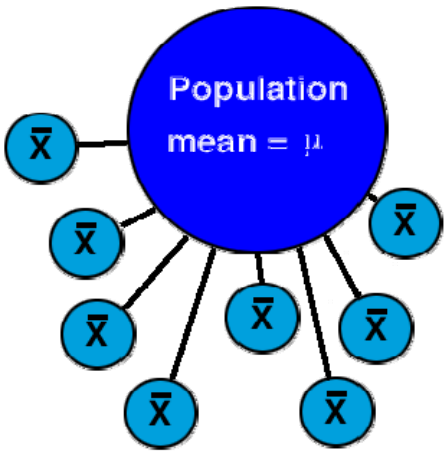
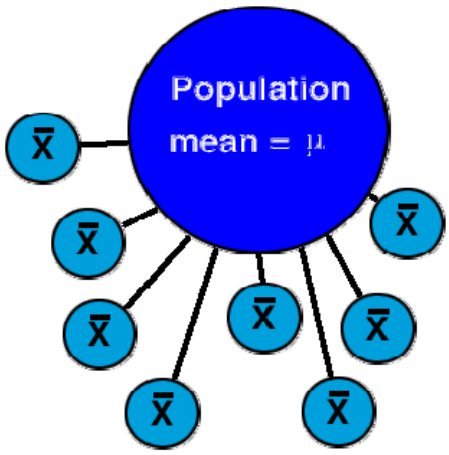
Chapter 7 Class Notes
Comparison of Two Independent Samples

In this chapter, we'll compare means from two independently sampled groups using HTs (hypothesis tests). As noted in Chapter 6, there are two paradigms/settings:

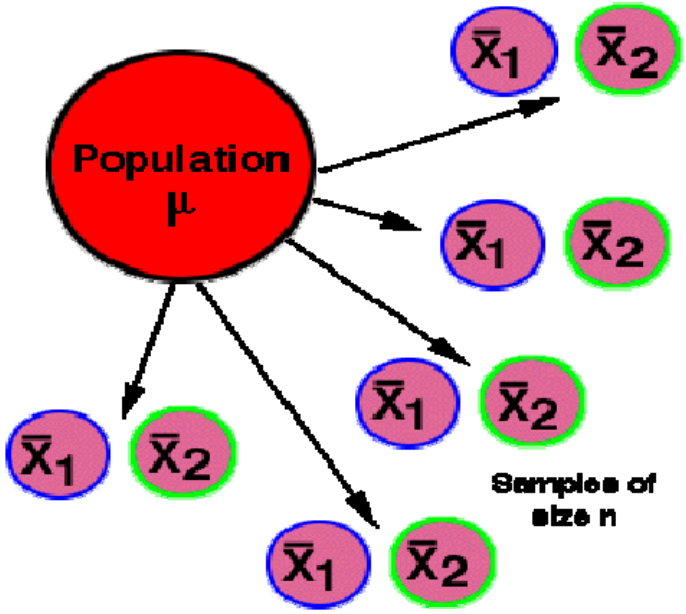
Group I (e.g., males)



Group II (e.g., females)



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



In §7.1 (pp. 218-222), the authors motivate hypothesis testing with a randomization test applied to Example 7.1.1 concerning flexibility of 7 women and two "methods", aerobics class (A) and dance (D). The flexibilities and methods are as follows:

38	45	48	58	59	61	64
A	A	D	A	D	D	A

The average flexibilities are: $\overline{y}_A$= 51.25 (A group) & $\overline{y}_D$= 56.00 (D group), and the observed difference is $\overline{y}_A - \overline{y}_D$= -4.75. *The important question is whether this is enough of a difference to declare that the population means differ*. Assume for the moment that there is no difference in the two methods in terms of flexibility (the "null hypothesis"). In that case, then the 4 "A" and 3 "D" labels could be randomly placed under the 7 flexibility numbers; there are 35 ways to do so, and each time you do, you can calculate a new $\overline{y}_A - \overline{y}_D$ value. When this is done, you get the dotplot on the bottom of p.220, and note that the number of differences at least as extreme as the observed one – i.e., where $|\overline{y}_A - \overline{y}_D| \geq 4.75$ – is 20. Next, since 20/35 = 0.57 is not very rare, it appears that our observed difference of -4.75 is not very extreme.

