AMS 7
10 Nov 09

this time: 2-independent samples with 0s + 1s

next time: correlation + regression

read: lecture notes pp. L-214 + L-231

1

---

lab 5 due by 5pm Mon 16 Nov in box outside my door

---

midterm solutions posted in glass case later today

---

no discussion sections, labs or office hours tomorrow (holiday); if you usually go to section and/or lab on Wed you need to go to another section and/or lab this week (e.g. today @ noon)

Thu 11 Am lab: apologies for mix up; running again as usual starting this week (Val TA)

②

(CA)

pop$_1$: all redwood trees in chosen grove

sample$_1$: the observed ~~redwood~~ trees

imag. data$_1$: all possible $\hat{p}_1$'s

1 = Y
0 = N

infected?

N_1 = thousands

$\begin{pmatrix} 1\text{s} \\ \& \\ 0\text{s} \end{pmatrix}$

(actual) SRS ~~+~~ $\doteq$ IID

infected?

$\begin{pmatrix} 1\text{s} \\ \& \\ 0\text{s} \end{pmatrix}$ $n_1 = 265$

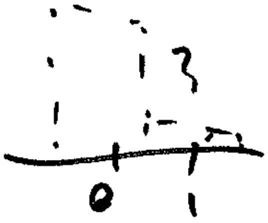
$\begin{pmatrix} 3.4\% \\ 3.6\% \\ \vdots \end{pmatrix}$ $M_1 = \infty$

mean $\hat{p}_1 = \frac{9}{265} \doteq 3.4\%$

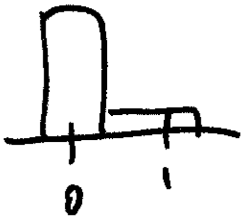
mean p_1 = ?

SD $\sigma_1 = \sqrt{p_1(1-p_1)}$ = ?

hyp. IID



pop. hist



sample hist

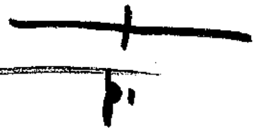
long run mean $\boxed{1}$ EV of $\hat{p}_1 = p_1$

est. long run SD $\boxed{2}$ $\hat{SE}$ of $\hat{p}_1 = 1.1\%$

$\begin{pmatrix} 1\text{s} \\ \& \\ 0\text{s} \end{pmatrix}$ $n_1 = 265$

mean $\hat{p}_1$ = ? (ex. 3.6%)

long run hist $\boxed{3}$



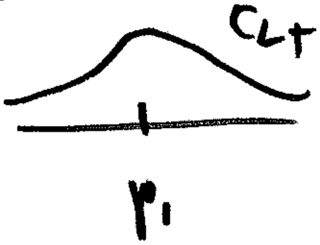
$\boxed{1}$ EV of $\hat{p}_1 = E_{IID}(\hat{p}_1) = p_1$

$\boxed{2}$ $\hat{SE}_{IID}(\hat{p}_1) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$

$= \sqrt{\frac{(.034)(.966)}{265}} = 0.011$

$\doteq 1.1\%$

$\boxed{3}$ $\hat{SE}$ 1.1%



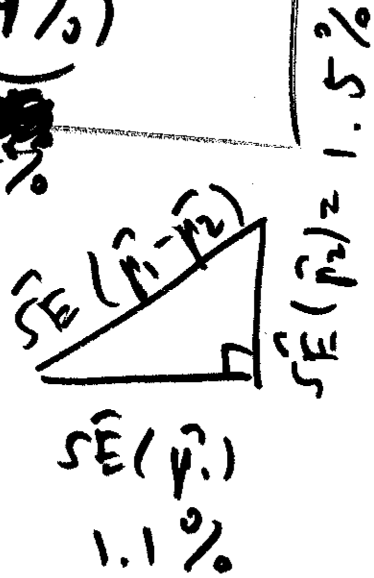
long run hist of $\hat{p}_1$

inferential summary

unknown pop. quantity of interest	$(p_1 - p_2)$ = pop. diff in % infection of (CA − OR)
estimate of $(p_1 - p_2)$	$\hat{p}_1 - \hat{p}_2 = 3.4\% - 7.1\% = -3.7\%$
give or take for $(\hat{p}_1 - \hat{p}_2)$ as est. of $(p_1 - p_2)$	$\hat{SE}(\hat{p}_1 - \hat{p}_2) = 1.9\%$
95% interval for $(p_1 - p_2)$	$-3.7\% \pm \underbrace{(1.96)(1.9\%)}_{3.8\%}$

