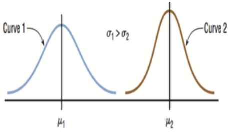
Two independent samples

$y_{11}, y_{12}, \cdots, y_{1n_1} \sim N(\mu_1, \sigma_1^2)$.
$y_{21}, y_{22}, \cdots, y_{2n_2} \sim N(\mu_2, \sigma_2^2)$.



$\bar{y}_{1\cdot}, \bar{y}_{2\cdot}$: sample mean.
$\bar{y}_{1\cdot} = \frac{y_{11}+y_{12}+\cdots+y_{1n_1}}{n_1}$; $\bar{y}_{2\cdot} = \frac{y_{21}+y_{22}+y_{2n_2}}{n_2}$
dot here indicates summation over the 2nd subscript.
s_1, s_2 : sample standard deviation.
Purpose: compare μ_1, μ_2.

Assume data come from

- an experiment where the experimental units for each treatment are randomly selected from the available units.
- or an observational study where the two samples are randomly and independently selected from two different populations.

sampling distribution of $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$

Recall that $\bar{y}_{1\cdot} \sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \bar{y}_{2\cdot} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$.
and because the two samples are independent, we have
$\bar{y}_{1\cdot} - \bar{y}_{2\cdot} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$.
(a linear combination of independent normal randoms variables also follows a normal distribution)
Note here $E(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = E(\bar{y}_{1\cdot}) - E(\bar{y}_{2\cdot}) = \mu_1 - \mu_2$, and
$Var(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = Var(\bar{y}_{1\cdot}) + Var(\bar{y}_{2\cdot}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$,
and
$\frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$.
Replacing σ^2 with s^2:
$\frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$ approximately.
ν is data based.

d.f. for the t distribution

ν is called the Satterthwaite approximation and is computed as

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Will use software for compuation.

CI for $\mu_1 - \mu_2$

A $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

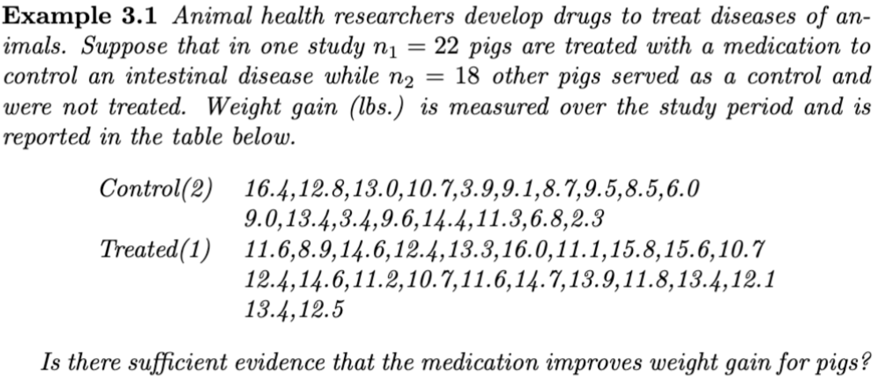
$$(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note the pattern for CI:
point estimate ± multiplier × standard error of the point estimate

t test

$H_0 : \mu_1 - \mu_2 \leq 0, H_a : \mu_1 - \mu_2 > 0.$
$H_0 : \mu_1 - \mu_2 \geq 0, H_a : \mu_1 - \mu_2 < 0.$
$H_0 : \mu_1 - \mu_2 = 0, H_a : \mu_1 \neq \mu_2.$
test statistic
$t = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Example 3.1



Example 3.1 *Animal health researchers develop drugs to treat diseases of animals. Suppose that in one study $n_1 = 22$ pigs are treated with a medication to control an intestinal disease while $n_2 = 18$ other pigs served as a control and were not treated. Weight gain (lbs.) is measured over the study period and is reported in the table below.*

Control(2)	*16.4,12.8,13.0,10.7,3.9,9.1,8.7,9.5,8.5,6.0* *9.0,13.4,3.4,9.6,14.4,11.3,6.8,2.3*
Treated(1)	*11.6,8.9,14.6,12.4,13.3,16.0,11.1,15.8,15.6,10.7* *12.4,14.6,11.2,10.7,11.6,14.7,13.9,11.8,13.4,12.1* *13.4,12.5*

Is there sufficient evidence that the medication improves weight gain for pigs?