This randomization test and these data would lead us to conclude we don't have enough evidence to reject the claim that the two methods produce the same (average) flexibility in women. (We need to be careful, however, since, to paraphrase Carl Sagan, lack of evidence of a difference is not equivalent to evidence of no difference.) Note also that (as demonstrated on pp.221-2) this randomization technique become unwieldy for larger problems, so we'll instead use the t-test method discussed below.

The important topic considered in this chapter is that of Hypothesis Tests, introduced and illustrated in Section 7.2 (p.223) – and we use this method for the rest of the text. Hypothesis testing uses a 'proof by contradiction' argument and places the burden of proof on the researcher seeking to show that the status quo is no longer the case.

For the Hypnosis example – where μ_1 is the average respiratory rate of all hypnotized male volunteers and μ_2 is the average respiratory rate of all non-hypnotized male volunteers – we may wonder if the populations mean are the same: $\mu_1 = \mu_2$? If *a priori* (ahead of time) we have no idea of which would be larger than the other, we may wonder if the means differ: $\mu_1 \neq \mu_2$? This is called a *directionless alternative* since no direction is specified; we'll take up directed alternatives later in Section 7.5.

Here, the Null Hypothesis (H_0) and Alternative Hypothesis (H_A) are:

$$H_0: \mu_1 = \mu_2 \text{ or equivalently: } \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 \neq \mu_2 \text{ or equivalently: } \mu_1 - \mu_2 \neq 0$$

In Statistical Test of Hypotheses (denoted HT), the procedure for assessing the compatibility of the data with H_0 ("the Null") uses a test statistic (denoted TS); HT's therefore answer the question of whether the observed difference (between the sample means here) is *real* or just *due to sampling error*. The TS here measures by how many SE's $\bar{y}_1$ and $\bar{y}_2$ differ, and thus is

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{SE_{\bar{y}_1 - \bar{y}_2}}$$

The reason for explicitly writing the "– 0" in the numerator of the above TS reflects the fact that the Null proposes that the difference of the population means μ_1 - μ_2 is equal to zero.

For the Hypnosis example from Chapter 6, our data summary is:

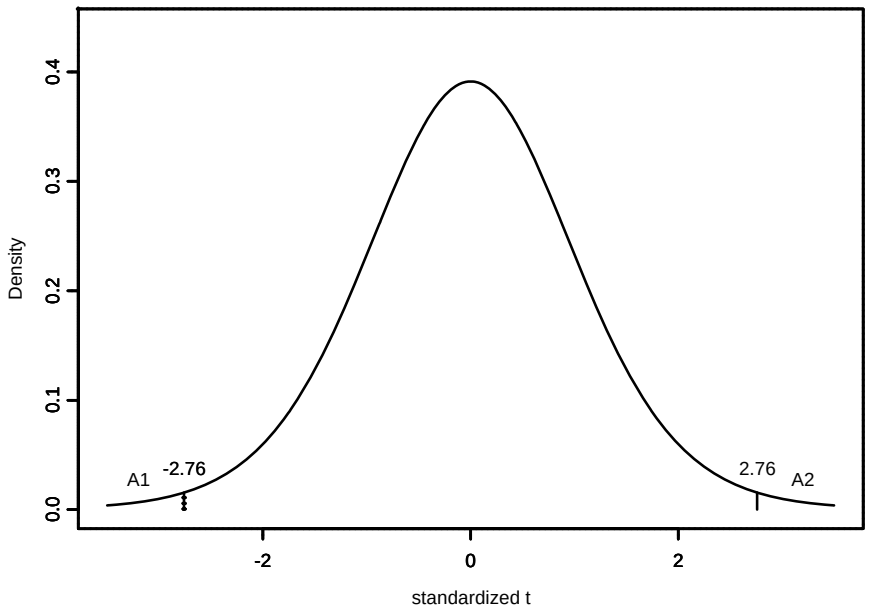
Y = respiratory (ventilation) rate

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
Sample size	$n_1 = 8$	$n_2 = 8$
Sample mean	$\bar{y}_1 = 6.169$	$\bar{y}_2 = 5.291$
Sample SD	$s_1 = 0.621$	$s_2 = 0.652$

We also found $SE_{\bar{y}_1 - \bar{y}_2}$ = 0.31834, so here, $t_s = \frac{(6.169 - 5.291) - 0}{0.31834} = 2.76$. This means that $\bar{y}_1$ and $\bar{y}_2$ differ by 2.76 SEs. But is that large?



