Outline

1. t-methods for 2 independent samples
   - Independent samples, assuming equal variances
   - Independent samples, allowing unequal variances

2. General aspects of design
   - Experimental vs. Observational studies
   - Blinding
   - Control group
   - Missing data
   - Randomization
   - Nesting

Independent samples

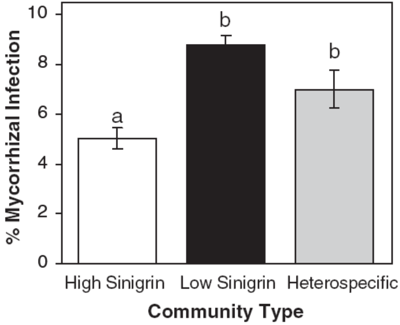


Fig. 5. Mycorrhizal infection potential (measured as the percentage of root sections colonized with mycorrhizal fungus) in soil from high- and low-sinigrin *B. nigra* communities and mixed heterospecific communities (means ± SE). Bars sharing the same letters are not statistically different.

Compare mycorrhizal colonization in soil from high-sinigrin black mustard communities (11 rep) and low-sinigrin black mustard

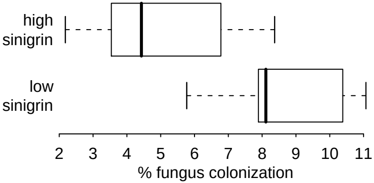
Is there evidence of an effect of black mustard (high/low sinigrin) on fungus colonization?

Mycorrhizal colonization example

Data: mycorrhizal colonization (% of root section), in high-sinigrin (hi) and low-sinigrin (lo) communities.

community	hi (Y_1)	lo (Y_2)
1	4.43	11.07
2	2.18	7.89
3	6.64	7.89
4	4.41	8.12
5	3.70	8.11
6	4.79	10.79
7	3.38	10.30
8	8.37	7.21
9	2.94	5.77
10	6.92	10.47
11	7.24	8.09

No pairing: observations can be permuted within each trt (column).



$\bar{y}_1 = 5$, $s_1 = 2.0$
$\bar{y}_2 = 8.7$, $s_2 = 1.7$

Mycorrhizal colonization

$\mu_1 =$ the population mean fungus colonization in communities assigned to high-sinigrin black mustard,
$\mu_2 =$ the population mean fungus colonization with low-sinigrin.

We want a confidence interval for $\mu_1 - \mu_2$, or test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$.
$\mu_1 = \mu_2$ means $\mu_1 - \mu_2 = 0$.

Main idea: use $\bar{Y}_1 - \bar{Y}_2$.

if $\bar{Y}_1 - \bar{Y}_2$ is close to 0, we will favor
if $\bar{Y}_1 - \bar{Y}_2$ is far from 0, we will favor

Here $\bar{y}_1 - \bar{y}_2 = 5 - 8.7 = -3.7$ (% root section difference).

Assumptions

1. Independent random samples Y_1 and Y_2: independence within a trt and between two trts

   > apart from the fact that all outcomes from the same sample share the same mean and variance.

2. Normality: the first sample $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ is from $\mathcal{N}(\mu_1, \sigma_1^2)$,
   second sample $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is from $\mathcal{N}(\mu_2, \sigma_2^2)$.
3. Equal variances: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

No need to have equal sample size.

Under these assumptions, we can pool the 2 samples to estimate their common variance:

Pooled estimated of σ^2

$$S_p^2 = \frac{\text{sum of all deviations}^2}{n_1 - 1 \ + n_2 - 1} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

weighted average of S_1^2 and S_2^2, weighted by the df's.

Standard error of $\bar{Y}_1 - \bar{Y}_2$

$$\text{SE}_{\text{pooled}} = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Confidence interval for $\mu_1 - \mu_2$

CI assuming equal variances

A $(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{\mathrm{df},\alpha/2} * \mathrm{SE}_{\mathrm{pooled}}$$

where df$= n_1 + n_2 - 2$, and recall $\mathrm{SE}_{\mathrm{pooled}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Black mustard: we had $\bar{y}_1 - \bar{y}_2 = 5 - 8.7 = -3.7$ (in % root section), $s_1 = 2.0$ and $s_2 = 1.7$, with $n_1 = n_2 = 11$.

