

read:
LN pp. L-214
→ L-282
today:
LN pp L-214
→ L-221

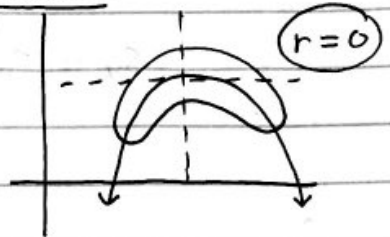
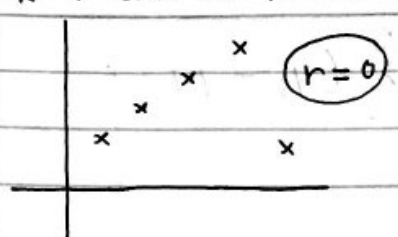
this time: correlation & regression
next time: ANOVA

* HW 3 due by
Sun 27 May 18
on canvas

Ashley Tai
AMS 7
24 May 18

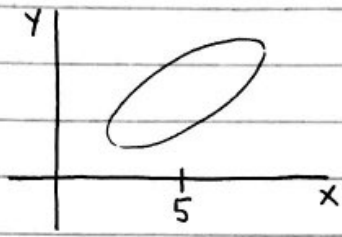
correlation coefficient (r)

* r can be fooled by outliers

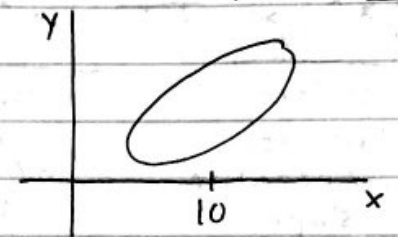


* especially
when n is
small

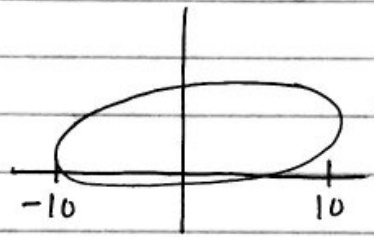
* if a constant is added to y or x values, r is unchanged



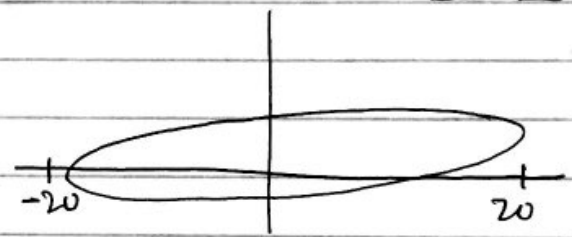
add
5 in
x



* if a constant is multiplied to y or x values, r is changed.



multiply
x by 2



* $r(x, y) = r(y, x)$

* $r(x, x) = 1$

Devil's Advocate: the only reason we got $r \neq 0$ is because of
unlucky random sampling
- must take seriously

Inference with r

population
all relevant birds of this species

* similar, in all relevant ways

sample
the observed birds

imaginary data
all possible values of r

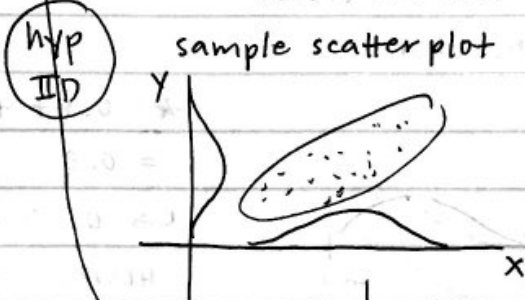
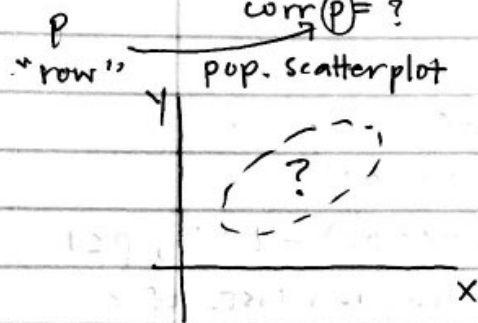
tail wing
 $N = ?$
 (big)
 y x
 mean $\mu_y = ?$ $\mu_x = ?$
 SD $\sigma_y = ?$ $\sigma_x = ?$

actual
like
SRS
= IID

tail wing
 y_1 x_1
 y_2 x_2
 \vdots \vdots
 y_n x_n
 $n = 12$
 mean $\bar{y} = 7.6$ $\bar{x} = 10.7$ cm
 SD $s_y = 0.35$ $s_x = 0.40$ cm