Is our data consistent with H_0? Subject to the needed assumptions and requirements (and not assuming equal variances), if H_0 is true, then the sampling distribution of the TS t_s is approximately a t-distribution with df (degrees of freedom) given by Equation 6.7.1; df = 13 here. We wonder if t_s is 'far out' in the tail. The yardstick is the P-value (denoted 'p' and not to be confused with the Binomial 'p'). For the Hypnosis example, $t_s = 2.76$ *appears* to be far out in the tail.

For this *directionless alternative* case, the P-value of the test is the area under the Student's t curve in the two tails to the left of $-t_s$ and the right of t_s; in the above graph it is the sum of the areas A1 and A2. The p-value is a measure of compatibility between the data and H_0.

> **The P-VALUE is the probability of observing a test statistic as extreme as or more extreme than the one actually observed (where 'extreme' is defined in H_A).**

For the Hypnosis example, since Pr{t > 2.76} = 0.0081, the p-value is p = 2(0.008) = 0.0162 (obtained from the computer).

Making Decisions using HT (see p.227):

1. **Fix α (the significance level) a priori; usually $\alpha = 5\%$**
2. **If (p-value) $p < \alpha$, then H_0 is rejected at the α significance level**
3. **If $p > \alpha$, fail to reject H_0 – i.e., retain H_0**

For the Hypnosis example, suppose we had picked $\alpha = 5\%$ before viewing the data. Here, the p-value (0.0162) is less than 0.05 (= α), so

we reject H_0: $\mu_1 = \mu_2$ at the 5% significance level. Conclusion: At the 5% level, there appears to be a significant difference between the average respiratory rate of hypnotized and non-hypnotized males.

Note: the above p-value was obtained using the computer; had we used Table 4, the best we could say is that $0.01 < p < 0.02$, but we still reach the same conclusion as above since $p < 0.05$.

D. Bacteria Example (p.233, ex. 7.2.16). Y = bacteria colony count

μ_1 = average bacterial colony count of all Petri dishes to which sterile water (control) has been or can be added

μ_2 = average bacterial colony count of all Petri dishes to which soap has been or can be added

	Control	Soap
Sample size	$n_1 = 8$	$n_2 = 7$
Sample mean	$\bar{y}_1 = 41.8$	$\bar{y}_2 = 32.4$
Sample SD	$s_1 = 15.6$	$s_2 = 22.8$

At the 5% significance level, let's test if the means differ.

H_0: $\mu_1 = \mu_2$ (i.e., $\mu_1 - \mu_2 = 0$)

H_A: $\mu_1 \neq \mu_2$ (i.e., $\mu_1 - \mu_2 \neq 0$)

Since it turns out here that $SE_{\bar{y}_1 - \bar{y}_2} = 10.23$, the test statistic is $t_s = \frac{(41.7-32.4)-0}{10.23} = 0.91$. Also, the RHS of Equation 6.7.1 is 10.42 so we use df = 10. Using Table 4 (t-table), let's approximate the p-value: from the table, we can bracket our TS ($t_s = 0.91$) between 0.879 and 1.372, so the RH area is between 0.10 and 0.20, and so the p-value is

between 0.20 and 0.40 (and close to the 0.40). Regardless of the exact p-value, we do know that $p > \alpha$.

We do not "accept H_0" – rather, we "fail to reject H_0" or "retain H_0". We can never "prove the null hypothesis is true". Again quoting Carl Sagan, "Absence of evidence is not evidence of absence"; 'absence of evidence' here refers to the lack of proof of a difference in the population means (based on the nearness of our sample means).

Conclusion: The data do not provide sufficient evidence at the 5% level of significance to conclude that ordinary soap and sterile water differ in terms of the average number of bacteria.

Section 7.3: There's an important connection between CI's and HT's. For the Hypnosis example, the 95% CI for $(\mu_1 - \mu_2)$ was $(0.190, 1.566)$. Also, since $p < \alpha = 5\%$, we reject H_0: $\mu_1 - \mu_2 = 0$ and accept H_A: $\mu_1 - \mu_2 \neq 0$. For the Bacteria example, the 95% CI for $(\mu_1 - \mu_2)$ is $(-13.5, 32.2)$, and since $p > \alpha$, we retain (fail to reject) H_0: $\mu_1 - \mu_2 = 0$.

