11/15/18
Wave
Moretto

*On HW 4

Lecture #15: 2 Independent Samples

- Sokal and Rohlf (1995) + R-22
- Analysis of 2 independent samples
- The diff does not seem large in practical terms (62 min diff in 2 weeks)
- Is this diff. large in statistical terms?

Genetic Cross I

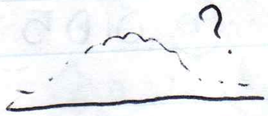
pop 1

All F clones from genetic cross

Time —

$N_1 = ?$ (BIG) $[\,\int\,]$ Like SRS = IID

mean $\mu_1 = ?$
SD $\sigma_1 = ?$



?

pop hist.

sample 1

I The obs. insect

mean $\bar{y} = 7.5143$
SD $s_1 = 0.7104$

$[\,\int\,]\ n_1 = ?$

sample hist. 1

$[\,\int\,]\ n_1 = ?$

mean $y_1 = ?$ (ex 7.58)

2 indp. samples

Imag. data

All possible $\bar{y}$s

$\begin{bmatrix} 7.5143 \\ 7.58 \\ \vdots \end{bmatrix}$ $M_1 = \infty$

Long run mean $E_{IID}(\bar{y}_1) = \mu_1$

Est. long run SD $\widehat{SE}_{IID}(\bar{y}) = \frac{s_1}{\sqrt{n_1}}$

long run hist

$\boxed{1}$

~~[illegible]~~

Genetic Cross II

Pop 2 $\boxed{\text{Ditto}}$

Time-

$N_2 = ?$ [∫]

Like SRS = IID

mean $\mu_2 = ?$

SD $\sigma_2 = ?$

pop. hist

sample 2 $\boxed{\text{Ditto}}$

Time-

[ ] $n_1 = ?$

mean $\bar{y}_2 = 7.5571$

SD $s_2 = 0.6399$

[∫] $n_2 = ?$

sample hist ②

imag. data 2 $\boxed{\text{possible } \bar{y}\text{s}}$

Time-

7.5571
7.54
⋮

$M_2 = \infty$

Long run mean $E_{IID}(\bar{y}_2) = \mu_2$

Est long run SD $\hat{SE}_{IID}(\bar{y}_2) = \frac{s_2}{\sqrt{n_2}}$

Long run hist ②

Inferential Summary

Ⓟ	Unknown quantity of interest	$(\mu_2-\mu_1)$ = pop mean diff in time to reproduce between gen I + II
Ⓢ	Estimate	$(\bar{y}_2-\bar{y}_1) = 7.5571 - 7.5143 = +0.0428$ days
Imag. data	Give or take for $(\bar{y}_2-\bar{y}_1)$ as $(\mu_2-\mu_1)$ est.	$\hat{SE}(\bar{y}_2-\bar{y}_1) = 0.3614$ days
	95% CI for $(\mu_2-\mu_1)$	$(\bar{y}_2-\bar{y}_1) \pm \underbrace{(2.179)}_{t\ 95\%,\ n_1+n_2-2} \hat{SE}(\bar{y}_2-\bar{y}_1)$

Formula 4

$\hat{SE}_{IID}(\bar{y}_1) = \frac{s_1}{\sqrt{n_1}} = \frac{0.7104 \text{ days}}{\sqrt{7}} = 0.2685$ days

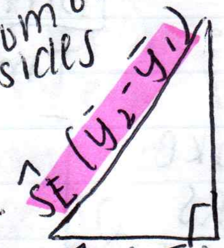
$\hat{SE}_{IID}(\bar{y}_2) = \frac{s_2}{\sqrt{n_2}} = \frac{0.6399 \text{ days}}{\sqrt{7}} = 0.2419$ days

• **Math fact:**

(pythagorus)

- answer must be ≥ 0.2686 ≤ 0.81

sum of sides



$SE(\bar{y}_2)$
0.2419 days

$\hat{SE}(\bar{y}_1)$
0.2685 days

$\sqrt{(0.2685\text{days})^2 + (0.2419\text{ days})^2}$

$= 0.3614$ days

$\hat{SE}(\bar{y}_1)$

$= \dfrac{s_1}{\sqrt{n_1}}$

$\hat{SE}(\bar{y}_2) = \dfrac{s_2}{\sqrt{n_2}}$

$\hat{SE}(\bar{y}_2 - \bar{y}_1)$

$$= \sqrt{[\hat{SE}(\bar{y}_2)]^2 + [\hat{SE}(\bar{y}_1)]^2}$$

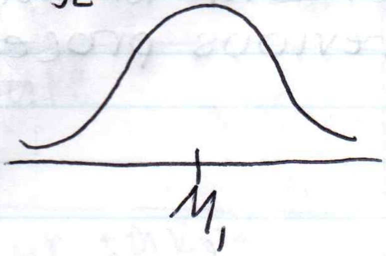
$$= \sqrt{\left(\frac{s_2}{\sqrt{n_2}}\right)^2 + \left(\frac{s_1}{\sqrt{n_1}}\right)^2}$$

R-24 formula!!

