34. Normal Mean CIs (variance unknown)

CI's for Normal Mean
Clothing Example

CI's for Difference of Two Normal Means (var's equal)
Anti-University of Georgia Example

CI's for Difference of Two Normal Means (var's $\neq$)
Numerical Example

CI's for Difference of Paired Normal Means
Parking Example

Confidence Intervals for Normal Mean

Now we'll look at the more-realistic case in which the variance of the underlying normal RV's is *unknown*.

This takes a little more work, but has many more applications.

Set-up: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$.

Facts:

(1) $$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \text{Nor}(0,1)$$

(2) $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \sim \frac{\sigma^2 \chi^2(n-1)}{n-1}$$

(3) $\bar{X}$ and S^2 are *indep* (tough to show)

Using Facts (1)–(3), we have, by the definition of the t distribution,

$$\frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{S^2}{\sigma^2}}} \sim \frac{\text{Nor}(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1).$$

In other words,

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t(n-1).$$

Note that this expression doesn't contain the unknown σ^2. It's been replaced by S^2, which we can calculate.

Now, by the **same** manipulations as in the known-var case, we can get a two-sided $100(1-\alpha)\%$ CI for μ:

$$\mu \ \in \ \bar{X} \pm t_{\alpha/2,n-1}\sqrt{S^2/n}.$$

Remark: This process is called "standardizing and Studentizing."

$100(1-\alpha)\%$ lower CI: $\mu \geq \bar{X} - t_{\alpha,n-1}\sqrt{S^2/n}$.

$100(1-\alpha)\%$ upper CI: $\mu \leq \bar{X} + t_{\alpha,n-1}\sqrt{S^2/n}$.

Remark: Here we use t distribution quantiles (instead of normal quantiles as in the known-var case from the previous module). The t quantile tends to be larger than the corresponding Nor(0,1) quantile, so these unknown-var CI's tend to be a bit longer than the known-var CI's. The longer CI's are the result of the fact that we lack precise info about the variance.

Example (H&M): Here are 20 residual flame times (in sec) of treated specimens of children's nightwear.

9.85	9.93	9.75	9.77	9.67
9.87	9.67	9.94	9.85	9.75
9.83	9.92	9.74	9.99	9.88
9.95	9.95	9.93	9.92	9.89

Let's get a 95% CI for the mean residual flame time.

After a little algebra, we get

$$\bar{X} = 9.8475 \text{ and } S = 0.0954.$$

Further, $t_{\alpha/2,n-1} = t_{.025,19} = 2.093$.

Then the half-length of the CI is

$$H = t_{\alpha/2,n-1}\sqrt{S^2/n} = \frac{(2.093)(0.0954)}{\sqrt{20}} = 0.0446.$$

Thus, the CI is $\mu \in \bar{X} \pm H$, or $9.8029 \le \mu \le 9.8921$.

CI's for the Difference of Two Normal Means (var's unknown and equal)

Compare the means from two competing normal populations.

Suppose we have samples of sizes n_x and n_y.

$$\begin{aligned} X_1, X_2, \ldots, X_{n_x} &\overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{(population 1)} \\ Y_1, Y_2, \ldots, Y_{n_y} &\overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2) \quad \text{(population 2).} \end{aligned}$$

We assume that the means μ_x and μ_y are *unknown*, and the variances σ_x^2 and σ_y^2 are also *unknown*.

Big Assumption: Suppose for now that $\sigma_x^2 = \sigma_y^2 = \sigma^2$, i.e., assume that the two variances are *equal*.

Also assume that the X_i's are *indep* of the Y_i's.

Let's find a CI for the difference in means, $\mu_x - \mu_y$.

