What we know so far:

We have seen how to set confidence intervals for the mean, or expected value, of a normal probability distribution, both when the variance is known (using the standard normal, or Z, table), and when it is not (using "Student's t" table). The two intervals are

$$\overline{X}_n \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad \text{and} \qquad \overline{X}_n \pm t_{\alpha/2}^{n-1} \frac{S}{\sqrt{n}}$$

We have also learned that, thanks to the Central Limit Theorem and the Law of Large Numbers,

$$\overline{X}_n \pm Z_{\alpha/2} \frac{S_n}{\sqrt{n}}$$

is an approximate confidence interval for the expected value, E[X], when the sample size (n) is large, even if the observations are coming from a distribution other than the normal.

Most commonly, the second interval (with the coefficient from the t table, ($t_{\alpha/2}^{n-1}$) is used only when the distribution of the individual observations is believed to be "nearly" normal.

Otherwise the third interval is used and is called a "*large sample confidence interval"* or an *"approximate confidence interval"* to emphasize that the coverage probability is only approximate.

Thus,

$\overline{X}_n \pm Z_{\alpha/2}\frac{S_n}{\sqrt{n}}$ is an "approximate (1-α)100% CI".

This all works because

$$\frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{S_n} \overset{approx}{\sim} N(0,1)$$

for `large' n.

Confidence Intervals for the difference between two means

In order to learn about how oral contraceptive use affects blood pressure, we find some women who use oral contraceptives and some who don't, observe their systolic blood pressures, and see what we see.

We conceptualize the two groups as different populations, from which we draw a sample:

Group 1: (OC users)
$X_1,\ldots,X_n$ are i.i.d. with E[X]= μ_X and Var[X]= σ_X^2

A sample of n=8 yields $\bar{x} = 132.86$ mm and $s_x = 15.34$ mm

Group 2: (non-users)
$Y_1,\ldots,Y_m$ are i.i.d. with E[Y]= μ_Y and Var[Y]= σ_Y^2

A sample of m=21 yields $\bar{y} = 127.44$ mm and $s_y = 18.23$ mm

☞ To analyze these observations we might use a probability model that says that blood pressures are normally distributed and we might not. We'll see later why this becomes important.

The quantity we are trying to estimate is E[X]-E[Y]= μ_X - μ_Y and our estimator is simply $\overline{X}-\overline{Y}$.

Because $\overline{X}$ and $\overline{Y}$ are random variables, so is $\overline{X}-\overline{Y}$. Hence we can standardize it like so:

$$Z=\frac{\overline{X}-\overline{Y}-E\left[\overline{X}-\overline{Y}\right]}{\sqrt{Var\left[\overline{X}-\overline{Y}\right]}}$$

Now we know that:

1) $E\left[\overline{X}-\overline{Y}\right]=E\left[X\right]-E\left[Y\right]=\mu_X-\mu_Y$

and

2)

$$Var\left[\overline{X}-\overline{Y}\right]=Var\left[\overline{X}\right]+Var\left[\overline{Y}\right]=\frac{Var\left[X\right]}{n}+\frac{Var\left[Y\right]}{m}=\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}$$

And because the CLT and LLN work on *averages of random variables* we have that

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \overset{approx}{\sim} N(0,1) \quad \text{as n gets large}*$$

Finally because P(-$Z_{\alpha/2}$ < Z < $Z_{\alpha/2}$)=1-α,

An approximate* (1-α)100% confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{X} - \overline{Y} \pm Z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

(*Which is exact if the underlying distributions of the X's and Y's are both normal.)

But the variance is unknown.

If we estimate $\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}$ with $\frac{S_X^2}{n}+\frac{S_Y^2}{m}$, we run into a problem because:

$$T=\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}} \quad \sim \quad ???$$

The distribution of T is unknown -- is not normal nor students-t. We have no way of calculating an exact or even approximate confidence interval. (Why?)

Fortunately, we can always use the CLT to save us in large samples because:

$$Z=\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}} \quad \overset{approx}{\sim} \quad N(0,1) \quad \text{as n getslarge}$$

Now, in large samples, an approximate (1-α)100% confidence interval for $\mu_X - \mu_Y$ given by

$$\overline{X} - \overline{Y} \pm Z_{\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

For our case this yields the following 95% CI :

$$132.86 - 127.44 \pm 1.96\sqrt{\frac{15.34^2}{8} + \frac{18.23^2}{21}}$$

or

$$5.42 \pm 13.18 = \left(-7.76, 18.60\right)$$

"At the 95% level the data suggest that the difference between the two populations means is at least -7.76mm but no more than 18.60 mm."