$$SE_{indep}(\hat{p}_1 - \hat{p}_2) = \text{Pythagoras again}$$

$\bar{Y}_1 \nearrow \hat{p}_1$, $\bar{Y}_2 \nearrow \hat{p}_2$



$$= \sqrt{(1.1\%)^2 + (1.5\%)^2} \doteq 1.9\%$$

$$= \sqrt{(.011)^2 + (.015)^2} = .019$$

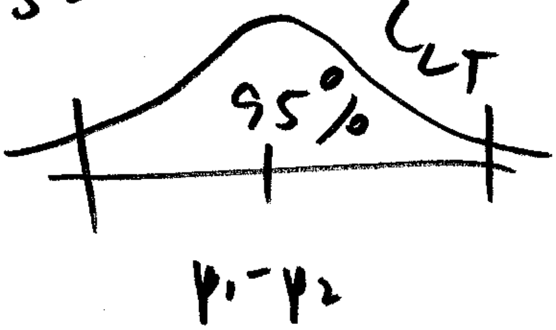
$$\hat{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\left[\hat{SE}(\hat{p}_1)\right]^2 + \left[\hat{SE}(\hat{p}_2)\right]^2}$$

$$= \sqrt{\left(\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}\right)^2 + \left(\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)^2}$$

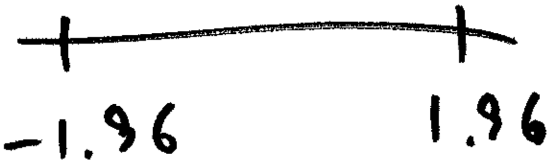
$$= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$ math fact

---

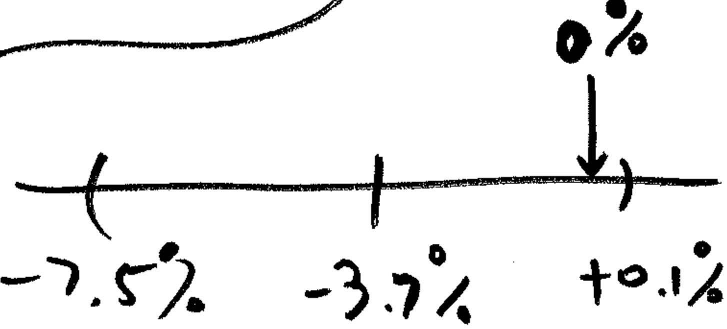
$\hat{SE}$ 1.9%



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



(barely) not stat sig



fix: just get a few more trees from each grove

	Y	X
1	7.4	10.4
2	7.6	10.8
	⋮	⋮
	8.3	11.4
mean	$\bar{Y} = 7.6$ cm	$\bar{X} = 10.7$ cm
SD	$s_y = 0.35$ cm	$s_x = 0.4$ cm
hist		

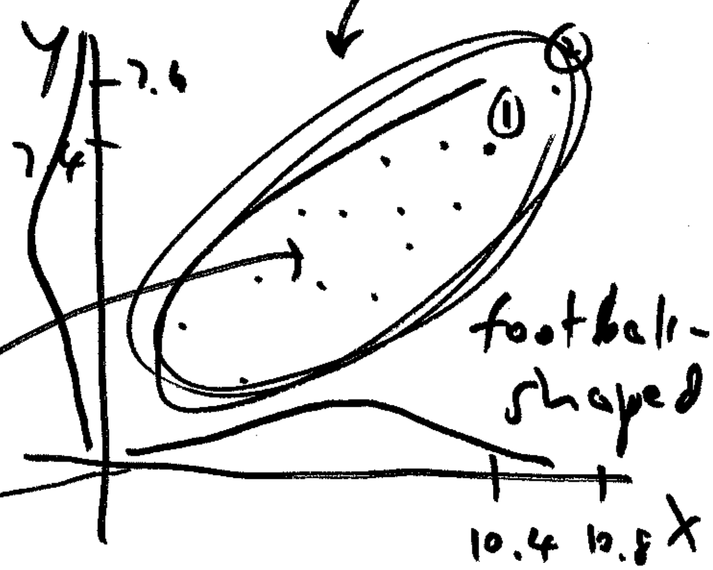
outcome → Y = tail length (cm)