Pooled estimate of σ: $s_p =$ $= 1.856$

Standard error of $\bar{Y}_1 - \bar{Y}_2$: $= .791$

df$=$ so t-multiplier: $t = 2.086$ for 95% confidence.

Interval: $-3.7 \pm 2.086 * 0.791 = (-5.35, -2.05)$

The two-sample t-test

1. Hypotheses: $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$
2. Test statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - 0}{\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2,\,\text{pooled}}}$$

If H_0 is really true then $T \sim$ t-distribution with $\text{df} = n_1 + n_2 - 2$

3. Get p-value: we had $\bar{y}_1 - \bar{y}_2 = 5 - 8.7 = -3.7$ (in % root section),
   Pooled standard error of $\bar{Y}_1 - \bar{Y}_2$: $\mathrm{SE} = .791$
   t-value: $t =$ $= -4.67$
   df= , so p-value: $2\mathbb{P}\{T_{20} < -4.67\} < .001$

4. We H_0. Or: There is
   evidence that fungus colonization is affected by the type of black mustard community.

The test and the CI are consistent

0 is outside the $(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$

$$\Leftrightarrow$$

$\mu_1 - \mu_2 = 0$ is rejected at level α, i.e. p-value for this test is $< \alpha$.

fungus colonization: p-value was < 0.001.

95% CI for $\mu_1 - \mu_2$:	$[-5.35, -2.05]$ % of root sections,
99% CI (check at home):	$[-5.95, -1.45]$
99.9% CI :	$[-6.74, -0.65]$

R command: t.test()

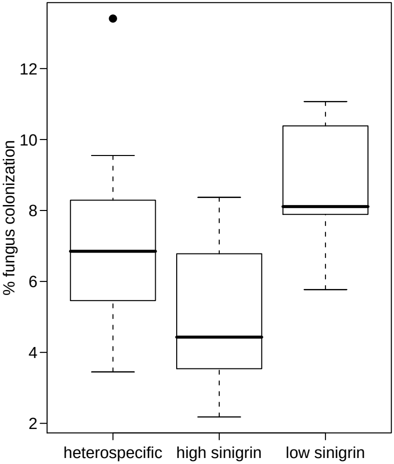
```
> # enter the data:
> lo = c(11.07,7.89,7.89,8.12,...,7.21,5.77,10.47,8.09)
> hi = c( 4.43,2.18,6.64,4.41,...,8.37,2.94, 6.92,7.24)
>
> # do the test:
> t.test(hi, lo, var.equal=T, conf.level=.99)

        Two Sample t-test

data:  hi and lo
t = -4.6786, df = 20, p-value = 0.0001444
alternative hypothesis: true difference in means is not
99 percent confidence interval:
 -5.951651 -1.450167
sample estimates:
mean of x mean of y
 5.000000  8.700909
```

Another example

Compare fungus colonization in high-sinigrin and in heterospecific communities (mixed species, no black mustard).



$\bar{y}_{\text{het}} = 7.0$, $s_{\text{het}} = 2.1$, $n_{\text{het}} = 33$
$\bar{y}_{\text{hi}} = 5.0$, $s_{\text{hi}} = 2.0$, $n_{\text{hi}} = 11$
$\bar{y}_{\text{lo}} = 8.7$, $s_{\text{lo}} = 1.7$, $n_{\text{lo}} = 11$.

Test $\mu_{\text{hi}} = \mu_{\text{het}}$.

Welch T-test allowing unequal variances

```
> liz
   tail.length  location
1          8.8   bigbend
2          9.7   bigbend
3         10.8   bigbend
...
38        10.3 boxcanyon
39         9.5 boxcanyon
40        11.4 boxcanyon
> with(liz, tapply(tail.length, location, mean))
  bigbend boxcanyon
   8.8958   10.1812
> with(liz, tapply(tail.length, location,   sd))
  bigbend boxcanyon
   1.4299    1.0316
```

	Big Bend	Box Canyon
n	24	16
mean	8.90	10.18
sd	1.43	1.03

$\bar{y}_1 - \bar{y}_2 = -1.28$, but how big could that be by chance alone?

