\mathbf{S}/\sqrt{\mathbf{n}})$$

CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO MEANS $\mu_x - \mu_y$

Suppose that we are interested in comparing the means of two normal distributions. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be, respectively, two independent random samples of sizes n and m from the two normal distributions $\mathbf{N(\mu_x, \sigma_x{}^2)}$ and $\mathbf{N(\mu_y, \sigma_y{}^2)}$. Suppose, for now, that $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ are **known**. The random samples are independent; thus, the respective sample means $\overline{X}$ and $\overline{Y}$ are also independent and have distributions $\mathbf{N(\mu_x, \sigma_x{}^2)}$ and $\mathbf{N(\mu_y, \sigma_y{}^2)}$**.** Consequently, the distribution of $W = \overline{X} - \overline{Y}$ is $\mathbf{N}\left(\boldsymbol{\mu_x - \mu_y}, \frac{\boldsymbol{\sigma_x}^2}{\mathbf{n}} + \frac{\boldsymbol{\sigma_y}^2}{\mathbf{m}}\right)$ and

$$\mathbf{P}\left(-\,\mathbf{z}_{\alpha/2} \le \frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu_x}-\boldsymbol{\mu_y})}{\sqrt{\frac{\boldsymbol{\sigma_x}^2}{\mathbf{n}}+\frac{\boldsymbol{\sigma_y}^2}{\mathbf{m}}}} \le \mathbf{z}_{\alpha/2}\right) = \mathbf{1} - \boldsymbol{\alpha}$$

which can be rewritten as

$$\mathbf{P}\left[(\overline{\mathbf{X}} - \overline{\mathbf{Y}}) - \mathbf{z}_{\alpha/2}\boldsymbol{\sigma}_{\mathbf{W}} \le \boldsymbol{\mu_x} - \boldsymbol{\mu_y} \le (\overline{\mathbf{X}} - \overline{\mathbf{Y}}) + \mathbf{z}_{\alpha/2}\boldsymbol{\sigma}_{\mathbf{W}}\right] = \mathbf{1} - \boldsymbol{\alpha}$$

where $\boldsymbol{\sigma}_{\mathbf{W}} = \sqrt{\frac{\boldsymbol{\sigma_x}^2}{\mathbf{n}} + \frac{\boldsymbol{\sigma_y}^2}{\mathbf{m}}}$ is the **standard deviation** of $\overline{\mathbf{X}} - \overline{\mathbf{Y}}$. Once the experiments have been performed and the means $\overline{X}$ and $\overline{Y}$ computed , the interval

$$\left[\, \overline{\mathbf{X}} - \overline{\mathbf{Y}} - \mathbf{z}_{\alpha/2}\boldsymbol{\sigma}_{\mathbf{W}}\,,\ \overline{\mathbf{X}} - \overline{\mathbf{Y}} + \mathbf{z}_{\alpha/2}\boldsymbol{\sigma}_{\mathbf{W}} \,\right]$$

or, equivalently, $(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sigma_W$ provides a $100(1-a)\%$ **confidence interval for $\boldsymbol{\mu_x - \mu_y}$** . Note that this interval is centered at the **point estimate** $\overline{X} - \overline{Y}$ of $\boldsymbol{\mu_x - \mu_y}$ and is completed by subtracting and adding the product of $z\alpha/_2$ and the **standard deviation** of the **point estimator**.

Example : In the preceding discussion, let $n = 15$, $m = 8$ $\overline{X} = 70.1$, $\overline{Y} = 75.3$, $\sigma_x{}^2 = 60$, $\sigma_y{}^2 = 40$ and $1 - \alpha = 0.90$. Thus , $1 - \alpha/_2 = 0.95 = \varphi(1.645)$. Hence ,

$$1.6450\sigma_W = 1.645\sqrt{\frac{60}{15} + \frac{40}{8}} = 4.935$$

and, since $\overline{X} - \overline{Y} = -5.2$, it follows that

$$[-5.2 - 4.935, -5.2 + 4.935] = [-10.135, -0.265]$$

is a 90% **confidence interval for** $\boldsymbol{\mu_x - \mu_y}$. Because the **confidence interval** does not include zero, we suspect that $\boldsymbol{\mu_y}$ is greater than $\boldsymbol{\mu_x}$.