Solutions

$H_0 : \mu_1 - \mu_2 = 0$
$H_a : \mu_1 - \mu_2 > 0$

$$t = \frac{\overline{y}_{1\cdot} - \overline{y}_{2\cdot}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{12.83 - 9.38}{\sqrt{\frac{1.87^2}{22} + \frac{3.88^2}{18}}} = \frac{3.48}{0.99} = 3.46$$

Degrees of freedom would be

$$\nu = \frac{\left(\frac{1.87^2}{22} + \frac{3.88^2}{18}\right)^2}{\frac{1}{22-1}\left(\frac{1.87^2}{22}\right)^2 + \frac{1}{18-1}\left(\frac{3.88^2}{18}\right)^2} = 23.4$$

If use t table, round down $\nu = 23$, the p-value is 0.001. There is evidence the weight gain is imporved with the medication at $\alpha = 0.05$.

If use R can use d.f.=23.4. Rcode to get the p-value:
1-pt(3.46,23.4)

CI

Suppose we also want to get a 95% CI for $\mu_1 - \mu_2$:
the t critical value is 2.069 (R code: qt(0.975,23))
and the CI is

$$(12.83 - 9.38) - 2.069(0.99) < \mu_1 - \mu_2 < (12.83 - 9.38) + 2.069(0.99)$$

or

$$3.45 - 2.05 < \mu_1 - \mu_2 < 3.45 + 2.05$$

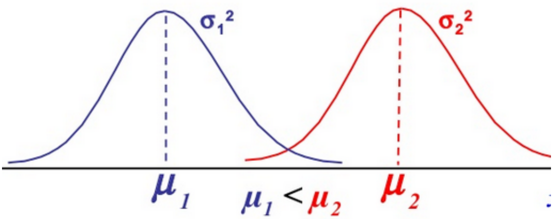
or

$$1.4 < \mu_1 - \mu_2 < 5.5$$

We will mainly use software for computing when we assume $\sigma_1^2 \neq \sigma_2^2$.

Assume equal population variances

Assume equal population variances: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.



Pooled sample variance

Then $Var(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$ and
$\sigma_{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. (1)
We will use the pooled sample variance s_p^2 to estimate σ^2: (before we use s_1^2 to estimate σ_1^2 and use s_2^2 to estimate σ_2^2)

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{n_1-1}{n_1+n_2-2}s_1^2 + \frac{n_2-1}{n_1+n_2-2}s_2^2$$

and $s_p = \sqrt{s_p^2}$.

Note s_p^2 is a weighted average of s_1^2 and s_2^2.
replace σ with s_p in (1), we get the standard error of $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$:
$s_{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Pooled t test

Fact: If the two independent samples are from normal distributions with equal variances, then

$$\frac{(\bar{y}_{1\cdot}-\bar{y}_{2\cdot})-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

CI for $\mu_1 - \mu_2$:
$\bar{y}_{1\cdot} - \bar{y}_{2\cdot} \pm t_{\alpha/2,\nu} \times s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$

Test statistic for $H_0 : \mu_1 - \mu_2 = 0$:
test statistic $t = \frac{\bar{y}_{1\cdot}-\bar{y}_{2\cdot}-0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

The following is a summary of stress limit for specimens using two types of wood.

Wood Type	sample size	sample mean	sample sd
Red Oak	15	8.50	0.80
Fir	15	7.70	1.25

Assume the two population variances are equal.

1. Get a 95% CI for the difference between the average stress limit for red oak and for fir.

2. Test whether the true average stress limit for red oak is different from that for fir at $\alpha = 0.05$.

$df = 28, t_{0.025,28} = 2.048$. (R code qt(0.975,28)).
$s_p^2 = \frac{14s_1^2+14s_2^2}{28} = 0.5 * 0.8^2 + 0.5 * 1.25^2 = 1.10, s_p = 1.05$
CI is $8.50 - 7.70 \pm 2.048 * 1.05\sqrt{\frac{1}{15} + \frac{1}{15}} = (0.01, 1.59)$.

$H_0 : \mu_1 - \mu_2 = 0$ vs $H_a : \mu_1 - \mu_2 \neq 0$.
$t = \frac{8.50-7.70}{1.05\sqrt{\frac{1}{15}+\frac{1}{15}}} = 2.09$.
p-value$= 2P(t > 2.09) = 0.046$.
Reject H_0 at $\alpha = 0.05$.
R code: 2*pt(-2.09,28) or 2* (1-pt(2.09,28))