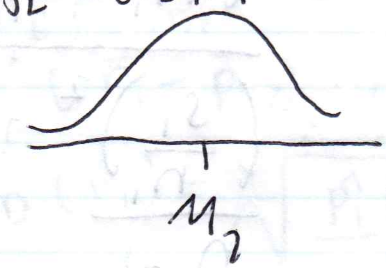
$$\boxed{= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \hat{SE}(\bar{y}_2 - \bar{y}_1)}$$

Long run histograms:

$\boxed{1}$ $\widehat{SE} = 0.2605$ $t_{n_1 - 1} = t_6$

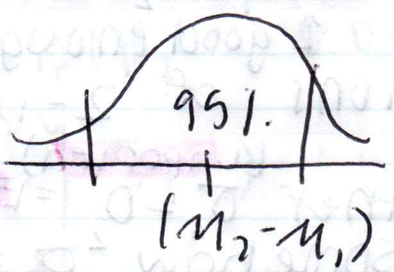


$\boxed{2}$ $\widehat{SE} = 0.2419$ $t_{n_2 - 1} = t_6$

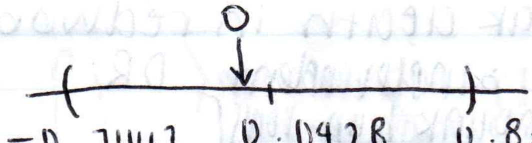


$t_{n_1 + n_2 - 2}$ → 12 degrees freedom

add t curves together



=> 0.0428 ± 0.7875



null: $\mu_2 - \mu_1 = 0$

Devil's advocate

NOT STAT SIG

- Squeaky clean formula for degrees of freedom of previous procedure is (zar, p.129):

(implemented by JMP)

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

$= n_1 + n_2 - 2$

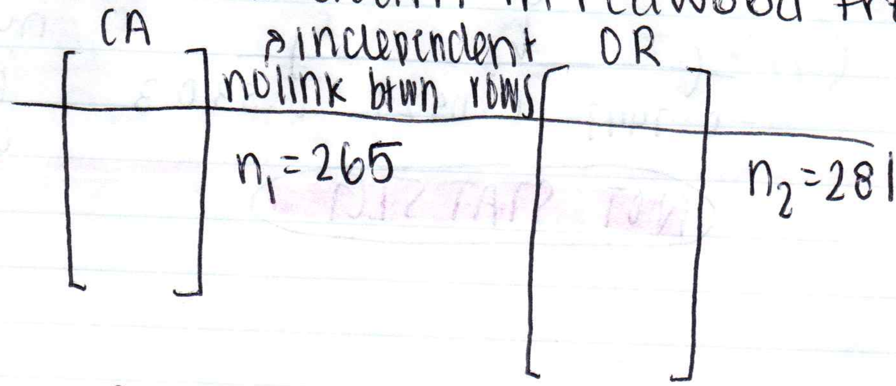
↑ good enough

- Actually 2 versions of 2-independent samples story:

Ⓐ Not sure whether $\sigma_1 = \sigma_2$ = Equal variances (unpooled)

Ⓑ Know ~~[illegible]~~ somehow $\sigma_1 = \sigma_2$ (pooled) = un equal variances

- 2 independent samples: dichotomous outcomes
  - sudden oak death in redwood trees



- Need 2 model diagrams

inferential summary

(P) unknown Q. interest	$(p_1 - p_2)$ = pop mean diff rates in OR (A)
(S) estimate	$(\hat{p}_1 - \hat{p}_2) = 3.4\% - 7.1\%$ $= -3.7\%$
Give or take	$\hat{SE}(\hat{p}_1 - \hat{p}_2) = 1.9\%$
95% CI for $(\bar{p}_1 - \bar{p}_2)$	$(-2.5\%, 0.1\%)$

imag. data

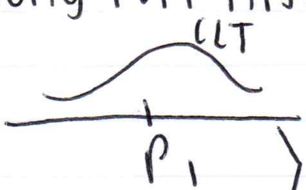
$$\hat{SE}_{IID}(\hat{p}_1) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = \sqrt{\frac{(0.034)(0.966)}{265}}$$

$= 0.0111 = 1.1\%$

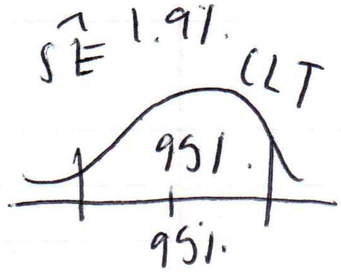
$\hat{SE}_{IID}(\hat{p}_2) = 0.0153 = 1.5\%$

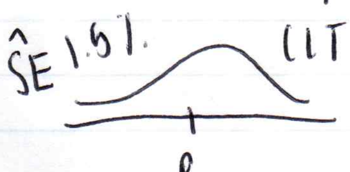
$\hat{SE}(p_2) = 1.5\%$

Long-run histograms

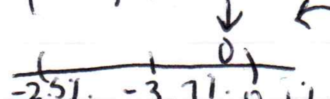


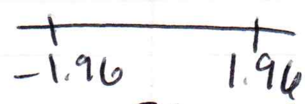
} normal curves ⇒





Not stat sig at 95%. So get more good data





95% = −3.7% ± 2(1.9%)

$$\boxed{\hat{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

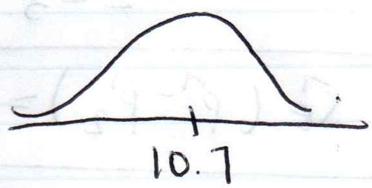
New Unit: Simple Correlation and Simple Linear Regression

- Study from Zar → sage sparrow

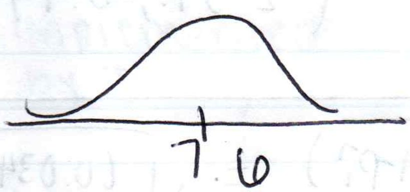
Y = Tail length (cm)
X = wing length (cm)

Y	X
7.4	10.4
7.6	10.3
⋮	⋮

$n = 12$



hist of X



hist of y

mean $\bar{y} = 7.6$ cm $\quad \bar{x} = 10.7$ cm

SD $s_y = 0.35$ cm $\quad s_x = 0.48$ cm

- Positively associated variables

scatterplot