If the sample sizes are large and $\boldsymbol{\sigma_x}$ and $\boldsymbol{\sigma_y}$ are unknown, we can replace $\boldsymbol{\sigma_x{}^2}$ and $\boldsymbol{\sigma_y{}^2}$ with $S_x{}^2$and $S_y{}^2$, where $\mathbf{S_x{}^2}$and $\mathbf{S_y{}^2}$ are the values of the respective unbiased estimates of the variances. This means that

$$\overline{\mathbf{X}} - \overline{\mathbf{Y}} \pm \mathbf{z}\alpha/_2\sqrt{\frac{\mathbf{S_x{}^2}}{\mathbf{n}} + \frac{\mathbf{S_y{}^2}}{\mathbf{m}}}$$

serves as an approximate $100(1 - \alpha)\%$ **confidence interval for** $\boldsymbol{\mu_x - \mu_y}$ **.**

Now consider the problem of constructing **confidence intervals** for the difference of the means of two normal distributions when the variances are **unknown** but the sample sizes are **small**. Let $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_m$ be two independent random samples from the distributions $\mathbf{N}(\boldsymbol{\mu_X}, \boldsymbol{\sigma_x{}^2})$ and $\mathbf{N}(\boldsymbol{\mu_y}, \boldsymbol{\sigma_y{}^2})$, respectively. If the sample sizes are not **large** (say, considerably smaller than 30), this problem can be a difficult one. However, even in these cases, if we can assume common, but **unknown**, **variances** (say, $\boldsymbol{\sigma_x{}^2 = \sigma_y{}^2 = \sigma^2}$) , there is a way out of our difficulty.

We know that

$$\mathbf{Z} = \frac{(\overline{\mathbf{X}} - \overline{\mathbf{Y}}) - (\boldsymbol{\mu}_\mathbf{x} - \boldsymbol{\mu}_\mathbf{y})}{\sqrt{\boldsymbol{\sigma}^2/\mathbf{n} + \boldsymbol{\sigma}^2/\mathbf{m}}}$$

is $\mathbf{N(0, 1)}$**.** Moreover, since the random samples are independent,

$$\mathbf{U} = \frac{(\mathbf{n-1})\mathbf{S_x}^2}{\boldsymbol{\sigma}^2} + \frac{(\mathbf{n-1})\mathbf{S_y}^2}{\boldsymbol{\sigma}^2}$$

is the sum of two independent **chi-square** random variables; thus, the distribution of U is $(n + m - 2)$. In addition, the independence of the sample means and sample variances implies that Z and U are independent. According to the definition of a T random variable,

$$\mathbf{T} = \frac{\mathbf{Z}}{\sqrt{\mathbf{U}/(\mathbf{n+m+2})}}$$

has a distribution with $n + m - 2$ **degrees of freedom**. That is,

$$\mathrm{T} = \frac{\dfrac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_\mathbf{x}-\boldsymbol{\mu}_\mathbf{y})}{\sqrt{\frac{\boldsymbol{\sigma}_\mathbf{x}^2}{\mathbf{n}}+\frac{\boldsymbol{\sigma}_\mathbf{y}^2}{\mathbf{m}}}}}{\sqrt{\left[\frac{(\mathbf{n-1})\mathbf{S_x}^2}{\boldsymbol{\sigma}^2}+\frac{(\mathbf{n-1})\mathbf{S_y}^2}{\boldsymbol{\sigma}^2}\right]\Big/(\mathbf{n+m-2})}}$$

$$= \frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_\mathbf{x}-\boldsymbol{\mu}_\mathbf{y})}{\sqrt{\left[\frac{(\mathbf{n-1})\mathbf{S_x}^2+(\mathbf{n-1})\mathbf{S_y}^2}{\mathbf{n+m-2}}\right]\left[\frac{\mathbf{1}}{\mathbf{n}}+\frac{\mathbf{1}}{\mathbf{m}}\right]}}$$

degrees of freedom. Thus, with

has a t distribution with to $r = n + m - 2$ **degrees of freedom**. Thus, with $\mathbf{t_0} = \mathbf{t}_{\alpha/2}(\mathbf{n+m-2})$, we have

$$\mathbf{P}(-\mathbf{t_0} \le \mathbf{T} \le \mathbf{t_0}) = \mathbf{1} - \boldsymbol{\alpha}$$

solving the inequalities for $\boldsymbol{\mu_x - \mu_y}$, yields

$$\mathbf{P}\left(\bar{\mathbf{X}} - \bar{\mathbf{Y}} - \mathbf{t_0 S_p}\sqrt{\frac{\mathbf{1}}{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{m}}} \leq \boldsymbol{\mu_x - \mu_y} \leq \bar{\mathbf{X}} - \bar{\mathbf{Y}} + \mathbf{t_0 S_p}\sqrt{\frac{\mathbf{1}}{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{m}}}\right)$$

where the pooled estimator of the common **standard deviation** is

$$\mathbf{S_p} = \sqrt{\frac{(\mathbf{n-1})\mathbf{S_x}^2 + (\mathbf{n-1})\mathbf{S_y}^2}{\mathbf{n+m-2}}}$$