exercise

Here are data on the number of tree species in 12 unlogged forest plots and 9 similar plots logged 8 years earlier:
unlogged 22 18 22 20 15 21 13 13 19 13 19 15
logged 17 4 18 14 18 15 15 10 12
Assess if logging significantly reduce the mean number of species in a plot after 8 years assuming $\sigma_1^2 = \sigma_2^2$. Use level of significance $\alpha = 0.05$.
Note this is a one-sided test.

```
> unlog = c(22,18,22,20, 15, 21,13,13,19, 13,19,15)
> log = c(17, 4, 18, 14, 18, 15, 15, 10, 12)
> mean1 = mean(unlog) #get sample mean
> mean1
[1] 17.5
> mean2 = mean(log)
> mean2
[1] 13.66667
> s1 = sd(unlog);s2 = sd(log) # sample standard deviation
> s1
[1] 3.5291
> s2
[1] 4.5

> sp2=(11/19)*s1^2+(8/19)*s2^2 #get pooled sample variance
> sp2
[1] 15.73684
```

```
> sp=sqrt(sp2)
> sp
[1] 3.966969    # pooled sample standard deviation

> tvalue=(mean1-mean2)/(sp*sqrt(1/12+1/9))  #test statistic
> tvalue
[1] 2.19139

> pvalue=1-pt(tvalue,19)  # get p-value
> pvalue
[1] 0.02054384
```

Two dependent samples

Two samples are dependent or paired.

Types of pairing:

re-using: each person gets two treatments at different times or occasions. Each person serves as a block.

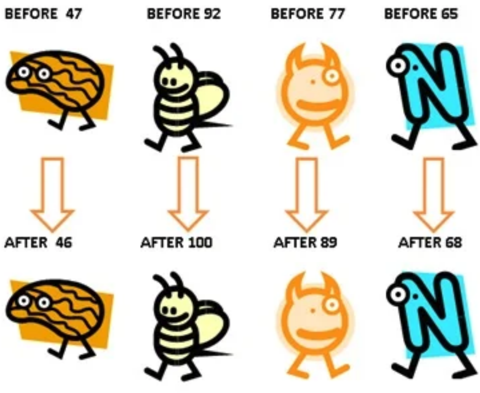
pairing: Two individuals are paired according to some extraneous variables. Two persons in each pair are randomly assigned to two treatments. Each pair of individuals serve as a block.

splitting: some material is split in half and the two halves are randomly assigned to two treatments. The two halves serve as a block.

n pairs: $(y_{11}, y_{21}) \cdots, (y_{1n}, y_{2n})$.

The two observations within each pair tend to be correlated.

Paired samples



Analysis: One sample t test using d values

Take difference of each pair: $d_1 = y_{11} - y_{21}, \cdots, d_n = y_{1n} - y_{2n}$.
Purpose: examine $\mu_d = \mu_1 - \mu_2$.

Assume $d_1, \cdots, d_n \sim N(\mu_d, \sigma_d^2)$, then a CI for μ_d is

$\bar{d} \pm t_{\alpha/2,n-1} \frac{s_d}{\sqrt{n}}$

Test $H_0 : \mu_d = \delta$

test statistic $t = \frac{\bar{d}-\delta}{s_d/\sqrt{n}}$

Connection between CI and hypothesis test

For a two sided test of significance level α and a $100(1-\alpha)\%$ CI:
If the test rejects $H_0 : \mu_1 = \mu_2$, then the CI does not include 0.
if the test does not reject H_0, then the CI contains 0.

paired samples

Each of six subjects were given both diets (diet 1: high fat, diet 2: high carbodydrates) in a random order. Each diet lasted 3 days. At the end of the 3 day period, the subject was put on the treadmill and the time to exhaustion, in seconds, was measured.

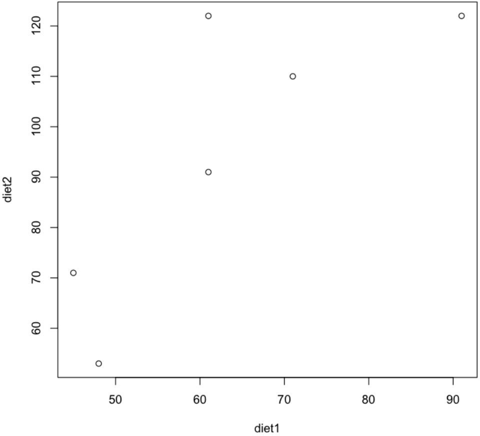
Subject	Diet1	Diet2	difference(d)
1	91	122	31
2	48	53	5
3	71	110	39
4	45	71	26
5	61	91	30
6	61	122	61
mean:	62.83	94.83	32

(standard deviation of d is 18.221)

Get a 95% confidence interval for the mean difference in treadmill time between the two diets. Also perform a t test to examine if the two diets produce the same mean treadmill time at $\alpha = 0.05$.