Thus, whenever the *α levels match* and the HT is non-directional, then (as the text proves on p.234-5):

- **Retain or Fail to reject H_0: $\mu_1 - \mu_2 = 0$ at the α level ← IFF → the $(1-\alpha)100\%$ CI for $(\mu_1 - \mu_2)$ contains zero;**
- **Reject H_0: $\mu_1 - \mu_2 = 0$ at the α level ← IFF → the $(1-\alpha)100\%$ CI for $(\mu_1 - \mu_2)$ does not zero**

With regard to interpreting α (p.236), note that α = Pr{reject H_0 given that H_0 is true}. It is best illustrated using a simulation study or a 'meta-experiment' as on p.237. Let μ_1 be that average IQ of all

Females, μ_2 be that average IQ of all Males, and suppose we know that $\mu_1 = \mu_2$ (i.e., that H_0 is true); say $\mu_1 = \mu_2 = \mu$. Then, repeatedly: (1) take samples of size n_1 from the Female $N(\mu, \sigma_1)$ population and find $\bar{y}_1$, (2) take samples of size n_2 from the Male $N(\mu, \sigma_2)$ population and find $\bar{y}_2$, (3) find the corresponding test statistics and decisions. Even though we know that H_0 is true ($\mu_1 = \mu_2$), α*100% (e.g., 5%) of the time we'll make the wrong decision and conclude $\mu_1 \neq \mu_2$.

In general, if we reject H_0, then either

- **H_0 is in fact false (i.e., we made a good decision), or**
- **H_0 is true and we were among the unlucky 5% who rejected H_0 anyway**

The probability of this latter event (α) is controlled for in that it is set before the study has started and data obtained.

Making Mistakes and Good Decisions in HT:

There are two types of errors we can make – called Type I and Type II errors, and displayed in the following table:

		H_0 is True	H_0 is False
Our Decision	Do not reject H_0	Correct ☺	Type II error
	Reject H_0	Type I error	Correct ☺

α = Pr{Type I error} – controlled at the outset

β = Pr{Type II error} – usually not controlled (but can be reduced by increasing the sample size)

POWER = 1 - β = Pr{Reject H_0 given that H_0 is False} – "the ability to see a difference when there really is a difference".

As discussed in Chapter 1, the Smoking and Birthweight example on pp.243-7 illustrates an observational study. Here:

- **Response variable: baby's birthweight**
- **Explanatory variable: whether mother smoked during pregnancy**
- **Extraneous variable(s): age, income, education, diet …**
- **Observational unit (OU): a mother-child pair**

As we encounter observational studies, we need to look for and consider possible sources of bias and then to imagine how these would impact the study and findings. We also need to ask ourselves whether the sample was truly random or simply a 'sample of convenience'; an example of the latter is Ex. 7.4.4 on p.245.

In the Smoking study, it could well be the case that diet is *confounded* with whether or not a person smokes – so too might alcohol consumption (AC) be confounded with smoking status.

S → low birthweight

S ---(???)→ low birthweight

AC

Recent smoking studies – and all epidemiological studies in general – measure many variables (e.g., amount of coffee consumed, age at onset of menstruation, age, weight, height, blood type, religion, education, income…), and the correct statistical analysis controls for these additional variables. These studies do find a link between smoking and baby birthweight.

But with all observational studies, remember

"Association is not causation."

Why then perform observation studies? First, may be the only way to study something like the effects of smoking (consider the alternative). Second, observational studies can be pooled or can give *suggestions* for subsequent experiments and research.

Spurious association (pp. 247-8): babies exposed to ultrasound in the womb were significantly lighter than babies not exposed to ultrasound. Does ultrasound cause reduced birthweight? Of course not – the lurking variable in this study was whether or not a mother was experiencing problems during pregnancy.

