This time: two independent samples (dichotomous outcomes)

Next time: correlation & regression

Read: LN pp. L-214 - L-224 Today: LN p. L-200 →

redwood example:

$\frac{9}{265} < \frac{20}{281}$ large in practical terms (2x as many)

Important questions to ask:

1. quant., qual., dichotomous? dichotomous
2. how many samples (1, 2, or more)? 2 samples
3. are they linked/matched (ex: pairs, repeated measures)? 2 independant samples

need to make 2 models (more in LN p. 201-202)

Case Study: Sudden Oak Death

where	sample	$\hat{p}$
CA	1	$\frac{9}{265} = 3.4\%$
OR	2	$\frac{20}{281} = 7.1\%$

Q: Practsig?

A: Overwhelmingly so

relative comparison: $\frac{7.1\% - 3.4\%}{3.4\%} = \frac{3.7\%}{3.4\%}$ (cancel out) $= 1.09 = 109\%$

sudden oak death rate in Oregon is 109% bigger than s.o.d. rate in California

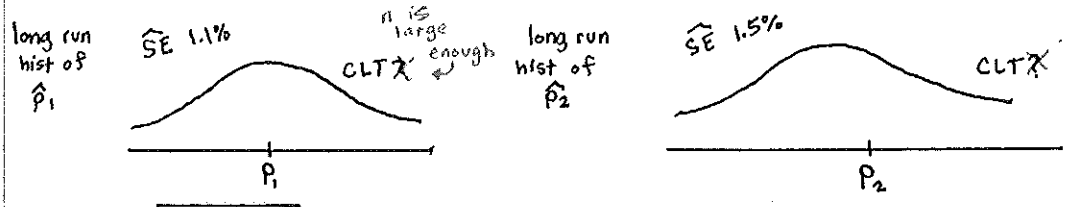
Tip: don't forget to distinguish your models as sample① or data②

make inf. summary for the comparison:

unknown Q of interest	$(p_1 - p_2)$
estimate	$(\hat{p}_1 - \hat{p}_2)$
give of take for $\bar{y}$ as est of μ_y	$\widehat{SE}(\hat{p}_1 - \hat{p}_2)$
95% CI	$(\hat{p}_1 - \hat{p}_2) \pm 2\widehat{SE}(\hat{p}_1 - \hat{p}_2)$

EV of $\hat{p}_1$ = $E_{SRD}(\hat{p}_1) = p_1$

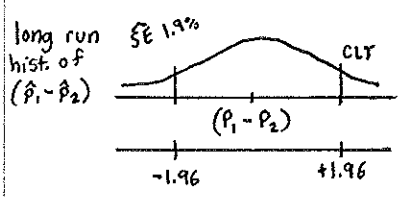
$E_{SRD}(\hat{p}_2) = p_2$



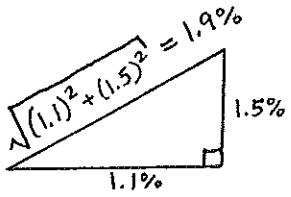
$\widehat{SE}(\hat{p}_1) = \sqrt{\frac{p_1(1-p_1)}{n_1}} = \frac{6}{\sqrt{n}}$ ← $\sqrt{p_1(1-p_1)}$

$= \sqrt{\frac{(0.34)(0.966)}{265}} = 0.0111 \doteq 1.1\%$

$$\widehat{SE}(\hat{p}_2) = \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{(0.071)(0.929)}{281}} = 0.0153 \doteq 1.5\%$$



$SE(\hat{p}_1 - \hat{p}_2) = ?$



$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{(SE(\hat{p}_1))^2 + (SE(\hat{p}_2))^2}$$

$$= \sqrt{\left(\sqrt{\frac{p_1(1-p_1)}{n_1}}\right)^2 + \left(\sqrt{\frac{p_2(1-p_2)}{n_2}}\right)^2}$$

formulas 13 & 14 on pg. 28

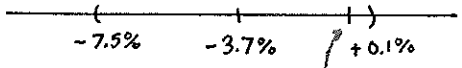
$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