Standard Error for $\bar{y}_1 - \bar{y}_2$, not requiring equal variance

SE's don't add up, but variances do:

$$\mathrm{SE}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\mathrm{SE}^2_{\bar{y}_1} + \mathrm{SE}^2_{\bar{y}_2}}$$

that is:

$$\mathrm{SE}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Welch t-test: not requiring equal variances

Test of H_0: $\mu_1 = \mu_2$, no variance requirement

Same test statistic: $T = \dfrac{\bar{Y}_1 - \bar{Y}_2}{\mathrm{SE}_{\bar{y}_1 - \bar{y}_2}}$ but $\mathrm{SE}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}$.
The p-value is obtained by comparing the value of T with a t-distribution with adjusted degree of freedom

$$\mathrm{df} = \frac{(\mathrm{v}_1 + \mathrm{v}_2)^2}{\frac{\mathrm{v}_1^2}{n_1 - 1} + \frac{\mathrm{v}_2^2}{n_2 - 1}}$$

where $\mathrm{v}_1 = \mathrm{SE}_1^2 = S_1^2 / n_1$ and $\mathrm{v}_2 = \mathrm{SE}_2^2 = S_2^2 / n_2$.

df will not necessarily be an integer, but round it down.
df always $\leq n_1 + n_2 - 2$ and $\geq$ the minimum of $n_1 - 1$ and $n_2 - 1$.

Lizard tail lengths

$\bar{y}_1 - \bar{y}_2 = -1.28$ cm longer tails in Big Bend lizards than Box Canyon lizards.

$\mathrm{v}_1 = 1.43^2/24 = .0852$, $\mathrm{v}_2 = 1.03^2/16 = .0665$, then

$\mathrm{SE}_{\bar{y}_1 - \bar{y}_2} = \sqrt{.0852 + .0665} = 0.389$ cm, and

df$=$ 37.7 ($>$ smallest of 23 and 15, and $\geq 23 + 15 = 38$)

$$t = -1.28/0.389 = -3.30$$

Table C with df$=$ 37: $.001 <$ p-value $< .01$ with a two-sided test.

There is **strong evidence** that the 2 lizard populations have different mean tail lengths.

Welch confidence interval

CI for $\mu_1 - \mu_2$ allowing unequal variances

A $(1-\alpha)$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \quad \pm \quad t * \mathrm{SE}_{\bar{y}_1 - \bar{y}_2}$$

where $t = t_{\mathrm{df},\alpha/2}$ and df is the **adjusted degree of freedom**, and

$$\mathrm{SE}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\,.$$

For lizard tails and 90% confidence we get $t = t_{37.7,.05} = 1.686$ and interval:

$$-1.28 \pm 1.686 * 0.389$$

i.e. [-1.94, -0.63] more cm on average for Big Bend lizards than Box Canyon lizards, i.e. [0.63, 1.94] more cm for Box Canyon lizards on average.

There are many ways to use t.test()

```
> bigbend
 [1]  8.8  9.7 10.8  7.1  6.6  9.9 10.2  8.6 10.4 11.9  7.6  8.0 ...
[16]  7.4  8.3  9.1  9.2  7.9  8.4 11.3  6.2  8.8
> boxcanyon
 [1] 10.7  8.8  9.9 10.9 10.4 11.1 12.0  9.5 10.9  8.1  9.0  9.8 ...
[16] 11.4

> t.test(bigbend, boxcanyon)

        Welch Two Sample t-test

data:  bigbend and boxcanyon
t = -3.3001, df = 37.699, p-value = 0.002119
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0741421 -0.4966912
sample estimates:
mean of x mean of y
 8.895833 10.181250

> t.test(bigbend, boxcanyon, var.equal=T)
... t = -3.0933, df = 38, p-value = 0.003701
... 95 percent confidence interval: -2.1266512 -0.4441821
```

```
> liz
   tail.length  location
1          8.8   bigbend
2          9.7   bigbend
...
39         9.5 boxcanyon
40        11.4 boxcanyon

> t.test(tail.length ~ location, data=liz)

        Welch Two Sample t-test

data:  tail.length by location
t = -3.3001, df = 37.699, p-value = 0.002119
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0741421 -0.4966912
sample estimates:
  mean in group bigbend mean in group boxcanyon
               8.895833               10.181250
```

Which t-test should I use?

Assumptions (Welch t-test, not requiring equal variances)

1. Independence, within and among samples,
2. Each sample comes from a Normal distribution or is large enough.

If software allows: Welch t-test by default.

On exams: either (unless indicated otherwise), but assess the equal variance assumption if using standard t-test: Definitely use the Welch t-test if

the sample SDs differ by 3-fold or more

or the sample sizes are very different.