Our observations are evidence that OC use causes an increase in blood pressure. (Our best estimate is that it increases mean blood pressure by 5.42 mm.) But the evidence is not very strong (because $\mu_X - \mu_Y$ = 0 means that there is <u>*no*</u> *increase, and the 95% CI includes this value).*

☞ Is our sample size large enough to invoke the CLT? Here we have n=8 and m=21, so probably not.

So what can we do for small samples?

Unless we are willing to make some additional assumptions then nothing more can be done.

So, if we assume that the underlying distributions of X and Y are *approximately* normal (symmetrical and not too skewed) then when the sample size is small (<u>either</u> n or m or both) then there are several available methods.

The price we pay is that our procedure is no longer `robust'. This is because all of our future calculations will depend on the fact that X's and Y's are normally distributed and if in fact they have some other distribution our calculations will be wrong.

(This is why the large sample interval discussed earlier is used so often, and also why us statisticians bug you doctors about getting a large sample size.)

The general problem is to come up with an estimate of $Var\left[\overline{X}-\overline{Y}\right]=\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}$, call it $\hat{V}$, so that

$$Z=\frac{\overline{X}-\overline{Y}-\left(\mu_X-\mu_Y\right)}{\sqrt{\hat{V}}} \quad \sim \quad Q$$

where the distribution Q is known.

However there is more than one "natural" way to estimate the variance of $\overline{X}-\overline{Y}$. Unfortunately, only one of these ways leads to a tidy solution (another Student's t interval), and it is often inappropriate.

Unknown Variance Method 1

(The Case of Equal Variances)

Assuming that both the X's and Y's are normally distributed, there is a neat, exact solution only for the case when the variances, although unknown, are assumed to be equal ($\sigma_X^2 = \sigma_Y^2 = \sigma^2$).

In this case $Var\left[\overline{X} - \overline{Y}\right] = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)$ is estimated

with $\hat{V} = S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)$ where $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$ is the `pooled' variance estimate.

It is a weighted average $S_p^2 = \left(\frac{n-1}{n+m-2}\right)S_X^2 + \left(\frac{m-1}{n+m-2}\right)S_Y^2$ of the two sample variances, S_X^2 and S_Y^2, with the one that is based on more observations getting more weight.

Now the standardized difference

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$$

has exactly a t-distribution with n+m-2 degrees of freedom.

Thus the (1-α)100% confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{X} - \overline{Y} \pm t_{\alpha/2}^{n+m-2} \sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}$$

Assuming that the variances in the two populations are equal and the underlying distributions are approximately normal.

However this interval is fairly robust to non-normality (That is, it continues to have approximately the correct coverage probability when the distributions are not normal).

In our example of how oral contraceptive use affects blood pressure, the pooled variance estimate is

$$s_p^2 = \frac{7(15.34)^2 + 20(18.23)^2}{8+21-2} = 307.1803$$

so for a confidence coefficient of 0.95 we find from Table A.2, $t_{27} = 2.052$, and the 95% confidence interval for the mean blood pressure difference between OC users and non-users is

$$132.86 - 127.44 \pm 2.052\sqrt{307.1803\left(\frac{1}{8} + \frac{1}{21}\right)}$$

$$\text{or} \quad 5.42 \pm 2.052(7.282),$$

$$\text{or} \quad 5.42 \pm 14.94.$$

Or (-9.52 mm, 20.36 mm).

However the variances are rarely, if ever, equal. So what can we do if we assume that the X’s and Y’s are normally distributed, but the variances are unequal.

Unknown Variance Method 2

(The Case of Unequal Variances)

Assuming that both the X's and Y's are normally distributed, and that ($\sigma_X^2 \neq \sigma_Y^2$), there is no neat solution (it remains unsolved today!)