In Section 7.5, we address Directional Alternatives or so-called One-Tailed t-tests. Previously, we have only considered non-directional alternatives, $H_A: \mu_1 \neq \mu_2$, which lead to the so-called 2-tailed test. Here, we consider either of these cases:

$H_0: \mu_1 = \mu_2$ $\quad$ $H_0: \mu_1 = \mu_2$

$H_A: \mu_1 < \mu_2$ $\quad$ **OR** $\quad$ $H_A: \mu_1 > \mu_2$

Both are called one-tailed or one-sided t-tests for two independent samples. (Note in passing: regardless of the case, the equal sign always appears in the null hypothesis.) It is important to realize that it is legitimate to use a directional alternative *only if H_A is formulated before seeing the data*. As we shall see, the difference is in calculating the associated p-value.

Here is the One-Tailed Alternative Procedure:

Step 1. Check the directionality of the data: see if the data deviate in the direction specified by H_A. If not, then the p-value exceeds 50% and we stop since $p > \alpha$. If it does, proceed to step 2.

Step 2. The p-value of the data is the ONE TAILED area beyond t_s.
Step 3. Again, reject the null and accept H_A if $p < \alpha$, and retain the null (fail to reject) if $p > \alpha$.

Thus, if H_A: $\mu_1 < \mu_2$, then we only look to the left of t_s; if H_A: $\mu_1 > \mu_2$, then we only look to the right of t_s. Also, since all our CI's are two-sided, the CI-HT connection is meaningless for one-tailed tests.

Do we reject the Null in each of the following four cases?

p. 258, ex. 7.5.3: H_0: $\mu_1 = \mu_2$ and H_A: $\mu_1 > \mu_2$

(a) $t_s = 3.75$, df = 19, $\alpha = 0.01$, decision? _____

(d) $t_s = 1.8$, df = 7, $\alpha = 0.05$, decision? _____

p. 258, ex. 7.5.4: H_0: $\mu_1 = \mu_2$ and H_A: $\mu_1 < \mu_2$

(c) $t_s = 0.4$, df = 16, $\alpha = 0.10$, decision? _____

(d) $t_s = -2.8$, df = 27, $\alpha = 0.01$, decision? _____

p. 259, ex. 7.5.9: Wounded Tomato Plants: Does wounding a tomato plant improve the plant's defense against subsequent insect attack? The researcher's *a priori* guess was "yes", and she wanted to perform the relevant HT at the 5% level. Here, Y = weight (in mg) of insect (tobacco hornworm) larvae after 7 days of attack.

μ_1 = average weight of all larvae on wounded plants

μ_2 = average weight of all larvae on non-wounded plants

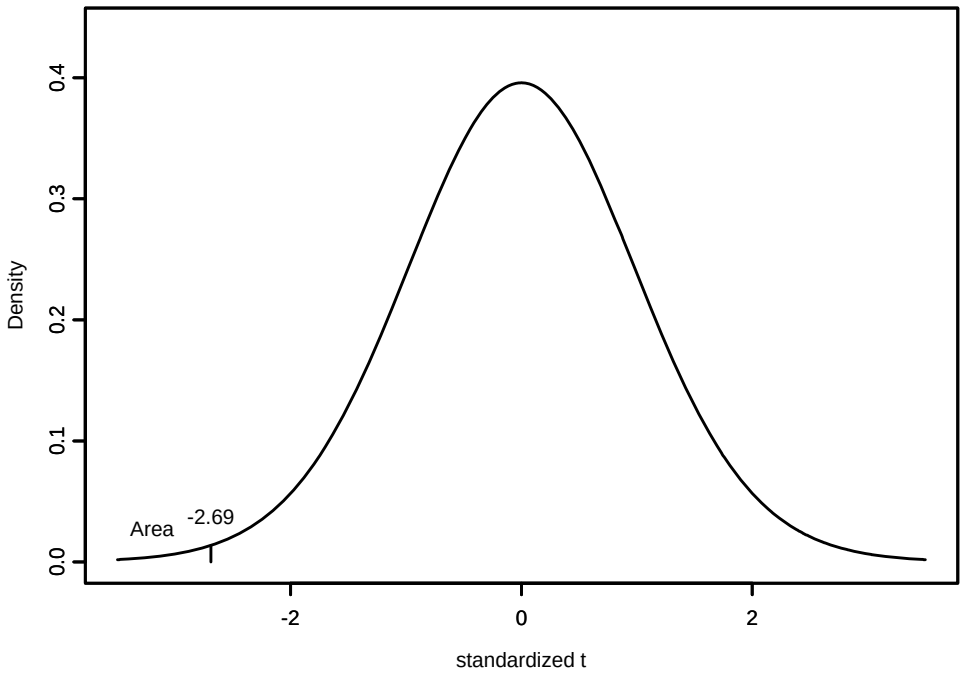
Here, we wish to test using $\alpha = 0.05$:

$$H_0: \mu_1 = \mu_2$$
$$H_A: \mu_1 < \mu_2$$

	Wounded	Control
Sample size	$n_1 = 16$	$n_2 = 18$
Sample mean	$\bar{y}_1 = 28.66$	$\bar{y}_2 = 37.96$
Sample SD	$s_1 = 9.02$	$s_2 = 11.14$

