AMS 7
10 Nov 15
①

this time: two independent samples (dichotomous outcomes)

read: LN pp. L-214 – L-244

next time: correlation & regression

today: LN p. L-200 →

if you have a ~~wed~~ disc sec., this week only see course web page about which section to go to (tomorrow is a holiday)

case study: sudden oak death (s.o.d.)

Q: practsig?

A: overwhelmingly so

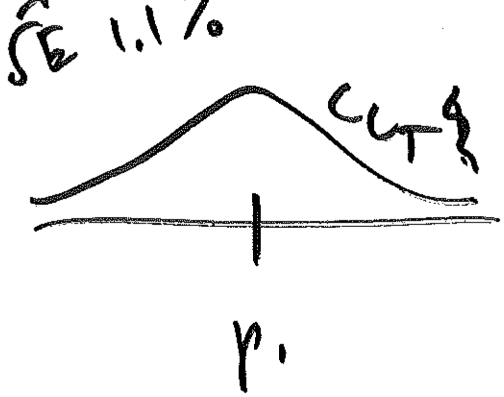
where	sample	$\hat{p}$
CA	1	$9/265 = 3.4\%$
OR	2	$20/281 = 7.1\%$

$$\frac{7.1\% - 3.4\%}{3.4\%} = \frac{3.7 \cancel{\%}}{3.4 \cancel{\%}} = 1.09 = 109\%$$

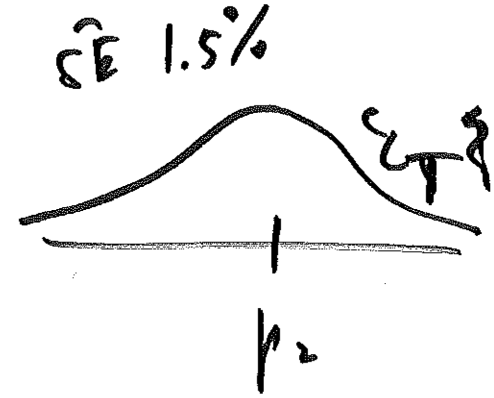
s.o.d. rate in OR is 109% bigger than s.o.d. rate in CA

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

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



long run hist of $\hat{p}_1$



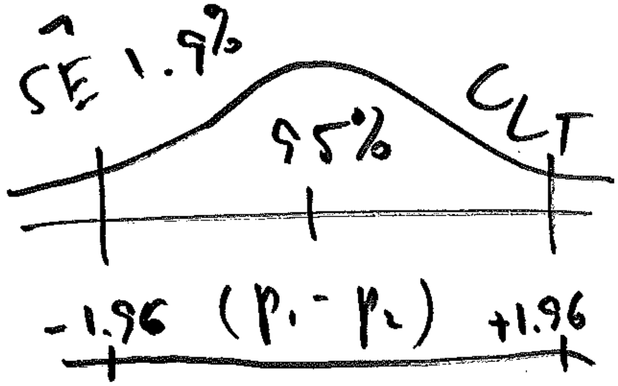
long run hist of $\hat{p}_2$

$$\hat{SE}(\hat{p}_1) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = \frac{\sigma}{\sqrt{n}} \qquad \sigma \leftarrow \sqrt{p_1(1-p_1)}$$

$$= \sqrt{\frac{(.034)(.966)}{265}} = .0111 \doteq 1.1\%$$

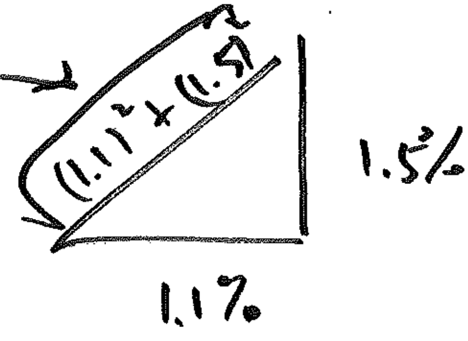
$$\hat{SE}(\hat{p}_2) = \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{(.071)(.929)}{281}}$$