To get things going, calculate sample means.

$$\bar{X} \equiv \frac{1}{n_x}\sum_{i=1}^{n_x} X_i \quad \text{and} \quad \bar{Y} \equiv \frac{1}{n_y}\sum_{i=1}^{n_y} Y_i.$$

Now, sample variances.

$$S_x^2 = \frac{\sum_{i=1}^{n_x}(X_i - \bar{X})^2}{n_x - 1}$$

$$S_y^2 = \frac{\sum_{i=1}^{n_y}(Y_i - \bar{Y})^2}{n_y - 1}.$$

Both S_x^2 and S_y^2 are estimators for σ^2 (the common variance). A better estimator is the **pooled** estimator for σ^2, which uses the info from both S_x^2 and S_y^2.

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}.$$

Theorem:

$$S_p^2 \sim \frac{\sigma^2 \chi^2(n_x + n_y - 2)}{n_x + n_y - 2}.$$

Proof: Not here — don't worry about it.

After some of the usual algebra, it turns out that

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t(n_x + n_y - 2).$$

So, after more of the usual algebra, we get a two-sided $100(1 - \alpha)\%$ CI for $\mu_x - \mu_y$:

$$\mu_x - \mu_y \in \bar{X} - \bar{Y} \pm t_{\alpha/2, n_x + n_y - 2} S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}},$$

where $t_{\alpha/2, n_x + n_y - 2}$ is the appropriate t distribution quantile.

One-Sided Lower CI:

$$\mu_x - \mu_y \ \geq \ \bar{X} - \bar{Y} - t_{\alpha,n_x+n_y-2} S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

One-Sided Upper CI:

$$\mu_x - \mu_y \ \leq \ \bar{X} - \bar{Y} + t_{\alpha,n_x+n_y-2} S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

Example: IQ's of students at Georgia Tech and the "Univ." of Georgia.

Georgia Tech students: $X_1, \ldots, X_{25} \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma^2)$.

Univ. of Georgia students: $Y_1, \ldots, Y_{36} \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma^2)$.

Note: We assume common σ^2.

Suppose it turns out that

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i = \frac{1}{25}\sum_{i=1}^{25} X_i = 120$$

$$\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i = \frac{1}{36}\sum_{i=1}^{36} Y_i = 80$$

$$S_x^2 = \frac{\sum_{i=1}^{n_x}(X_i - \bar{X})^2}{n_x - 1} = 100$$

$$S_y^2 = \frac{\sum_{i=1}^{n_y}(Y_i - \bar{Y})^2}{n_y - 1} = 95$$

The two sample var's are pretty close, so we'll go ahead and feel good about our common σ^2 assumption.

Use the pooled var estimator,

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

$$= \frac{(24)(100) + (35)(95)}{59} = 97.03.$$

Thus, a two-sided 95% CI for $\mu_x - \mu_y$ is

$$\begin{aligned} \mu_x - \mu_y &\in \bar{X} - \bar{Y} \pm t_{\alpha/2, n_x + n_y - 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 120 - 80 \pm 2.00\sqrt{97.03}\sqrt{0.0678} \\ &= 40 \pm 5.13. \end{aligned}$$

So the 95% CI is $34.87 \le \mu_x - \mu_y \le 45.13$.

So it's pretty obvious that GT students are smarter than UGA students! No surprise.

CI's for the Difference of Two Normal Means (var's unknown and unequal)

We'll again compare the means from two competing normal populations, but now they might have *unequal* variances.

Again suppose we have samples of sizes n_x and n_y.

$$\begin{aligned} X_1, X_2, \ldots, X_{n_x} &\overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{(population 1)} \\ Y_1, Y_2, \ldots, Y_{n_y} &\overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2) \quad \text{(population 2).} \end{aligned}$$

We assume that the means μ_x and μ_y are *unknown*, and the variances σ_x^2 and σ_y^2 are also *unknown* and possibly *unequal*.

Also assume that the X_i's are *indep* of the Y_i's.

Find a CI for the difference in means, $\mu_x - \mu_y$.

As before, start by calculating sample means and sample variances, $\bar{X}$, $\bar{Y}$, S_x^2, S_y^2.