Outline

1. t-methods for 2 independent samples
   - Independent samples, assuming equal variances
   - Independent samples, allowing unequal variances

2. General aspects of design
   - Experimental vs. Observational studies
   - Blinding
   - Control group
   - Missing data
   - Randomization
   - Nesting

Experimental study:

Compare 2 drugs and determine which is more effective in controlling some form of cancer:

> take a # of patients available,
>
> randomly assign them to drug type.

The experimenter controls the randomization to experimental conditions (who gets what), although there can be other factors (age, weight, etc.) that affect the outcome.

Observational study:

on the effect of asbestos on causing some form of cancer.

> Take a group of people with this type of cancer, see what proportion had exposure to asbestos.
>
> Look at a control group without cancer, see what proportion had exposure.

No random assignment of individuals to treatment.

How is the control group selected? Difficult!

Experimental versus Observational studies

In general, observational studies are more difficult to carry out, analyze and interpret.

Babies and Smoking Example.
Pregnant women (smoking and control: non-smoking) were followed with their babies. There was strong statistical evidence that the mean birth weight of smoker’s babies is lower than mean birth weight of non-smokers babies.

Association is not causation - Confounded effects

We can conclude that smoking and light babies are **associated**.
We cannot conclude that smoking **causes** lower weight.

Ice cream and drowning deaths: There is strong evidence that the average # of drowing deaths is higher on days when ice cream sales are higher than on days when ice cream sales are lower.

Association ice cream sales $\leftrightarrow$ # drowing deaths
No causal relationship!

Possible confounder:

Association is not causation - Confounded effects

We can conclude that smoking and light babies are **associated**.
We cannot conclude that smoking **causes** lower weight.

Possible effects of smoking could be confounded with many other explanations:

Woman's weight,

nutritional habits,

age, activity, etc.

To turn association into causation, need to compare 2 groups (smoking/non smoking) very similar with respect to all other things.

Second study:

A large # of variables were measured. A complex statistical method that simultaneously estimates the effects of several explanatory variables (weight, activity, nutritional habits, race, etc.) found that even after making adjustments for these other factors, smoking still had an effect on birth weight.

(Did these people think about **all** explanatory variables?)

Third study:

found differences in the placenta between smokers and non-smokers, and some of the differences were associated with chemicals found in cigarettes. Also found that having smokers not smoke for 3 hours caused a change in blood flow to the placenta.

Fourth study:

159 women smoked during the 1st pregnancy but not during 2nd pregnancy. Matched with 159 other women who had smoked during both pregnancies and for whom other explanatory variables (age, etc.) were similar. Found that those who quit smoking had heavier 2nd babies than those who continued to smoke.

It takes all this to address confounding and establish causal relationship!

Experimental studies more powerful than observational studies:

to show causation,
if the units are randomized
to avoid confounding effects.

(more on randomization later)

Importance of blinding

Blind experiment: The experimental subject does not know what treatment he/she is receiving.
Double-blind: the physician, or the person making the measurements does not know the treatment.

One gets excited about a new treatment (researchers and physicians especially!), and may give better care to patients receiving the new treatment, either consciously or not.

Blinding is extremely important!

Importance of control groups

College students volunteered in experiment to test a vaccine for preventing the common cold.

		Mean # of colds	
	n	previous year	current year
Vaccine	201	5.6	1.7

Placebo effect is strong! The drop in the vaccine group might be due to the placebo effect only.

The # of colds were based on students' memories (previous year) or active count (current year). Can it make a difference?

Missing data should not be ignored

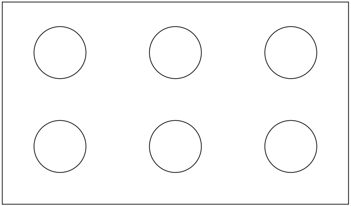
A study compares 2 drugs. Some patients died over the course of the study. No data about their blood components.
If they died because of the treatment they received, it would not be appropriate to drop them from the analysis.

Some patients dropped from the study. For these patients, partial data is available.
If they dropped because they thought the treatment was not working, it is not appropriate to exclude them from the analysis.

In general, ignoring missing data can bias the conclusions.

Completely randomized design (CRD)

Consider 2 treatments (A and B), a greenhouse bench, and 6 available pots.