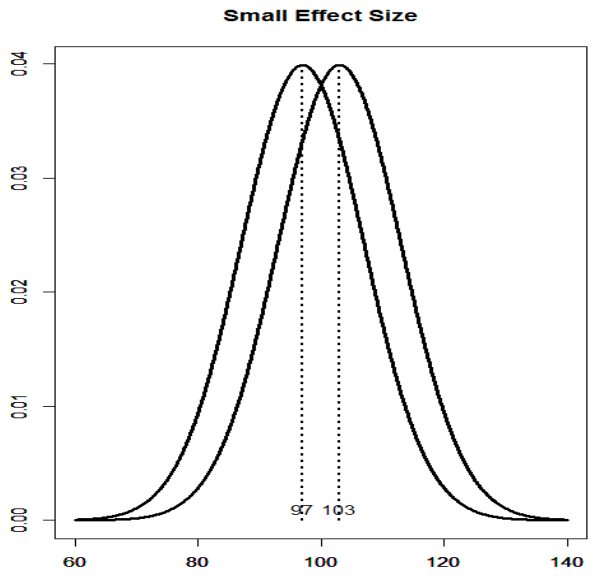
The RHS of Equation 6.7.1 equals 31.8 so df = 31. Making no assumptions about equality of the variances we have $SE_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{9.02^2}{16} + \frac{11.14^2}{18}} = 3.46$, and the test statistic is $t_s = \frac{28.66 - 37.96}{3.46} = -2.69$. From Table 4 (t-table), 0.005 < p < 0.01; since p < α, we reject the null. (Computer confirms that p = 0.0057 here.)

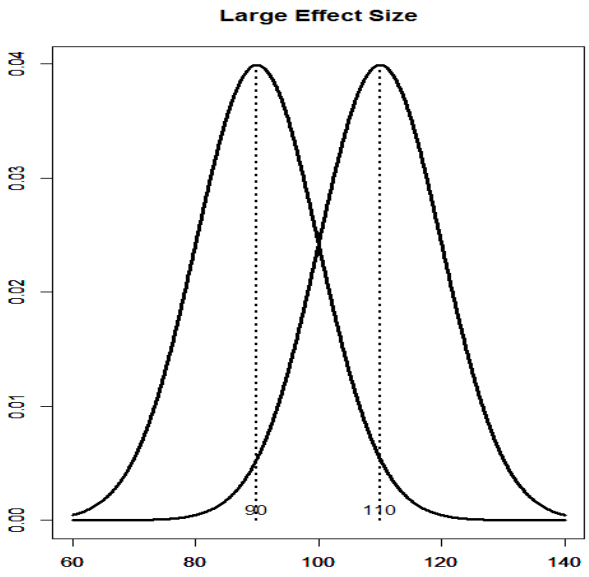
Conclusion: assuming that the respective larvae weight populations are Normally distributed, there is significant evidence to conclude that wounding a tomato plant does diminish average larvae growth.



Section 7.6 addresses the differences between significance and importance. We tend to use phrases like: "the effect of the drug was highly significant," or "the wheat yields did not differ significantly between the two fertilizers," or "no significant toxicity was found" (meaning the null hypothesis of no difference of the means was not rejected). But it is the researcher (consumer) who decides on importance and relevance – not the statistical test. For example, a study involving 10000's of volunteers in each of the two treatment arms may find the drug A cancer rate of 7.2% and the drug B cancer rate of 7.4% and the difference may be statistically significant, but consumers (the public, and perhaps MD's) need to decide for themselves if the difference is important enough to switch from the generic drug (A) to the expensive one (B).