Since the var's are possibly unequal, we can't use the pooled estimator S_p^2. Instead, use an *approximation*.

$$t^\star \equiv \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}} \approx t(\nu),$$

where the approximate degrees of freedom is given by

$$\nu \equiv \frac{\left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right)^2}{\frac{(S_x^2/n_x)^2}{n_x+1} + \frac{(S_y^2/n_y)^2}{n_y+1}} - 2.$$

Proof: Uses “moment matching” — don't worry.

After the usual algebra, we obtain an approx two-sided $100(1-\alpha)\%$ CI for $\mu_x - \mu_y$:

$$\mu_x - \mu_y \ \in \ \bar{X} - \bar{Y} \pm t_{\alpha/2,\nu} \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}.$$

This is the confidence interval for the difference in means that has the widest application in practice.

Example: Two normal populations. Let's get a 95% CI for $\mu_x - \mu_y$.

Suppose it turns out that

$$n_x = 25 \quad \bar{X} = 100 \quad S_x^2 = 400$$

$$n_y = 16 \quad \bar{Y} = 80 \quad S_y^2 = 100$$

You can tell from S_x^2 and S_y^2 that there's no way that the two var's are equal. So we'll have to use the approximation method.

The approx degrees of freedom is

$$\begin{aligned}
\nu &= \frac{\left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}\right)^2}{\frac{(S_x^2/n_x)^2}{n_x+1} + \frac{(S_y^2/n_y)^2}{n_y+1}} - 2 \\
&= \frac{\left(\frac{400}{25} + \frac{100}{16}\right)^2}{\frac{(400/25)^2}{26} + \frac{(100/16)^2}{17}} - 2 \\
&= 38.77 \approx 38,
\end{aligned}$$

where we round the d.f. *down* to be *conservative* (i.e., slightly longer CI's).

Since the confidence coefficient is 0.95, we have

$$t_{\alpha/2,\nu} = t_{.025,38} = 2.02.$$

This gives us the following CI for $\mu_x - \mu_y$.

$$\mu_x - \mu_y \in \bar{X} - \bar{Y} \pm t_{\alpha/2,\nu} \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}$$

$$= 20 \pm 2.02\sqrt{22.25} = 20 \pm 9.53.$$

CI's for Difference of Paired Normal Means

Again consider two competing normal pop'ns. Suppose we collect obs'ns from the two pop'ns in *pairs*.

The RV's within *different* pairs are *indep*. The two obs'ns within the *same* pair may *not* be indep.

Example: Think sets of twins. One twin takes a new drug, the other takes a placebo.

In symbols,

$$\text{indep}\begin{cases} \text{Pair } 1: & (X_1, Y_1) \\ \text{Pair } 2: & (X_2, Y_2) \\ \quad\vdots & \quad\vdots \\ \text{Pair } n: & \underbrace{(X_n, Y_n)}_{\text{not indep}} \end{cases}$$

Pair i is indep of pair j, i.e., the pair (X_i, Y_i) is indep of the pair (X_j, Y_j).

But within pair i, it may be that X_i and Y_i are not indep.

Idea: By setting up such experiments, we hope to be able to capture the difference between the two normal pop'ns more precisely, since we're using the pairs to eliminate extraneous noise.

Here's the set-up:

$$\begin{aligned} X_1, X_2, \ldots, X_n &\overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{(population 1)} \\ Y_1, Y_2, \ldots, Y_n &\overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2) \quad \text{(population 2).} \end{aligned}$$

Note that the sample size n is common.

We assume that the means μ_x and μ_y are *unknown*, and the variances σ_x^2 and σ_y^2 are also *unknown* and possibly *unequal*.

Also assume that pair i is indep of pair j (between pairs).

But the X_i may not be indep of Y_i (within a pair).

Find a CI for the difference in means, $\mu_x - \mu_y$.