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

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\, \widehat{SE}(\hat{p}_1 - \hat{p}_2)$$

$$-3.7\% \pm (2)(1.9\%) = -3.7\% \pm 3.8\%$$

95% CI for $(p_1 - p_2)$



0% is in 95% CI, so not strictly speaking statsig, but evidence is trending toward the conclusion that this diff. is real

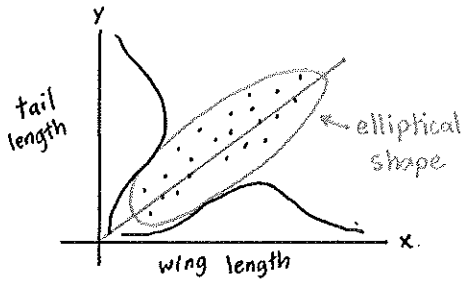
statsig?

is 0 in the interval? Yes, strictly speaking it's not statsig (but is close)

↑ boring devil's advocate

Ending on pg. 213: extra notes, includes discussion question answers

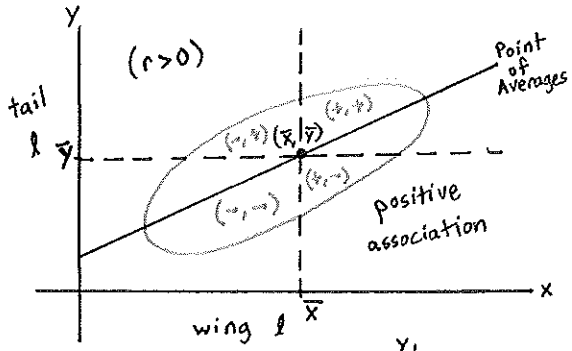
Unit 6: Simple Correlation & Simple Linear Regression

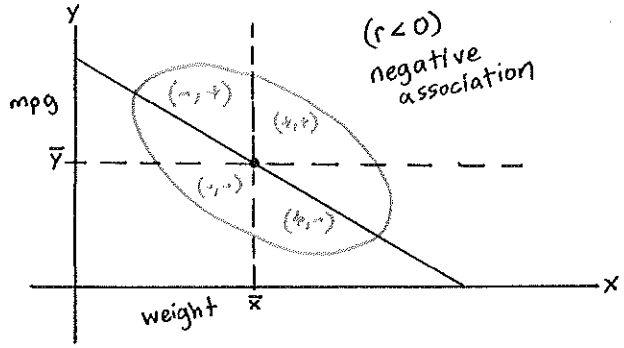


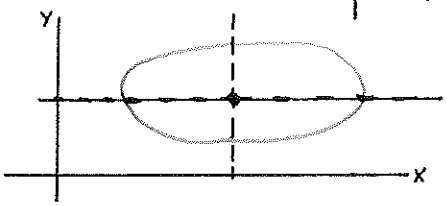
Scatterplot
(x, y) bivariate (2 variables)

use x to predict y

cars







(r=0)
no (linear) association

Karl Pearson (1890)

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_x^*}\right)\cdot\left(\frac{y_i-\bar{y}}{S_y^*}\right)$$

r = correlation coefficient

outcome = dependent variable

predictor = indep. var.

$$\begin{bmatrix} y_1 & x_1 \\ y_2 & x_2 \\ \vdots & \vdots \\ y_n & x_n \end{bmatrix}$$

mean $\bar{y}$ $\bar{x}$

SD s_y s_x

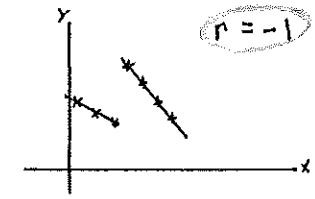
$$S_x^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \& \quad S_y^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

(not n-1)

facts about r

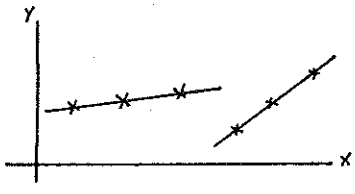
① r is a pure number without units

② $-1 \le r \le +1$



r = −1

all points lie on strait line w/neg. slope



r = +1

all points lie on strait line w/positive slope

O has 3 diff. ways to unfold. We'll talk about them next time.