If $\bar{X}, \bar{Y},$ and S_p are the observed values of $\bar{X}$, $\bar{Y}$. and S_p , then

$$\left[\bar{\mathbf{X}} - \bar{\mathbf{Y}} - \mathbf{t_0 S_p}\sqrt{\frac{\mathbf{1}}{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{m}}}\ ,\ \bar{\mathbf{X}} - \bar{\mathbf{Y}} + \mathbf{t_0 S_p}\sqrt{\frac{\mathbf{1}}{\mathbf{n}} + \frac{\mathbf{1}}{\mathbf{m}}}\right]$$

is a $100(1-\alpha)\%$ **confidence interval for** $\boldsymbol{\mu_x - \mu_y}$.

Example : Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $\mathbf{N}(\boldsymbol{\mu_X}, \boldsymbol{\sigma_x}^2)$ and $\mathbf{N}(\boldsymbol{\mu_y}, \boldsymbol{\sigma_y}^2)$, respectively, where $\boldsymbol{\sigma^2}$ is **unknown**. If a random sample of $n = 9$ students from large high schools yielded $\bar{X} = 81.31$, $\sigma_x^2 = 60.76$, and a random sample of $m = 15$ students from small high schools yielded $\bar{Y} = 78.61$, $\sigma_y^2 = 48.24$, then the endpoints for a 95% **confidence interval for** $\boldsymbol{\mu_x - \mu_y}$ are given by

$$\mathbf{81.31 - 78.61 \pm 2.074}\sqrt{\frac{\mathbf{8(60.76) + 14(48.24)}}{\mathbf{22}}}\sqrt{\frac{\mathbf{1}}{\mathbf{9}} + \frac{\mathbf{1}}{\mathbf{15}}}$$

because $t_{0.025}(22) = 2.074$. The 95% **confidence interval** is $[-3.65, 9.05]$.

REMARKS The assumption of equal variances, namely, $\sigma_x^2 = \sigma_y^2$. can be modified somewhat so that we are still able to find a confidence interval for $\mu_x - \mu_y$. That is, if we know the ratio $\boldsymbol{\sigma_x^2/\sigma_y^2}$ of the variances, we can still make this type of statistical inference by using a random variable with a t distribution. However, if we do not know the ratio of the variances and yet suspect that the unknown $\sigma_x{}^2$ and $\sigma_y{}^2$ differ by a great deal, what do we do? It is safest to return to

$$\frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_{\mathbf{x}}-\boldsymbol{\mu}_{\mathbf{y}})}{\sqrt{\frac{\sigma_x{}^2}{n}+\frac{\sigma_y{}^2}{m}}}$$

for the inference about $\boldsymbol{\mu}_{\mathbf{x}}-\boldsymbol{\mu}_{\mathbf{y}}$ but replacing $\sigma_x{}^2$ and $\sigma_y{}^2$ by their respective estimators $S_x{}^2$ and $S_y{}^2$. That is, consider

$$\mathbf{W}=\frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_{\mathbf{x}}-\boldsymbol{\mu}_{\mathbf{y}})}{\sqrt{\frac{S_x{}^2}{n}+\frac{S_y{}^2}{m}}}$$

What is the distribution of W? As before, we note that if n and m are large enough and the underlying distributions are close to normal (or at least not badly skewed), then W has an approximate normal distribution and a **confidence interval for $\boldsymbol{\mu}_{\mathbf{x}}-\boldsymbol{\mu}_{\mathbf{y}}$** can be found by considering

$$\mathbf{P}\left(-\mathbf{z}\alpha_{/2} \leq \mathbf{W} \leq \mathbf{z}\alpha_{/2}\right) \approx \mathbf{1}-\boldsymbol{\alpha}$$

However, for smaller n and m, Welch has proposed a Student's t distribution as the approximating one for W. Welch's proposal was later modified by Aspin. (See A. A. Aspin, "Tables for Use in Comparisons Whose Accuracy Involves Two Variances, Separately Estimated," Biometrika , 36 (1949), pp. 290-296, with an appendix by B. L. Welch in which he makes the suggestion used here.] The approximating Student's t distribution has r degrees of freedom, where

$$\frac{\mathbf{1}}{\mathbf{r}}=\frac{\mathbf{c}^{\mathbf{2}}}{\mathbf{n}-\mathbf{1}}+\frac{(\mathbf{1}-\mathbf{c})^{\mathbf{2}}}{\mathbf{m}-\mathbf{1}} \quad \text{and} \quad \mathbf{c}=\frac{\frac{S_x{}^2}{n}}{\frac{S_x{}^2}{n}+\frac{S_y{}^2}{m}}$$

