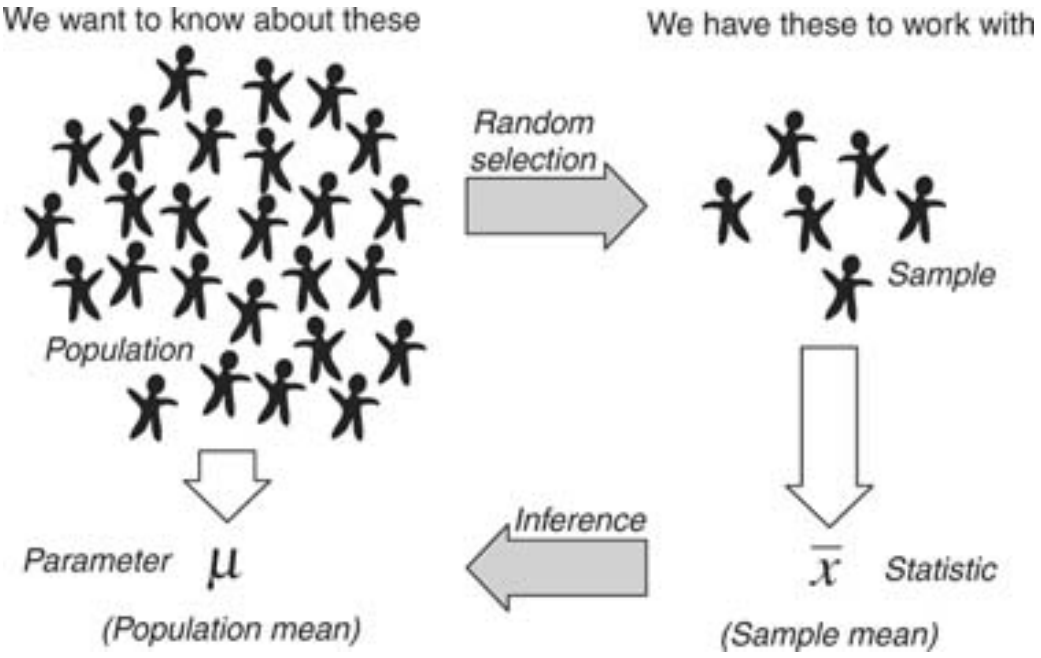
Chapter 6 Class Notes –Confidence Intervals



Consider a population and a RV Y, where the population average is μ and SD is σ; we take a random sample, measure the variable and calculate the sample statistics $\bar{y}$ and s. One goal here is to estimate μ and to comment on the precision of that estimate; to do so, we'll make a confidence statement by finding a CI (confidence interval).

To illustrate (p.185, ex.6.3.3), thymus weights (in mg) of n = 5 chick embryos at 14 days (29.6, 21.5, 28.0, 34.6, 44.9) yields $\bar{y}$ = 31.72 and s = 8.73. Then, μ is the (population) mean thymus weight of <u>all</u> chick embryos at 14 days and σ is the (population) SD of thymus weights of <u>all</u> chick embryos at 14 days.

→ $\bar{y}$ = 31.72 mg is an estimator of μ
→ s = 8.73 mg is an estimator of σ

But what is the reliability of $\bar{y}$ as an estimator of μ?

The answer is the *theoretical* SEM (standard error of the mean), which we saw = $\frac{\sigma}{\sqrt{n}}$ – it is the theoretical magnitude of sampling error. But, like μ it too is unknown since σ is unknown.

So, instead we use the SEM (standard error of the mean),

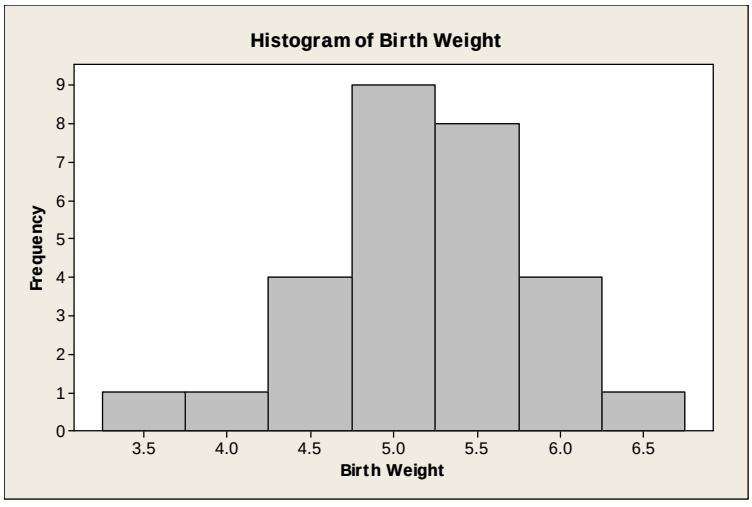
$$SE_{\overline{Y}} = SE = \frac{s}{\sqrt{n}}$$