Define the pair-wise differences,

$$D_i \equiv X_i - Y_i, \quad i = 1, 2, \ldots, n.$$

Note that

$$D_1, D_2, \ldots, D_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_d, \sigma_d^2),$$

where $\mu_d \equiv \mu_x - \mu_y$ and

$$\sigma_d^2 \equiv \sigma_x^2 + \sigma_y^2 - 2\text{Cov}(X_i, Y_i).$$

Now the problem sort of reduces to the plain old Nor(μ, σ^2) case with unknown μ and σ^2. So let's calculate the sample mean and variance as before.

$$\bar{D} \equiv \sum_{i=1}^{n} D_i/n \sim \text{Nor}(\mu_d, \sigma_d^2/n)$$

$$S_d^2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2 \sim \frac{\sigma_d^2 \chi^2(n-1)}{n-1}.$$

Just like before, get

$$\frac{\bar{D} - \mu_d}{\sqrt{S_d^2/n}} \sim t(n-1).$$

After the usual algebra, we obtain an approx two-sided $100(1-\alpha)\%$ CI for $\mu_d = \mu_x - \mu_y$:

$$\mu_d \ \in \ \bar{D} \pm t_{\alpha/2,n-1}\sqrt{S_d^2/n}.$$

One-sided lower: $\mu_d \geq \bar{D} - t_{\alpha,n-1}\sqrt{S_d^2/n}$.

One-sided upper: $\mu_d \leq \bar{D} + t_{\alpha,n-1}\sqrt{S_d^2/n}$.

Example: Times for people to parallel park two cars.

Person	Park Honda	Park Cadillac	Difference
1	10	20	-10
2	25	40	-15
3	5	5	0
4	20	35	-15
5	15	20	-5

Clearly, the people are indep, but the times for the same individual to park the two cars may not be indep.

Let's assume that all times are normal. We want a 90% two-sided CI for $\mu_d = \mu_h - \mu_c$.

We see that $n = 5$, $\bar{D} = 9$, $S_d^2 = 42.5$.

Thus, the 90% two-sided CI is

$$\begin{aligned} \mu_d &\in \bar{D} \pm t_{.05,4} \sqrt{S_d^2/n} \\ &= -9 \pm 2.13 \sqrt{42.5/5} = -9 \pm 6.21. \end{aligned}$$

So why didn't we just use the "usual" CI for the difference in two means? Namely,

$$\mu_x - \mu_y \ \in \ \bar{X} - \bar{Y} \pm t_{\alpha/2,\nu} \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}.$$

Main reason: The above CI requires that the X's must be indep of the Y's. (Recall that the paired-t method allows X_i and Y_i to be dependent.)

So what might happen if I went and managed to make the X's and Y's indep?

Good thing: The approx d.f. ν from the "usual" method would probably be larger than the d.f. $n-1$ from the paired-t method. This would make the CI smaller.

Bad thing: You'd introduce much more noise into the system by using the "usual" method. This could make the CI much larger.

Example: Back to the car example.

A guy parks Honda X_i	Same guy parks Caddy	Different guy parks Caddy Y_i
10	20	30
25	40	15
5	5	40
20	35	10
15	20	25

Just concentrate on the X_i and Y_i columns. (The middle column is from the last example for comparison purposes from the last example.)

Now all of the X_i's are indep of all of the Y_i's.

$$\bar{X} = 15, \quad \bar{Y} = 24, \quad S_x^2 = 62.5, \quad S_y^2 = 142.5.$$

Then we have

$$\begin{aligned}
\nu &= \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{n}\right)^2}{\frac{(S_x^2/n)^2}{n+1} + \frac{(S_y^2/n)^2}{n+1}} - 2 \\
&= \frac{6\left(62.5 + 142.5\right)^2}{(62.5)^2 + (142.5)^2} - 2 \\
&= 8.4 \approx 8.
\end{aligned}$$

This gives us the following 90% CI for $\mu_x - \mu_y$.

$$\mu_x - \mu_y \;\in\; \bar{X} - \bar{Y} \pm t_{.05,8} \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{n}}$$

$$= \; -9 \pm 1.86\sqrt{41} \; = \; -9 \pm 11.91.$$

This CI is *wider* than the paired-t version, even though we have more d.f. here.

Moral: Use paired-t when you can.