Sections 6.6 and 6.7

Note made by: Timothy Hanson
Instructor: Peijie Hou

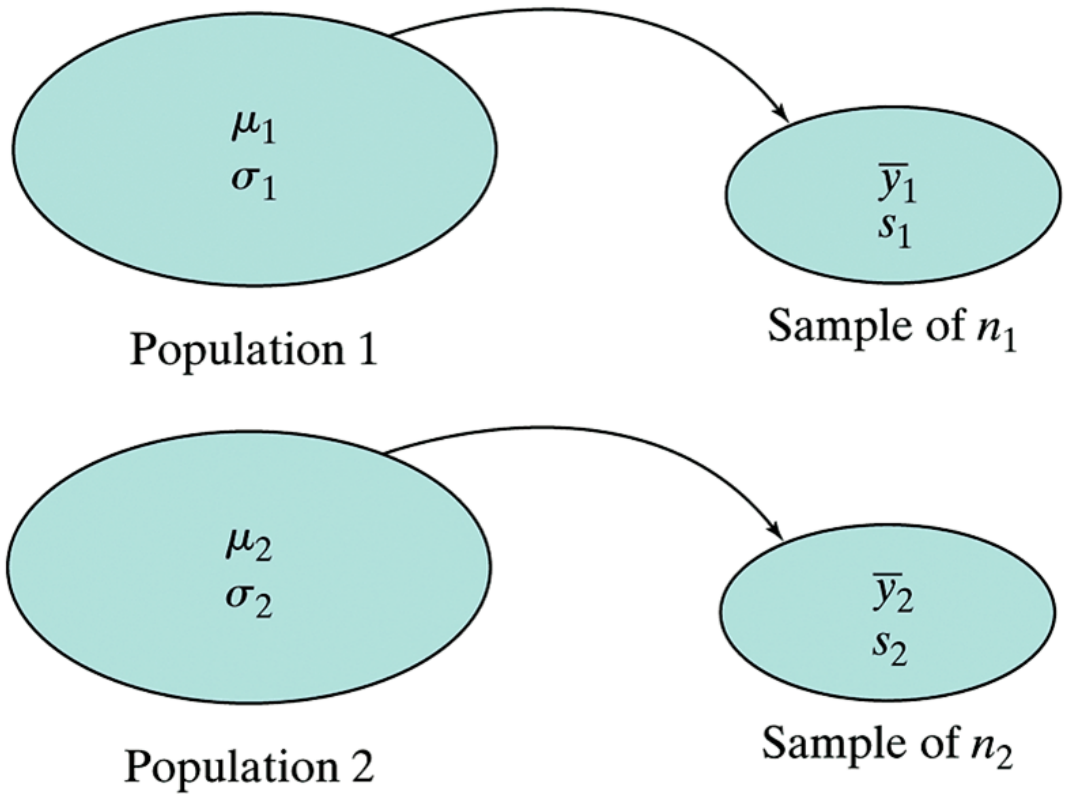
Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Comparing two populations

- Much of scientific research is focused on *comparing populations*.
- Any aspect of populations can be compared: mean, median, 90*th* percentile, number of modes, skew, overall shape, standard deviation, etc.
- Most common is to compare *population* means.
- We now have *two populations*, 1 and 2, that differ according to some aspect such as treatment received, gender, etc.
- From population 1, $\bar{y}_1$ estimates μ_1 and s_1 estimates σ_1 (sample size n_1).
- From population 2, $\bar{y}_2$ estimates μ_2 and s_2 estimates σ_2 (sample size n_1).

Notation for comparison of two samples



Example 6.6.1 Vital capacity

Amount of air expelled after a deep breath was measured on $n_1 = 8$ brass instrument (trumpet, trombone, french horn, etc.) players compared to $n_2 = 5$ controls (don't play brass instrument).

Table 6.6.1 Vital capacity (liters)

	Brass player	Control
	4.7	4.2
	4.6	4.7
	4.3	5.1
	4.5	4.7
	5.5	5.0
	4.9	
	5.3	
n	7	5
$\bar{y}$	4.83	4.74
s	0.435	0.351

- Brass players, $\bar{y}_1 = 4.83$ estimates μ_1 and $s_1 = 0.435$ estimates σ_1.
- Control group, $\bar{y}_2 = 4.74$ estimates μ_2 and $s_2 = 0.351$ estimates σ_2.