For the above data, we get SE = $\frac{8.73}{\sqrt{5}}$ = 3.904 mg.

Note the distinction between the SD and the SE (see pp.172-6):

- **the SD describes the dispersion of the data**
- **the SE describes the uncertainty in the mean of the data**

Example 6.2.3. Lamb Birthweights (p.173): n = 28 Rambouillet lamb weights are given and the histogram follows:



The sample statistics are: $\overline{y}$ = 5.168, s = 0.654, SE = 0.124 (all in kg). As pointed out on p.173, s = 0.654 describes the variability from

lamb to lamb in terms of weight whereas SE = 0.124 indicates how much we expect $\bar{y}$ to vary from μ.

Students are asked to understand the discussion of SD versus SE (related to graphics) in Example 6.2.4 on pp.174-6.

Section 6.3 on pp.177-187 addresses setting confidence intervals for μ, and it introduces a new distribution and table (Table 4 on p.618 and in the back cover of the textbook). The authors use an analogy of an invisible man walking his visible dog (see p.177); the leash is spring loaded so that

- **the dog ($\bar{y}$) is within one SE of the man (μ) 68% of the time**
- **the dog ($\bar{y}$) is within two SEs of the man (μ) 95% of the time**

Obviously, we can only see the dog ($\bar{y}$), and we wish to estimate the location of the man (μ). The concept used for CI's is that if the model is correct (i.e., no bias, etc.) then

$$\text{(position of the dog)} \pm 2 \times \text{SE}$$

is an approximately 95% CI for the position of the man.

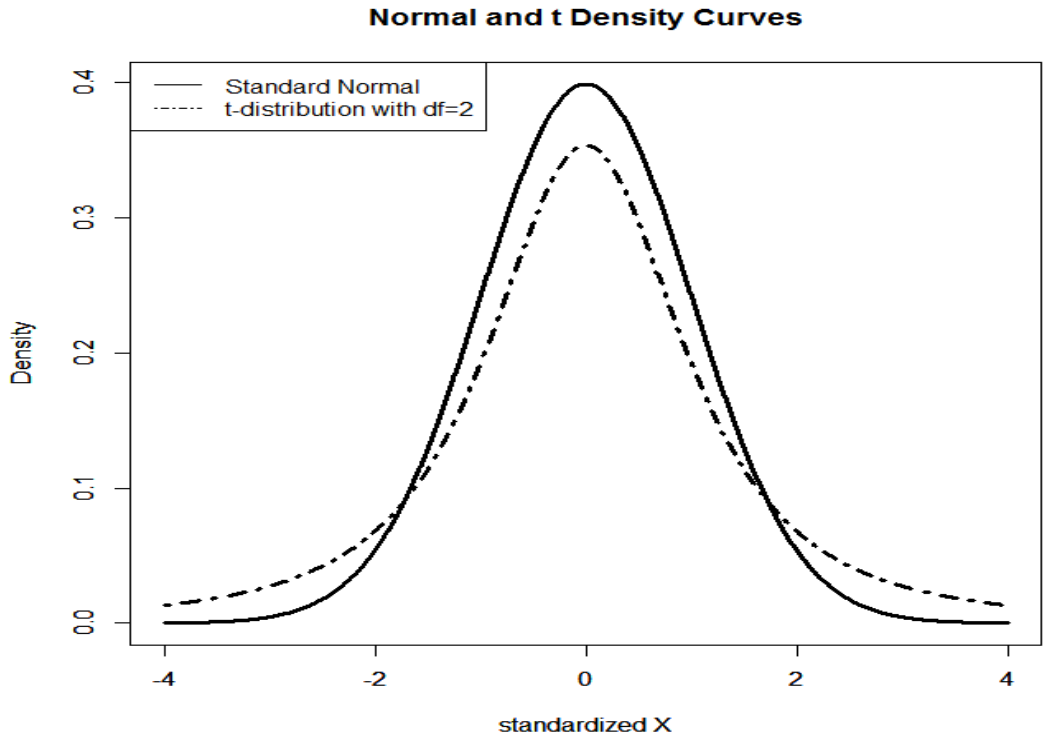
More specifically: for Z ~ N(0,1), then Pr{-1.96 < Z < 1.96} = 0.95. So, if Y ~ N(μ,σ) or if n is large (in which case we can use the CLT), then