$$\doteq .0153 \doteq 1.5\%$$



long-run hist of $(\hat{p}_1 - \hat{p}_2)$

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



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

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

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

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

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

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

95% CI for $(p_1 - p_2)$

$$-7.5\% \quad -3.7\% \quad +0.1\%$$

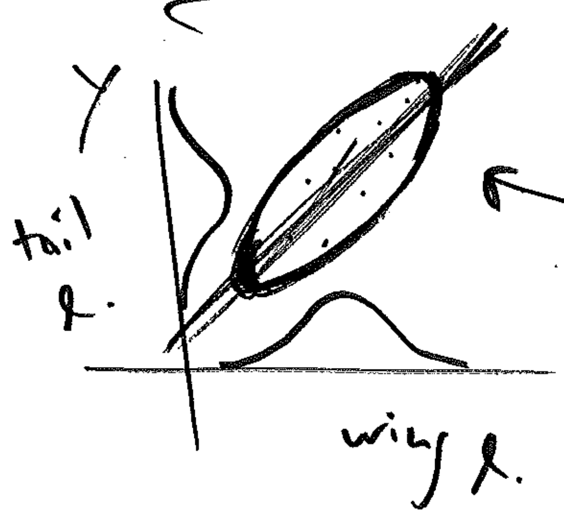
0% is in 95% CI,

so not strictly speaking

statsig, but

evidence tending toward

conclusion that this diff. is real

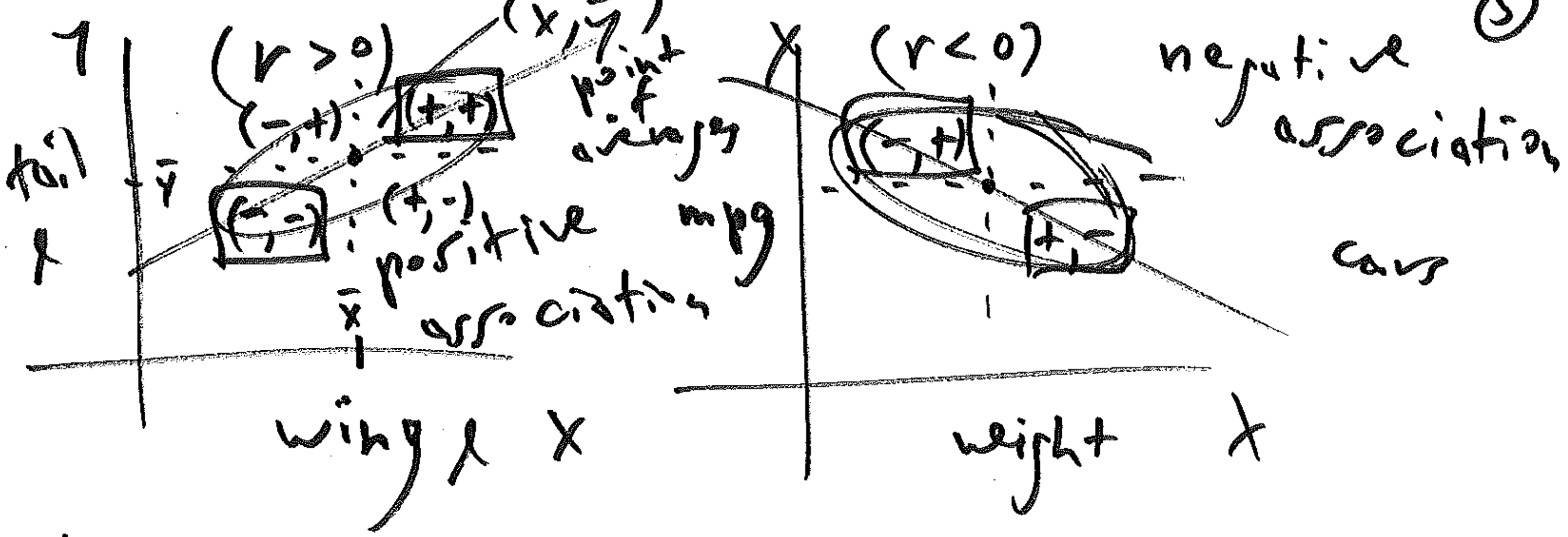


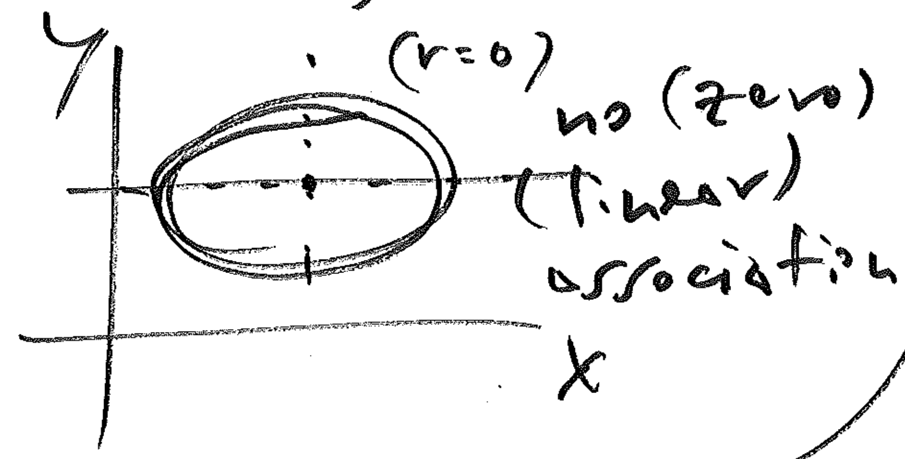
Scatterplot

elliptical shape

(X, Y) bivariate

use x to predict y





Karl Pearson (1890)

$$
\begin{array}{cccc}
\text{outcome} = \text{dependent var.} & y_1 & x_1 & \text{predictor} = \text{indep. var.} \\
 & y_2 & x_2 & \\
 & \vdots & & \\
 & y_n & x_n & \\
\text{mean} & \bar{y} & \bar{x} & \\
\text{SD} & s_y & s_x &
\end{array}
$$

$\boxed{\text{correlation}}$ (coefficient)

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

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

facts about r

1. r is a pure number without units

2. $-1 \le r \le +1$