Paired samples tend to be correlated

This is a paired samples design. Each subject is used twice.
The two samples are postively correlated. The two values produced by the same subject tend to be bigger or smaller together.



sample correlation $=0.79$

Paired t test

$\bar{d} \pm t\frac{s_d}{\sqrt{n}} = 32 \pm 2.571 * 18.221/\sqrt{6} = 32 \pm 19.12 = (12.88, 51.12).$
$H_0 : \mu_d = 0$ or $\mu_1 = \mu_2$
$H_1 : \mu_d \neq 0$ or $\mu_1 \neq \mu_2$
$t = \frac{32-0}{\frac{18.221}{\sqrt{6}}} = 4.30$
p-value $= 2P(t > 4.30) = 0.008$
Reject H_0.

```
> qt(0.975,5) # t critical value
[1] 2.570582
> 2*pt(-4.30,5) # two tailed p-value
[1] 0.007715449
```

Solution to prob 3.2

This is a paired samples design. A pair of candles burned on the same day serve as a block.
$\bar{d} = 25, s_d = 58.21, df = 9$
A 95% CI is:
$\bar{d} \pm t * s_d/\sqrt{n} = 25 \pm 2.262 * 58.25/\sqrt{10} = (-16.64, 66.64)$.
Test $H_0 : \mu_1 = \mu_2(\mu_d = 0)$, vs $H_a : \mu_1 \neq \mu_2(\mu_d \neq 0)$.
$t = \frac{25-0}{58.21/\sqrt{10}} = 1.358$.
p-value$=2P(t > 1.358) = 0.2075$.
Fail to reject H_0. There is no evidence in the data that the mean burning time of scented and unscented candles differ.

R code

Two sample t test

```
t.test(x,y)  # two sample t test unequal variance

t.test(x,y,alternative="greater") #right sided test

t.test(x,y,alternative="less")  # left sided test

t.test(x,y,var.equal=TRUE) #pooled t test

t.test(x,y,mu=10,paired=TRUE)   #paired t test,
                           #null difference=10
```

Two independent samples with unequal population variances

```
> unlog = c(22,18,22,20, 15, 21,13,13,19, 13,19,15)
> log = c(17, 4, 18, 14, 18, 15, 15, 10, 12)
> t.test(unlog, log)
Welch Two Sample t-test
data:  unlog and log
t = 2.1141, df = 14.793, p-value = 0.05192
alternative hypothesis: true difference in means is not
                        equal to 0
95 percent confidence interval:
 -0.03622486  7.70289153
sample estimates:
mean of x mean of y
 17.50000  13.66667
```

Two independent samples: pooled t test

```
> t.test(unlog,log,var.equal=TRUE)
Two Sample t-test
data:  unlog and log
t = 2.1914, df = 19, p-value = 0.04109
alternative hypothesis: true difference in means
                        is not equal to 0
95 percent confidence interval:
 0.1720715 7.4945951
sample estimates:
mean of x mean of y
 17.50000  13.66667
```

Paried t test

```
> diet1=c(91,48,71,45,61,61)
> diet2=c(122,53,110,71,91,122)
> plot(diet1,diet2) #plot the two samples
> cor(diet1,diet2) #correlation between the two samples
[1] 0.7942231
> t.test(diet1,diet2,paired=TRUE)

Paired t-test

data:  diet1 and diet2
t = -4.3019, df = 5, p-value = 0.007702
alternative hypothesis: true difference in means is
                        not equal to 0
95 percent confidence interval:
 -51.12163 -12.87837
sample estimates:
mean of the differences
                    -32
```

Power of the test and the needed sample size

$y_{11}, \cdots, y_{1n_1} \sim N(\mu_1, \sigma_1^2)$,
$y_{21}, \cdots, y_{2n_2} \sim N(\mu_2, \sigma_2^2)$,
Assume $n_1 = n_2 = n = 25, \sigma_1 = \sigma_2 = \sigma = 1$.
Test $H_0 : \mu_1 - \mu_2 = 0$ vs $H_a : \mu_1 - \mu_2 = 1.2$.
$\alpha = 0.05$.
Reject H_0 if $\frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot} - 0}{\sigma\sqrt{1/n + 1/n}} > 1.645$.
or if $\bar{y}_{1\cdot} - \bar{y}_{2\cdot} > 1.645\sigma\sqrt{\frac{2}{n}}$.
power$= P(\bar{y}_{1\cdot} - \bar{y}_{2\cdot} > 1.645\sigma\sqrt{\frac{2}{n}}; \mu_1 - \mu_2 = 1.2)$
$= P(z > \frac{1.645\sigma\sqrt{2/n} - 1.2}{\sigma\sqrt{2/n}}) = P(z > -2.60) = 0.995$.
What common sample size is needed if $H_a : \mu_1 - \mu_2 = 1$ and we want $1 - \beta \geq 0.80, \alpha \leq 0.05$?