$$0.95 = \Pr\left\{-1.96 < \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} < 1.96\right\}$$

$$= \Pr\left\{\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right\}$$

It follows that the interval $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}} = (\bar{y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}})$ will contain the true population mean (μ) 95% of the time.

The problem above is that we *don't know* σ. If we use s instead of σ, then we cannot use the Z (Normal) distribution – rather, we use the Student's t-distribution named after W.S. Gosset, who worked at Guinness Brewery in Dublin (and published under the pen name "Student"). The t-distribution, tallied in Table 4, is indexed by a value for df (degrees of freedom). For one-sample problems such as the above chick embryos and Rambouillet lambs, we use df = n - 1, but this will change in other situations. The t-distribution is "heaver-tailed" than the Normal distribution and the t approaches the Normal as df → ∞. Here's the plot:



Let's adopt the notation: $z_{0.025} = 1.96$, so that the subscript denotes the area under the Normal curve to the right of $z_{0.025} = 1.96$. Using Table 4 and df = 4, note that $t_{0.025} = 2.776$. Thus, for the above thymus weight example, the 95% CI for μ is

$$\bar{y} \pm t_{0.025}\frac{s}{\sqrt{n}} = 31.7 \pm 2.776 \times \frac{8.73}{\sqrt{5}} = 31.7 \pm 10.8 = \underline{(20.9, 42.5)}$$

Conclusion: Assuming that the parent population of the chick embryo thymus weights is *Normal*, we're 95% confident that the average thymus weight at 14 days of *all* chicks lies between 20.9mg and 42.5mg.

Had we desired instead a 90% CI for μ we would have found:

$$\bar{y} \pm t_{0.05}\frac{s}{\sqrt{n}} = 31.7 \pm 2.132 \times \frac{8.73}{\sqrt{5}} = 31.7 \pm 8.3 = \underline{(23.4, 40.0)}$$

→ *Notice that the 90% CI is narrower than the 95% CI*.

Note that with all other things equal:

1. **as we increase the confidence level (e.g., going from 95% to 99%), the CI gets wider**
2. **as we increase the sample size, the CI gets narrower.**

Technical Note (pp.182-183): it is the CI which is random, not μ: the 95% probability comes in via *the method*; don't treat μ as if it is a random variable. A simulation study helps here as in the text.

For the Rambouillet lamb birthweight example, we don't need to assume Normality since n = 28 is large (but as always the sample weights must represent a *random sample*). Then, the 95% CI for μ is

$$\bar{y} \pm t_{0.05}\frac{s}{\sqrt{n}} = 5.17 \pm 2.052 \times \frac{0.65}{\sqrt{28}} = 5.17 \pm 0.252 = \underline{(4.92, 5.42)}$$

Conclusion: Provided these data represent a random sample, we're 95% confident that the average birthweight *all* Rambouillet lambs lies between 4.92kg and 5.42kg.

Sample Size Determination (Section 6.4, pp.187-189)

If σ is know (e.g. from a previous estimate), then $\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is a $(1-\alpha)$100% CI for μ. Let's call $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ the associated Margin of Error, denoted 'ME'. Note that the ME is half the length of the CI. Now, suppose that we wish to choose a sample size (n) so that the ME is at most equal to k (some given constant). Then,

$$\text{ME} \le k \rightarrow z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le k \rightarrow n \ge \left(\frac{z_{\alpha/2}\times\sigma}{k}\right)^2$$

For example, for the Thymus illustration with σ = 9kg, suppose we want a 95% margin of error not to exceed 5kg. Then,

$$n \ge \left(\frac{z_{\alpha/2}\times\sigma}{k}\right)^2 = \left(\frac{1.96\times 9}{5}\right)^2 = 3.528^2 = 12.45 \rightarrow \text{choose } \underline{n = 13}.$$

Section 6.5 (*Conditions for the Validity of Estimation Methods* on pp.190-199) highlight some caveats and points related to the above.