In this case we estimate $\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}$ with $\frac{S_X^2}{n}+\frac{S_Y^2}{m}$, but run into a problem because:

$$T=\frac{\overline{X}-\overline{Y}-\left(\mu_X-\mu_Y\right)}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}} \quad \sim \quad ???$$

(Remember that we are assuming that the sample sizes are small enough that T would not be approximately normal, by the CLT)

We can use an approximation such as

$$T = \frac{\overline{X} - \overline{Y} - \left(\mu_X - \mu_Y\right)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \quad \overset{approx}{\sim} \quad t^*$$

where t^* is approx a t-dist with

$$DF = \frac{\left(S_X^2 / n + S_Y^2 / m\right)^2}{\frac{\left(S_X^2 / n\right)^2}{n-1} + \frac{\left(S_Y^2 / m\right)^2}{m-1}} \text{ rounded down}$$

called Satterthwaite's correction for df.
(most computer programs do this – but too much of a pain to do by hand)

Alternatively there is a conservative approximation

$$T = \frac{\overline{X} - \overline{Y} - \left(\mu_X - \mu_Y\right)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \quad \overset{approx}{\sim} \quad t_{\min[n-1,m-1]}$$

That is, just use the df for the smallest population.

☞ Why is this conservative?

Each approach suggests the interval

$$\overline{X} - \overline{Y} \pm t^{*}_{\alpha/2}\sqrt{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)}$$

Where the degree of freedom is calculated either as

DF = Min[n-1,m-1] or DF= $\dfrac{\left(S_X^2/n + S_Y^2/m\right)^2}{\dfrac{\left(S_X^2/n\right)^2}{n-1} + \dfrac{\left(S_Y^2/m\right)^2}{m-1}}$

Both intervals are approximately correct under the assumption of X's and Y's normally distributed with unequal variances. But neither is the exact solution.

Take home message:

When we have independent samples from two normal distributions with unknown and unequal variances, we cannot find sensible exact confidence intervals for the difference between the means (in the small sample case).

Fortunately the Central Limit Theorem and the Law of Large Numbers still apply, and they provide the basis for approximate CI's when both sample sizes, n and m, are large. These two results (CLT and LLN) can be used to prove that

$$\overline{X} - \overline{Y} \pm Z_{\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

is an *approximate* (1-α)100% for μ_X - μ_Y (As we saw earlier). *That is, the probability that this random interval will include* μ_X - μ_Y *approaches 0.95 as the sample sizes grow.*

If the *X*'s and Y's are normal and the two variances are equal, then the Student's t CI,

$$\overline{X} - \overline{Y} \pm t_{\alpha/2}^{n+m-2}\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)},$$

is exact for all sample sizes, large or small.

If the variances are very unequal, the coverage probability of this interval might not be even approximately correct, even when the *X*'s are normally distributed and the samples are large.

The coverage probability of this interval can be seriously wrong if the two variances σ_X^2 and σ_Y^2 are not equal, because the pooled variance estimate, $S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)$ estimates

$$E\left(S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)\right) = E\left(\frac{(n-1)\,S_X^2+(m-1)\,S_Y^2}{n+m-2}\right)\left(\frac{1}{n}+\frac{1}{m}\right)$$

$$= \frac{(n-1)\,\sigma_X^2+(m-1)\,\sigma_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)$$

not the correct quantity $Var\left[\overline{X}-\overline{Y}\right]=\left(\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}\right)$

In fact the two are the same only when the variances are equal, i.e., when $\sigma_X^2 = \sigma_Y^2$.

Interestingly enough, the main source of the problem with the Student's t confidence interval that is caused by unequal variances disappears when the two sample sizes are equal (n=m)!

In that special case, the pooled variance is estimating the right quantity after all, because the two variance estimates are identical:

$$S_p^2\left(\frac{1}{n}+\frac{1}{n}\right)=\frac{(n-1)\,S_X^2+(n-1)\;S_Y^2}{n+n-2}\left(\frac{1}{n}+\frac{1}{n}\right)$$

$$=\frac{(n-1)(\;S_X^2+\;S_Y^2)}{2(n-1)}\left(\frac{2}{n}\right)$$

$$=\frac{S_X^2}{n}+\frac{S_Y^2}{n}.$$

The distribution is still not exactly Student's t, so the coverage probability won't be exactly the value shown in the t-table. But because the variance estimate is estimating the right thing, the possibility of serious discrepancy between the table value and the actual coverage probability of the interval is avoided when the two sample sizes are roughly equal.

Summary

Interval	When
$\overline{X} - \overline{Y} \pm Z_{\alpha/2}\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$	both n and m are large
$\overline{X} - \overline{Y} \pm t_{\alpha/2}^{n+m-2}\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}$	For n,m small and X’s & Y’s normal; variances equal or n=m
$\overline{X} - \overline{Y} \pm t_{\alpha/2}^{\min[n-1,m-1]}\sqrt{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)}$	For n,m small and X’s & Y’s normal; variances unequal; (can also use Satterthwaite’s df)