Allocate 3 A’s and 3 B’s, at random:

1-3 receive A, 4-6 receives B.

Then randomly assign numbers to pots.

```
> sample(1:6)
[1] 1 3 4 6 5 2
```

Here number of pots known in advance.

In medical studies, very often, we don’t know how many patients will accept to enter the study. Each time a patients comes up, we can toss a coin. Heads: A, Tail: B. We might not get as many A’s as B’s at the end of the study.

Or: randomize (toss a coin) every other patient.

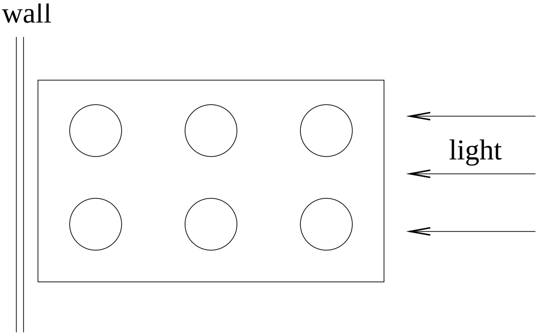
Why must we randomize?

To **avoid bias**, known or unknown.

Randomization will avoid confounding with other variables. Age, gender, weight, sunlight, temperature, etc., will tend to be equally balanced in the 2 groups.

There are many different ways to randomize. Be creative!

Randomized complete block design (RCBD)



Light is known to affect outcome, even if we don’t care!

Pots divided into blocks, such that conditions are similar within a block.

Randomize within each block: $\begin{array}{c}\text{A}\\\text{B}\end{array}$ or $\begin{array}{c}\text{B}\\\text{A}\end{array}$.

Toss a coin only once for each block → reduction of randomization.

Randomized complete block design (RCBD)

Medical study: individuals are grouped into blocks, or strata.

young men	young women
old men	old women

A, B B, A B, A	B, A A, B B
B, A A, B A	B, A

Randomize within blocks: ensure about the same number of men and women, and about the same number of old and young subjects in all treatment. No bias toward younger people in group A, or toward one gender in group A.

When should we block?

We use blocking if it helps remove variability.

We should use blocks when

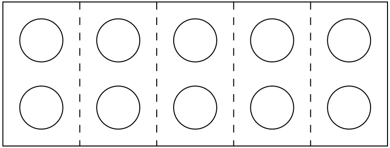
Homogeneity within blocks: outcome expected to be similar within blocks,
Major variability expected between blocks.

Example: Pairing. Study of effect of coffee on pulse rate.

1 block = 1 subject

before coffee

after coffee

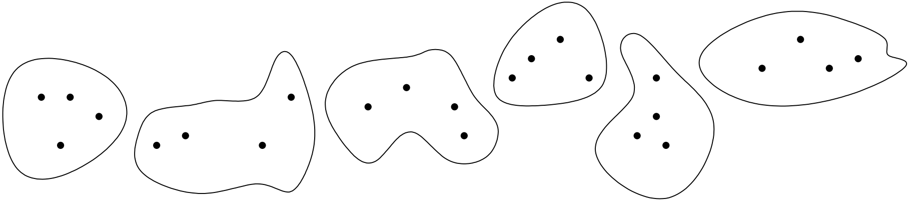


5 sujects total

With pairing, variability within “columns” is much reduced.

Nesting

Bioremediation of contaminated soil. We wish to compare 2 treatments on a lead compound. We have 6 “large” areas. We assign treatments by CRD.



Only 1 treatment can be assigned to each area, but 4 measures (soil cores) on lead compound are taken from each area: makes 24 measures.

Nesting

treatment A			treatment B		
Area 1	Area2	Area 3	Area 4	Area 5	Area 6
y_{11}	y_{21}	y_{31}	y_{41}	y_{51}	y_{61}
y_{12}	y_{22}	y_{32}	y_{42}	y_{52}	y_{62}
y_{13}	y_{23}	y_{33}	y_{43}	y_{53}	y_{63}
y_{14}	y_{24}	y_{34}	y_{44}	y_{54}	y_{64}

Is it reasonable to group all “A” observations (left), group all B observations (right), and perform a 2-independent sample t-test to compare the 2 groups?

No! The 4 readings on each plot are expected to be similar to each other. Readings are **nested** within plots.

1. Use a more complex model and analysis,
2. Or: reduce data to plot means, and do a 2-sample t-test on these data.