An equivalent formula for r is

$$\mathbf{r} = \frac{\left(\frac{S_x{}^2}{n}+\frac{S_y{}^2}{m}\right)^2}{\frac{1}{\mathbf{n-1}}\left(\frac{S_x{}^2}{n}\right)^2+\frac{1}{\mathbf{m-1}}\left(\frac{S_y{}^2}{m}\right)^2}$$

In particular, the assignment of r by this rule provides protection in the case in which the smaller sample size is associated with the larger variance by greatly reducing the number of **degrees of freedom** from the usual $n + m - 2$. Of course, this reduction increases the value of $t_{\alpha/2}$. If r is not an integer, then use the greatest integer in r ; that is, use [r] as the number of degrees of freedom associated with the approximating Student's

t-distribution. An approximate $100(1-\alpha)\%$ **confidence interval for** $\boldsymbol{\mu_x - \mu_y}$ is given by

$$\overline{\mathbf{X}} - \overline{\mathbf{Y}} \pm \mathbf{t}_{\alpha/2}(\mathbf{r})\sqrt{\frac{S_x{}^2}{n}+\frac{S_y{}^2}{m}}$$

It is interesting to consider the two-sample T in more detail. It is

$$\mathbf{T} = \frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_\mathbf{x}-\boldsymbol{\mu}_\mathbf{y})}{\sqrt{\frac{(\mathbf{n-1})\mathbf{S_x}^2+(\mathbf{m-1})\mathbf{S_y}^2}{\mathbf{n+m-2}}\left(\frac{1}{\mathbf{n}}+\frac{1}{\mathbf{m}}\right)}}$$

$$= \frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_\mathbf{x}-\boldsymbol{\mu}_\mathbf{y})}{\sqrt{\left[\frac{(\mathbf{n-1})\mathbf{S_x}^2}{\mathbf{nm}}+\frac{(\mathbf{m-1})\mathbf{S_y}^2}{\mathbf{nm}}\right]\left[\frac{\mathbf{n+m}}{\mathbf{n+m-2}}\right]}}$$

Now, since $(n-1)/n \approx 1$, $(m-1)/m \approx 1$, and $(n+m)/(1+m-2) \approx 1$, we have

$$\mathbf{T} \approx \frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu}_\mathbf{x}-\boldsymbol{\mu}_\mathbf{y})}{\sqrt{\frac{\mathbf{S_x}^2}{\mathbf{n}}+\frac{\mathbf{S_y}^2}{\mathbf{m}}}}$$

We note that, in this form, each variance is divided by the wrong sample size! That is, if the sample sizes are large or the variances **known,** we would like

$$\sqrt{\frac{S_x{}^2}{n}+\frac{S_y{}^2}{m}} \qquad \textbf{or} \qquad \sqrt{\frac{\sigma_x{}^2}{n}+\frac{\sigma_y{}^2}{m}}$$

in the denominator; so T seems to change the sample sizes. Thus, using this T is particularly bad when the sample sizes and the variances are unequal; hence, caution must be taken in using that T to construct a **confidence interval for $\boldsymbol{\mu_x - \mu_y}$** . That is, if $n < m$ and $\sigma_x{}^2 < \sigma_y{}^2$, then T does not have a t- distribution which is close to that of a Student t-distribution with

$n + m - 2$ degrees of freedom: Instead, its spread is much less than the Student t's as the term $\sigma_y{}^2/n$ in the denominator is much larger than it should be. By contrast, if $m < n$ and $\sigma_x{}^2 < \sigma_y{}^2$, then $S_x{}^2/m + S_y{}^2/n$ is generally smaller than it should be and the distribution of T is spread out more than that of the Student t.

There is a way out of this difficulty, however: When the underlying distributions are close to normal, but the sample sizes and the variances are seemingly much different, we suggest the use of

$$\mathbf{W} = \frac{(\overline{\mathbf{X}}-\overline{\mathbf{Y}})-(\boldsymbol{\mu_x}-\boldsymbol{\mu_y})}{\sqrt{\frac{S_x{}^2}{n}+\frac{S_y{}^2}{m}}}$$

where Welch proved that W has an approximate t distribution with [r] degrees of freedom, with the number of degrees of freedoms.