- If a given dataset is not the result of a random sample, all our analysis and efforts may be futile. If the sample is only representative of a *sub-group* (such as only biology or science students), then we need to be mindful of this when we state our conclusion(s). Samples of convenience are often potentially misleading and should be avoided.
- For sampling without replacement from a finite population of size N, the SE of $\bar{Y}$ is actually $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$. Here, $\sqrt{\frac{N-n}{N-1}}$ is called the FPC (for finite population correction) factor. When N is very large or $\frac{n}{N} \le 5\%$, the FPC factor can be ignored.

- **Be careful of nested or hierarchical measurements – their analysis requires specialized techniques and those methods discussed in Chapters 6 and 7 do not apply – we require independent measurements here. To illustrate, it is obvious that in a 6-point longitudinal study, the six measurements e.g. at times t = 0, 2, 4, 8, 16, and 24 hours are probably correlated, and so we cannot treat (analyze) them as independent measurements. Another example is the fungus spore germination study discussed on p.191: 129 batches of spores were placed on 129 Petri dishes with one of 43 treatments (3 replicates of each). 10 plugs were taken (cut) from each Petri dish, incubated, and the percentage of germinated spores was assessed. A portion of the data (for TRT1) looks like:**

	Dish A	Dish B	Dish C
	49	66	49
	58	84	60
	48	83	54
	69	69	72
	45	72	57
	43	85	70
	60	59	65
	44	60	68
	44	75	66
	68	68	60
Mean	52.8	72.1	62.1
SD	10.1	9.5	7.4

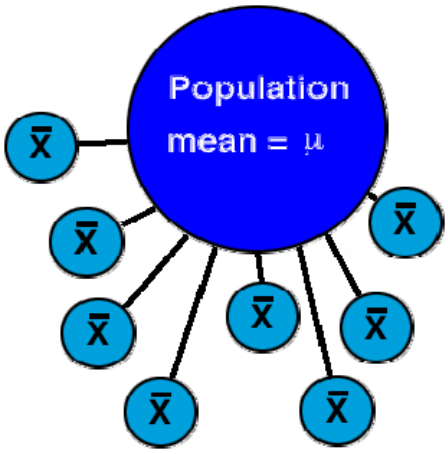
The average of all thirty numbers above is 62.33 and the SD is 11.88. Is it correct to use as the SE for TRT1: $\frac{11.88}{\sqrt{30}}$ = 2.2?

No, of course not. The EU (experimental unit) here is the dish, so we only have three independent measurements – thus the correct SE would be the SD of 52.8, 72.1 and 62.1 divided by $\sqrt{3}$ - this turns out to be 9.65/$\sqrt{3}$ = 5.6.

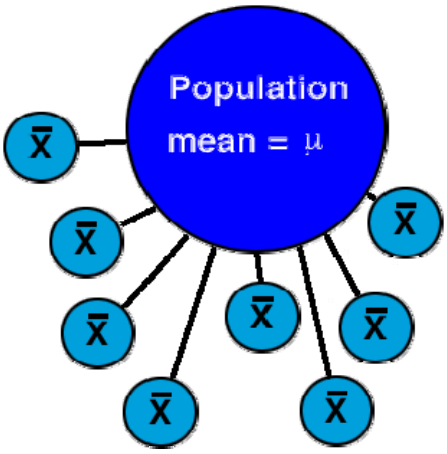
So, when we're reviewing a study, we need to scrutinize the experimental design (if possible): What are potential biases? Are the measurements independent? Is the "largeness of n" criterion met or is it reasonable to assume Normality of the response variable? Examine the NPP and histogram and check for outliers.