$\bar{Y}_1 - \bar{Y}_2$ estimates $\mu_1 - \mu_2$

- A natural estimate of $\mu_1 - \mu_2$ is $\bar{Y}_1 - \bar{Y}_2$.
- The variability in $\bar{Y}_1 - \bar{Y}_2$ is estimated by

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

  the **standard error of** $\bar{Y}_1 - \bar{Y}_2$.
- For the vital capacity data,

$$SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{0.435^2}{7} + \frac{0.351^2}{5}} = 0.23 \text{ liter}.$$

A rough CI for $\mu_1 - \mu_2$

- If both populations are normal and/or the sample sizes are big

$$\bar{Y}_1 \sim N(\mu_1, SE_{\bar{Y}_1}) \text{ and } \bar{Y}_2 \sim N(\mu_2, SE_{\bar{Y}_2}).$$

- The difference of two normals is also normal

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, SE_{\bar{Y}_1 - \bar{Y}_2}).$$

- As in Section 6.3, a normal is within 2 standard errors of its mean 95% of the time, so...
- A rough CI for $\mu_1 - \mu_2$ is $\bar{Y}_1 - \bar{Y}_2 \pm 2SE_{\bar{Y}_1 - \bar{Y}_2}$.
- For the vital capacity data, $\bar{y}_1 - \bar{y}_2 = 4.83 - 4.74 = 0.09$ liter and $SE_{\bar{Y}_1 - \bar{Y}_2} = 0.23$ liter. A rough 95% CI for $\mu_1 - \mu_2$ is $(0.09 - 2(0.23), 0.09 + 2(0.23)) = (-0.37, 0.55)$.

6.7 Confidence interval for $\mu_1 - \mu_2$

- The rough CI from the last section can be refined (W.S. Gosset again).
- A 95% CI for $\mu_1 - \mu_2$ is given by $\bar{y}_1 - \bar{y}_2 \pm t_{0.025} SE_{\bar{Y}_1 - \bar{Y}_2}$ where $t_{0.025}$ is the multiplier from a t distribution with degrees of freedom given by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4/n_1^2}{n_1 - 1} + \frac{s_2^4/n_2^2}{n_2 - 1}}.$$

- This *df* formula is due to Welch (1947) and Satterthwaite (1946). It doesn't give an integer; people generally round down.

CI for $\mu_1 - \mu_2$ in R

- R takes care of these details for us. If your two samples are called sample1 and sample2, t.test(sample1,sample2) will provide a 95% CI.
- The t interval approach is valid if the samples sizes are large enough ($n_1 > 30$ and $n_2 > 30$, say), or if the data populations are normal to begin with. For small sample sizes, we need to check that *both samples* are approximately normal.

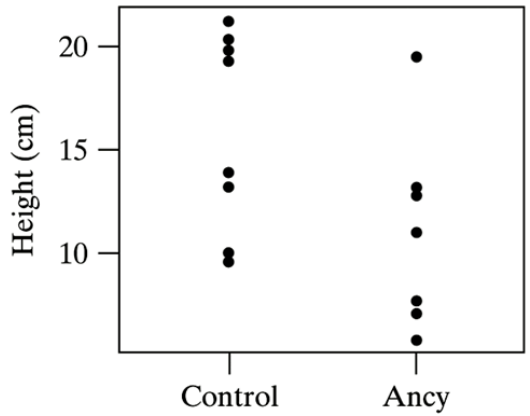
Example 6.7.1 Two-week height of control & ancy plants

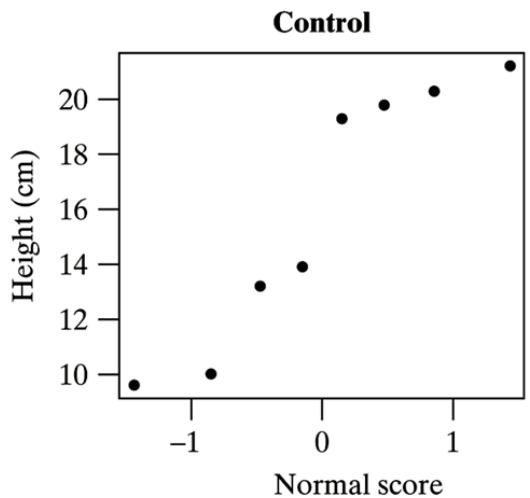
The Wisconsin Fast Plant grows fast. Ancymidol (ancy) slows growth. $n_1 = 8$ control (no ancy) and $n_2 = 7$ plants treated with ancy were measured (cm) after two weeks. We want to estimate the mean difference in growth between all regular and all ancy-treated plants, i.e. the two populations of plants.