For two treatment groups, if we can assume $\sigma_1 = \sigma_2 = \sigma$, then we define the Effect Size to be $\frac{|\mu_1 - \mu_2|}{\sigma}$, and we estimate this by the sample estimate $\frac{|\bar{y}_1 - \bar{y}_2|}{s}$; s = max(s_1,s_2). It is a signal-to-noise ratio.





Planning for Adequate Power is the topic of §7.7. Recall that power is the probability of rejecting H_0 when H_0 is indeed false (so H_A is true). If we have H_0: $\mu_1 = \mu_2$ vs. H_A: $\mu_1 \neq \mu_2$, then power is the ability to see the difference between the means when there really is one. Thus, power depends upon α, σ, n (total sample size) and $(\mu_1 - \mu_2)$ or the effect size. Taken one at a time (i.e., all other things equal) we have:

(a) **Dependence on α: If for example instead of $\alpha = 5\%$, we choose $\alpha = 1\%$, we'd reject H_0 less often, so the *power would drop*;**

(b) **Dependence on σ: A larger σ means less precision and so a *drop in power* – to increase power, control outside factors (so-called "noise") as much as possible;**

(c) **Dependence on n: As n increases, precision increases since SE's ($\sigma/\sqrt{n}$) drop, so *power increases*;**

(d) **Dependence on $(\mu_1 - \mu_2)$: Lots of shift in means → *increase in power* – investigator needs to ask him or her-self about the magnitude of the difference s/he is looking for; and this is captured in the *effect size*.**

p.270, ex.7.7.2 – Male and Female Heights. From a previous study, we have $\bar{y}_M$ = 69.1″, $\bar{y}_F$ = 64.1″, s = 2.5″, so our estimate of the effect size is $(69.1 - 64.1)/2.5 = 2.0$. In a future study, we wish to test H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 \neq \mu_2$ at the $\alpha = 5\%$ level. How many Males and Females do we need to sample so the power is 99%?

Using Table 5 on pp. 619-620, we see that we need $n_M = 11$ and $n_F = 11$ (so 22 total subjects). For additional examples work through ex. 7.7.3 on pp.270-1 and pp.271-2 ex. 7.7.2 (homework exercise).

§7.8 (Summary) reminds us of our assumptions/requirements:

- **Design – random samples, independence, representative;**
- **Distribution – Normality or large n; no outliers.**

If another design is used → use another statistical analysis (e.g., paired t-test in Chapter 8).

Violations could result in

(1) α is really higher than the specified 5%, and/or
(2) chosen t-test is not very powerful – i.e., another more powerful test exists and should be used.

Alternatives to the above t-test on the raw data include:

(a) performing the t-test on transformed data (see the text example on pp. 274-6), or
(b) performing a nonparametric test (see below and p.282).

§7.9 recaps hypothesis-testing strategies. Often – but not always – the null hypothesis is the "status quo" and the alternative hypothesis is "what the researcher wants to show." Recall, the "=" sign always occurs in H_0 in this course (although not so in *bioequivalence studies*). Read carefully the discussion on p.279: the probability that H_0 is true cannot be calculated – this is *not* what the p-value is equal to.

The Wilcoxon-Mann-Whitney (WMW) Test – is a two-independent-sample nonparametric test discussed in §7.10 (pp.282-289). Here, "nonparametric" means no assumption of a distribution such as the Normal distribution. This is an unfortunate choice of words since the WMW test is equivalent to a test of equality of the two population medians (which are indeed parameters).

Sample 1 measurements come from population 1 – denoted Y_1
Sample 2 measurements come from population 2 – denoted Y_2

H_0: the populations of Y_1 and Y_2 are the same
H_A: depends on the exercise – sometimes looking for a difference, sometimes for a Right or Left shift

Soil Respiration and Microbial Activity Example (p.282 ex.7.10.1). Here, Y is amount of CO_2 given off from core samples (in mol/g-soil/hour) at two locations in a forest: (1) under an opening in the forest canopy ("gap" locations), or (2) at a nearby area under heavy tree growth ("growth" locations). For this example, we have:

H_0: Gap and Growth areas do not differ w/r/to soil respiration.
H_A: Soil respiration rates tend to be DIFFERENT in the growth area and the gap area.