X = wing length (cm)

↑ predictor

$n = 12$

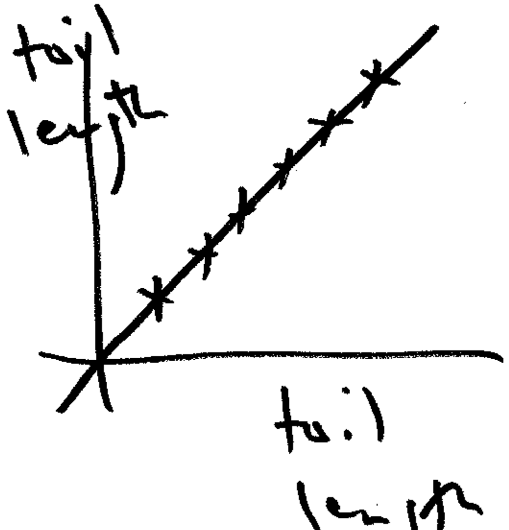
elliptical

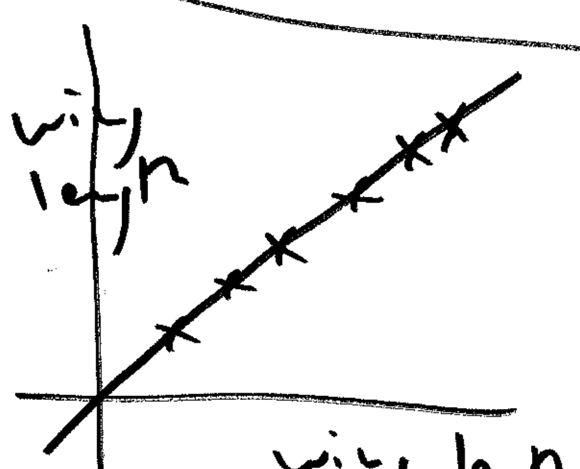


scatterplot or scatter diagram

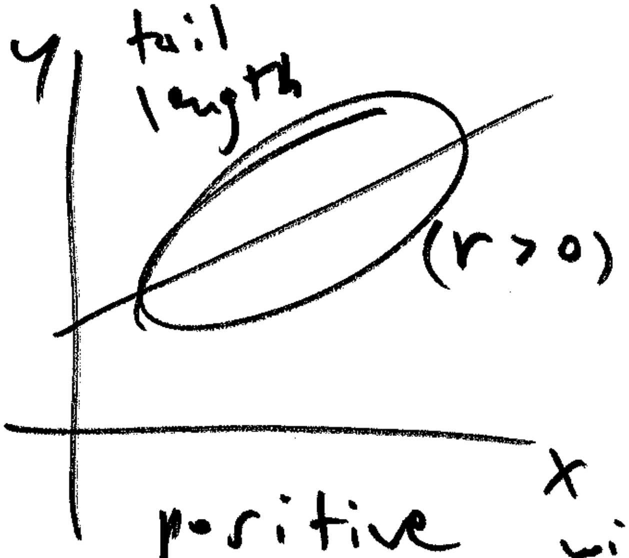
bivariate normal distribution

(elliptical scatterplot)