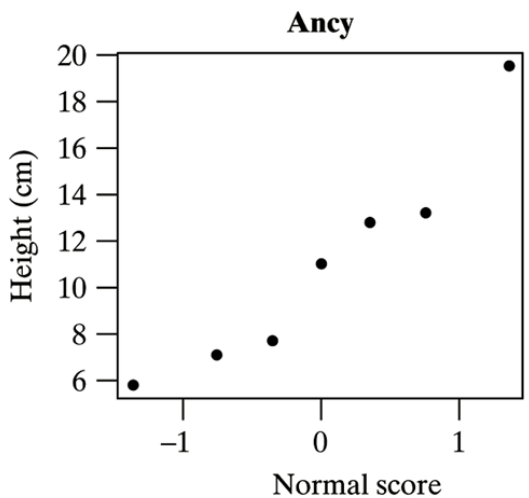
Table 6.7.1 Fourteen-day height of control and of ancy plants (cm)

	Control (Group 1)	Ancy (Group 2)
	10.0	13.2
	13.2	19.5
	19.8	11.0
	19.3	5.8
	21.2	12.8
	13.9	7.1
	20.3	7.7
	9.6	
n	8	7
$\bar{y}$	15.9	11.0
s	4.8	4.7
SE	1.7	1.8

Checking assumptions







CI in R

```
> control=c(10.0,13.2,19.8,19.3,21.2,13.9,20.3,9.6)
> ancy=c(13.2,19.5,11.0,5.8,12.8,7.1,7.7)
> t.test(control,ancy)

        Welch Two Sample t-test

data:  control and ancy
t = 1.9939, df = 12.783, p-value = 0.06795
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4182434 10.2146719
sample estimates:
mean of x mean of y
 15.91250  11.01429
```

We are 95% confident that the true mean difference is between -0.42 and 10.21 cm.

Vital capacity example

```
> brass=c(4.7,4.6,4.3,4.5,5.5,4.9,5.3)
> control=c(4.2,4.7,5.1,4.7,5.0)
> t.test(brass,control)

        Welch Two Sample t-test

data:  brass and control
t = 0.3899, df = 9.761, p-value = 0.705
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4192838  0.5964267
sample estimates:
mean of x mean of y
 4.828571  4.740000
```

We are 95% confident that the true difference in vital capacity between brass and non-brass is between -0.42 and 0.60 liter. How does the t interval $(-0.42, 0.60)$ compare to our "rough" interval $(-0.37, 0.55)$?