The data are repeated in the following table in black:

GROWTH		GAP	
# of GAP smaller	Y_1: GROWTH data	Y_2: GAP data	# of GROWTH smaller
5	17	6	0
6	20	13	0
6.5	22	14	0
8	64	15	0
8	170	16	0
8	190	18	1
8	315	22	2.5
		29	3

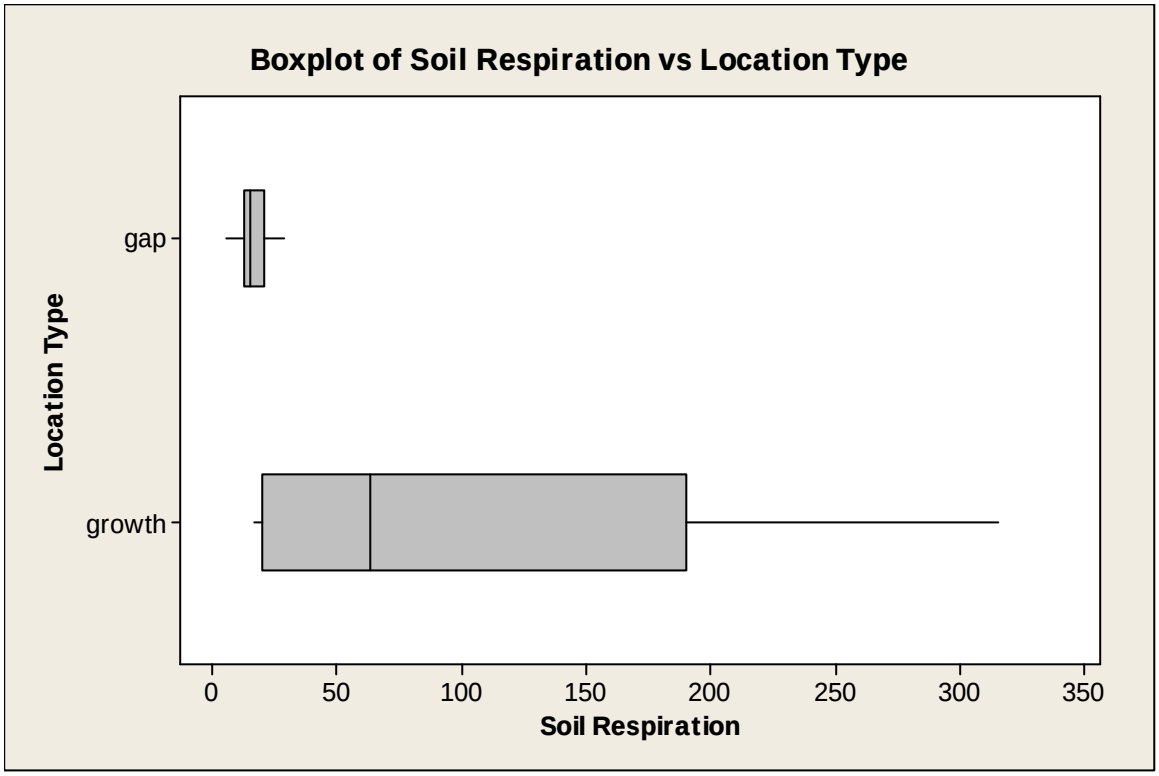
Now add up the red numbers: $K_1 = 49.5$ and $K_2 = 6.5$ (check that the sum of these, 56 here, is equal to the product of the above sample sizes $n_1 * n_1 = 7 * 8 = 56$ for this exercise). Then the Wilcoxon-Mann-Whitney test statistic is

$$U_s = \max\{K_1, K_2\}$$

Here, $U_s = 49.5$. Since $48 < U_s < 50$, from Table 6 on p.621, we have

$$0.0093 < \text{p-value} < 0.021$$

The following Box-Plot and MTB output confirm these results.



Mann-Whitney Test and CI: growth, gap

```
        N  Median
growth  7    64.0
gap     8    15.5

Point estimate for ETA1-ETA2 is 48.5
95.7 Percent CI for ETA1-ETA2 is (3.0,176.0)
W = 77.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0151
The test is significant at 0.0150 (adjusted for ties)
```

Conclusion: With a 5% level of significance, there is sufficient evidence to conclude that soil respiration rates are different in the gap and the growth areas. This conclusion is based on the assumptions/requirements that the soil measurements and locations in this study constitute a random sample and are therefore representative and free of bias.