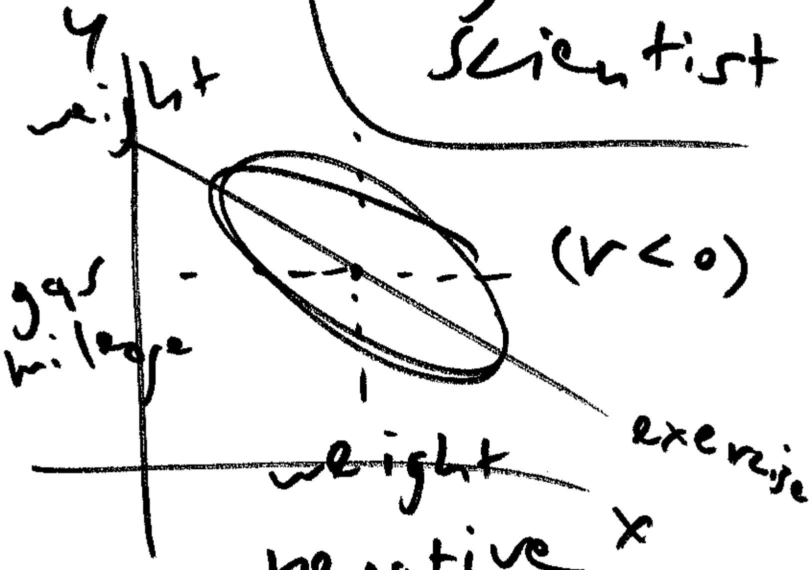




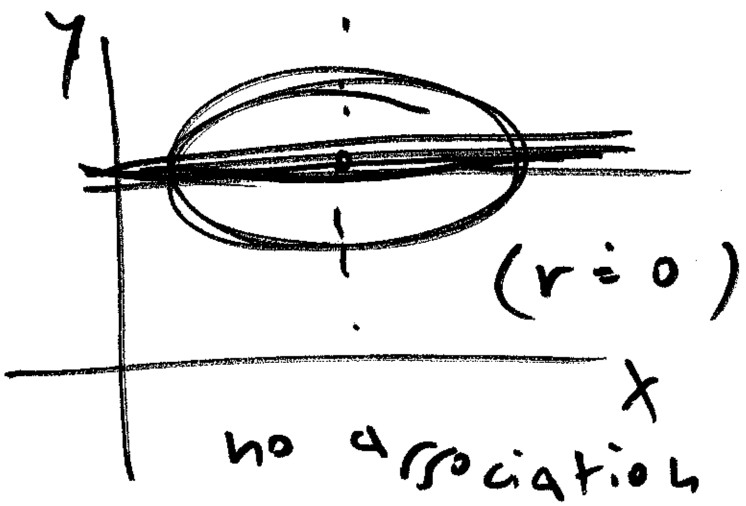
Karl Pearson (1880) German + English Scientist



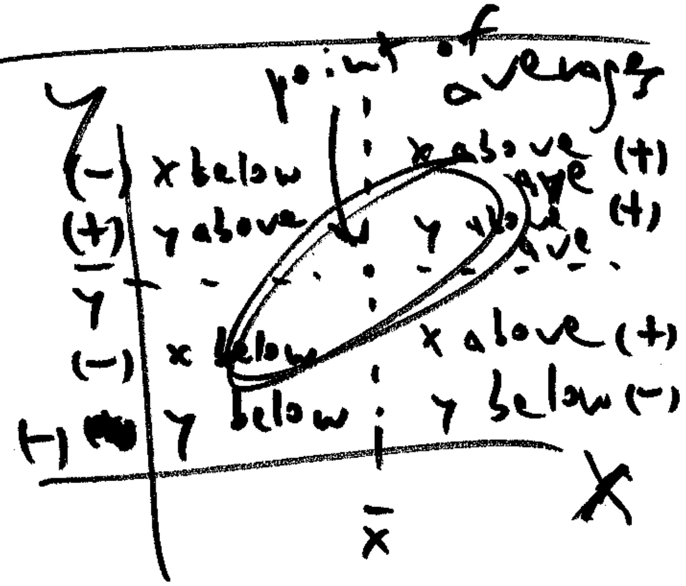
positive association



negative association



no association



(pearson's product-moment) $\boxed{\text{correlation}}$ (coefficient)

y_1	x_1
y_2	x_2
$\vdots$	$\vdots$
y_n	x_n

n

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

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

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