In Sections 6.6 and 6.7, we'll compare means from two independently sampled groups using CIs (confidence intervals). There are two paradigms or settings:

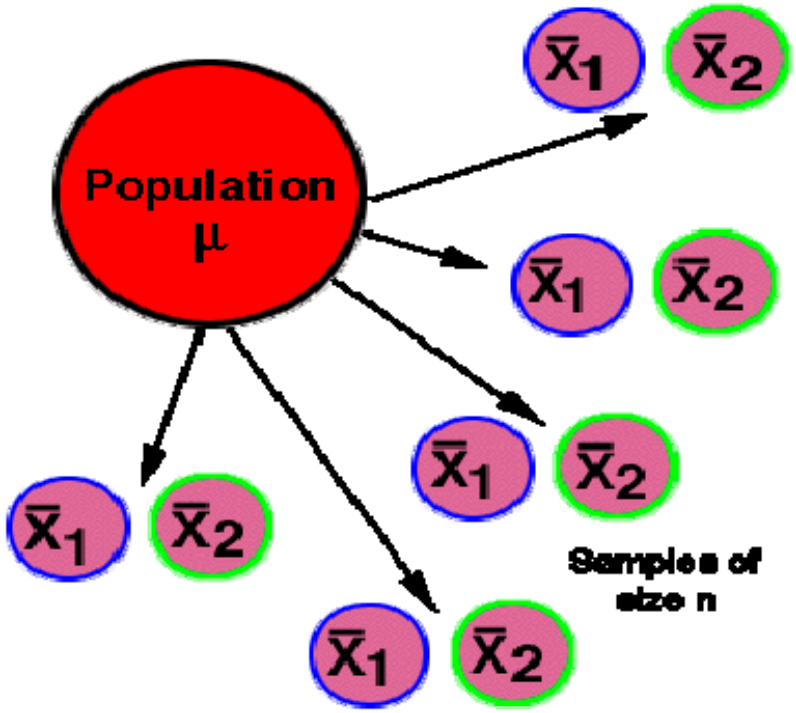
Group I (e.g., males)



Group II (e.g., females)



Randomize some subjects to treatment A and some to treatment B:



We illustrate with these examples:

A. p.205, ex.6.6.8: compare total dry weight (in grams) of the leaves of two varieties of lettuce: the "Salad Bowl" and "Bibbs" varieties.

B. p.258, ex.7.5.6: researchers compare the respiratory responses (liters of air/minute/m^2 body area) of hypnotized ("experimental") and non-hypnotized ("control") male volunteers.

In the second example, note the *active role* of the researcher in randomizing the volunteers to the two groups, whereas for the first example, the partition occurs naturally and no randomization is necessary. The text illustration on p.242 ex.7.4.1 resembles our first example in that the groups (males and females) occur naturally.

Summary Data for the first three illustrations:

A. p.205, ex.6.6.8: Y = total dry weight of leaves

Population I: *all* lettuce plants of the Salad Bowl variety

Population II: *all* lettuce plants of the Bibbs variety

	Salad Bowl variety	Bibbs variety
n	$n_1 = 9$	$n_2 = 6$
$\bar{y}$	$\bar{y}_1 = 3.259$	$\bar{y}_2 = 1.413$
s	$s_1 = 0.400$	$s_2 = 0.220$

Plot of the data shows a shift of centers and some change in SD's

B. p.258, ex.7.5.6: Y = respiratory (ventilation) rate

Population I: *all* male volunteers to who could be hypnotized

Population II: *all* male volunteers to who the non-hypnotized (control) treatment could be applied

	Experimental	Control
n	$n_1 = 8$	$n_2 = 8$
$\bar{y}$	$\bar{y}_1 = 6.169$	$\bar{y}_2 = 5.291$
s	$s_1 = 0.621$	$s_2 = 0.652$

A plot of the data shows same shape and SD, shift in center

C. p.242, ex.7.4.1: Y = hematocrit levels

Population I: *all* male 17-year-olds in U.S.

Population II: *all* female 17-year-olds in U.S.

Data summary and graphs show approximately the same shape and SD's, and a possible shift in center.

In examples A and C, the populations occur naturally so these are observational studies; in illustration B, the groups are created by the researcher and this is an experimental study. Regardless, we use the same statistical methods (given below and in Chapter 7) to analyze these sets of data.

The standard error of $(\bar{y}_1 - \bar{y}_2)$ is addressed on p.201:
Let $SE_1 = \frac{s_1}{\sqrt{n_1}}$ and $SE_2 = \frac{s_2}{\sqrt{n_2}}$ be the respective standard errors, then the SE for the difference of the sample means is

$$SE_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

On p.202, the authors draw an analogy to Pythagoras' Theorem (length of the hypotenuse of a right triangle).