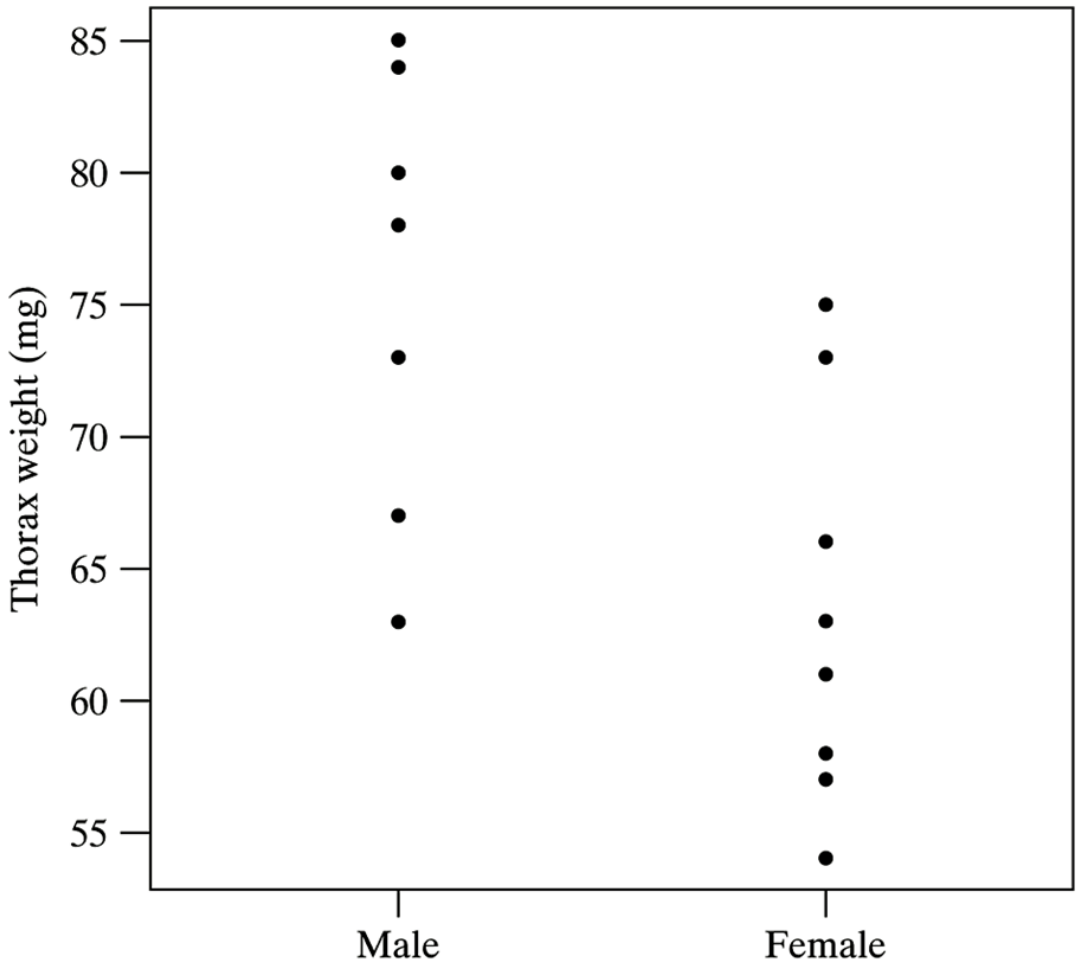
Example Thorax weight

Biologists think that male Moncarch butterflies have, on average, a larger thorax than females.

Table 6.7.2 Thorax weight (mg)

	Male	Female
	67	73
	73	54
	85	61
	84	63
	78	66
	63	57
	80	75
		58
n	7	8
$\bar{y}$	75.7	63.4
s	8.4	7.5
SE	3.2	2.7

Thorax weights



95% and 90% CI for $\mu_1 - \mu_2$ thorax weights

```
> male=c(67,73,85,84,78,63,80)
> female=c(73,54,61,63,66,57,75,58)
> t.test(male,female)

95 percent confidence interval:
  3.325484 21.353087

> t.test(male,female,conf.level=0.90)

90 percent confidence interval:
  4.962093 19.716479
```

We are 95% confident that all male Monarch butterflies have a thorax weight between 3.3 and 21.4 mg greater than females.

We are 90% confident that all male Monarch butterflies have a thorax weight between 5.0 and 19.7 mg greater than females.

Almost always, people report 95% CI's.

Interval for σ_1/σ_2

- Sometimes people want to see how population spreads compare.
- R provides a 95% CI for the ratio $\dfrac{\sigma_1}{\sigma_2}$.
- For example, comparing the spreads of the thorax weights from male to female

```
> var.test(male,female)
95 percent confidence interval:
 0.2425657 7.0714732
sample estimates:
ratio of variances
          1.241596
```

We estimate $\sigma_1/\sigma_2 = 1.24$ and are 95% confident that σ_1/σ_2 is between 0.24, and 7.07.

Review

- A *confidence interval* provides a plausible range for $\mu_1 - \mu_2$.
- Since $\bar{Y}_1 - \bar{Y}_2$ is normal, the 68/95/99.7 rule says $\mu_1 - \mu_2$ is within $\bar{Y}_1 - \bar{Y}_2 \pm 2SE_{\bar{Y}_1 - \bar{Y}_2}$ 95% of the time.
- This interval is too small; Gosset introduced the t distribution to make the interval more accurate $\bar{Y}_1 - \bar{Y}_2 \pm 2SE_{\bar{Y}_1 - \bar{Y}_2}$; the *df* for the t distribution is computed using the Welch-Satterthwaite formula.
- `t.test(sample1,sample2)` in R takes care of the details.
- For $n_1 < 30$ or $n_2 < 30$ the data must be normal; check this with two normal probability plots.
- Interpretation is important. "With 95% confidence the true mean difference in population characterstic is between a and b units."