For the Lettuce example, $SE_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{0.220^2}{6} + \frac{0.400^2}{9}} = 0.16076$.

For the Hypnosis example, $SE_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{0.621^2}{8} + \frac{0.652^2}{8}} = 0.31834$.

In what follows, note that when we assume that $\sigma_1^2 \neq \sigma_2^2$, the best estimate of the true SE of $(\bar{y}_1 - \bar{y}_2)$ is indeed $SE_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

But if we can assume that $\sigma_1^2 = \sigma_2^2$ [$= \sigma^2$], then the best estimator of σ^2 is the so-called pooled estimator, which is:

$$s^2_{POOLED} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$$

Then, in calculating the SE, s^2_{POOLED} takes the place of both s_1^2 and s_2^2 in the above formula, so the estimated SE for the difference of the means is:

$$SE_{POOLED} = \sqrt{s^2_{POOLED}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

(Only) In this case, the associated *degrees of freedom* is $n_1 + n_2 - 2$

Verify these calculations: For the Lettuce example, $s^2_{POOLED} = 0.3422^2$, so $SE_{POOLED} = 0.18034$ with df = 13. It is a simple exercise in algebra to show that whenever the sample sizes from the two groups are equal (viz, $n_1 = n_2$), both estimates of the SE of the difference of the means are the same; it therefore follows for the Hypnosis example that $SE_{POOLED} = 0.31834$ (here with df = 14).

In §6.7, we address the important task of setting a CI (confidence interval) for $(\mu_1 - \mu_2)$ in a manner similar to that used in §6.3. Here we use the analogous formula:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{0.025} \times SE_{(\bar{Y}_1 - \bar{Y}_2)}$$

Unfortunately, except in the above "pooled" case, it is not known exactly how many degrees of freedom to associate with this t-statistic and an approximation must be used. This approximation is called Welch's method and is used by most software packages and some graphing calculators. As given in Equation (6.7.1) on p.206, we choose the degrees of freedom (df) as the largest integer satisfying the equation:

$$df \leq \frac{\left(SE_1^2 + SE_2^2\right)^2}{\frac{SE_1^4}{n_1 - 1} + \frac{SE_2^4}{n_2 - 1}} \qquad \text{(Equation 6.7.1)}$$

Again, $SE_1 = \frac{s_1}{\sqrt{n_1}}$ and $SE_2 = \frac{s_2}{\sqrt{n_2}}$. On exams and quizzes, the right-hand side ("RHS") will be provided. For the Lettuce example, RHS = 12.718 so df = 12, and for Hypnosis, RHS = 13.968 so df = 13.

With regard to CIs for $(\mu_1 - \mu_2)$ – if we make no assumptions about equality of variances – we get:

- For the Lettuce example (order is Salad Bowl then Bibbs)

$(3.259 - 1.413) \pm 2.179 \times 0.16076 \rightarrow 1.846 \pm 0.3503 \rightarrow (1.495, 2.196)$

Conclusion: We are 95% confident that the population average total dry weight for the Salad Bowl variety is larger than the population average total dry weight for the Bibbs variety by an amount that might be as small as 1.495g and as much as 2.196g.

- For the Hypnosis example (order is Experimental then Control)

$(6.169 - 5.291) \pm 2.160 \times 0.31834 \rightarrow 0.878 \pm 0.688 \rightarrow (0.190, 1.566)$

Conclusion: We are 95% confident that the population average respiratory rate for the Experimental group is between 0.190 and 1.566 liters of air per minute per m^2 of body area greater than the population average respiratory rate for the Control group.

The key assumptions/requirements here are that both samples are (1) random samples, (2) independently selected (or allocated), and

(3) come from Normal populations (i.e., for the response variable). In order to invoke the CLT (i.e., relax the Normality assumption or requirement) here, we need both $n_1 \geq 25$ and $n_2 \geq 25$. Whenever we do the pooled approach, we need to add the assumption that the population variance are equal. Note that in both the Lettuce and the Hypnosis examples, we need to assume Normality of the respective parent populations due to low sample sizes.

Finally, note that since it is rarely used in practice, we will skip the "conservative t-test" discussed on p.206 – this approach uses as the df the minimum of $(n_1 - 1)$ and $(n_2 - 1)$, and is thus overly conservative.