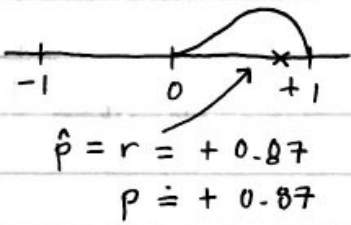
$\begin{bmatrix} 0.87 \\ 0.82 \\ \vdots \end{bmatrix}$
 $M \rightarrow \infty$

long run mean
 $E_{IID}(r) \doteq p$
 est. long run SD
 $\hat{SE}_{IID}(r) = 0.16$
 long run hist.



hyp IID

y_1 x_1
 y_2 x_2
 \vdots \vdots
 y_n x_n
 $n = 12$
 ex. corr $r = +0.82$



Inferential Summary

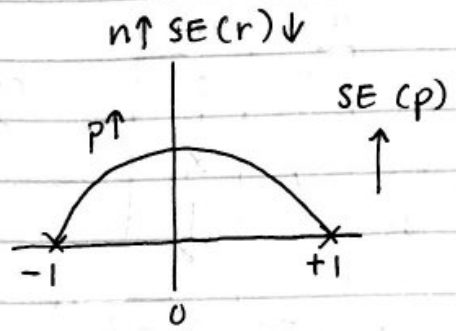
pop	unknown pop. quantity of main interest	$p =$ pop. correlation between wing & tail length of this species
sample	estimate of p	$r = +0.87$
↑ imag data	give or take for r as est. of p	$\hat{SE}(r) = 0.16$
↓	95% CI for p	approx: (0.55, 1.0) exact: (0.59, 0.96)

facts:

- ① $E_{\text{IID}}(r) \doteq p$
- ② $SE_{\text{IID}}(r) = \frac{1-p^2}{\sqrt{n-3}}$

$$\hat{SE}_{\text{IID}}(r) = \frac{1-r^2}{\sqrt{n-3}}$$

here, $\hat{SE}_{\text{IID}}(r) = \frac{1-(0.8)^2}{\sqrt{12-3}} = 0.081$



③ long run hist. of r

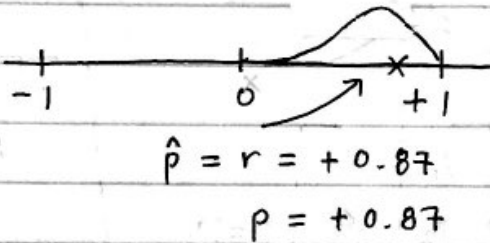
★ $r \pm 2\hat{SE}(r)$

$= 0.87 \pm 2(0.08)$

↳ $0.87 + 2(0.08) \doteq 1.03, p \leq 1$

HOWEVER: long run hist. of r

isn't close to normal when $r = \pm 1$



Fisher (1915) Idea

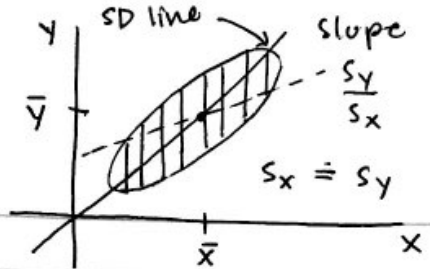
- 1) find some other quantity related to r whose long run hist. does follow the normal curve
- 2) do est. $\pm 2SE$ story on the derived quantity
- 3) back-transform to find r

★ $0.87 \pm 2(0.08) = (0.71, 1.0)$

approx 95% CI for p

STATSIG ✓

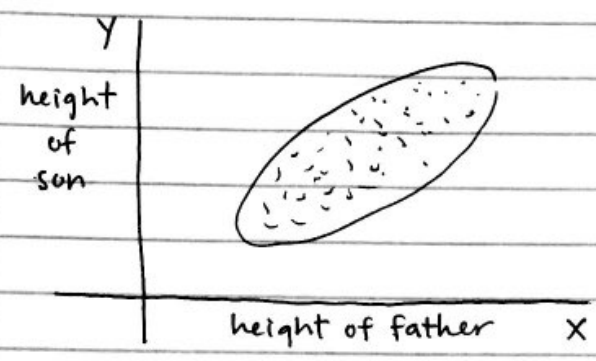




$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

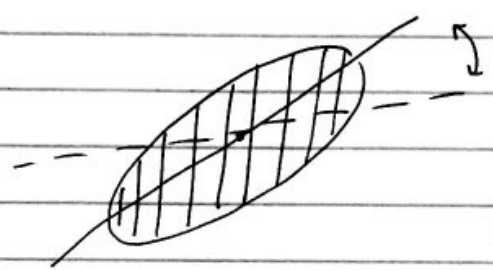
↑ predicted y-value ↑ est. intercept ↑ est. slope

Galton
1890



n = 1,000 UK families
with at least
one son

* regression to mediocrity



- tall fathers tend to have sons not as tall as them
 - short fathers tend to have sons not as short as them
- * the extremes fall back to place

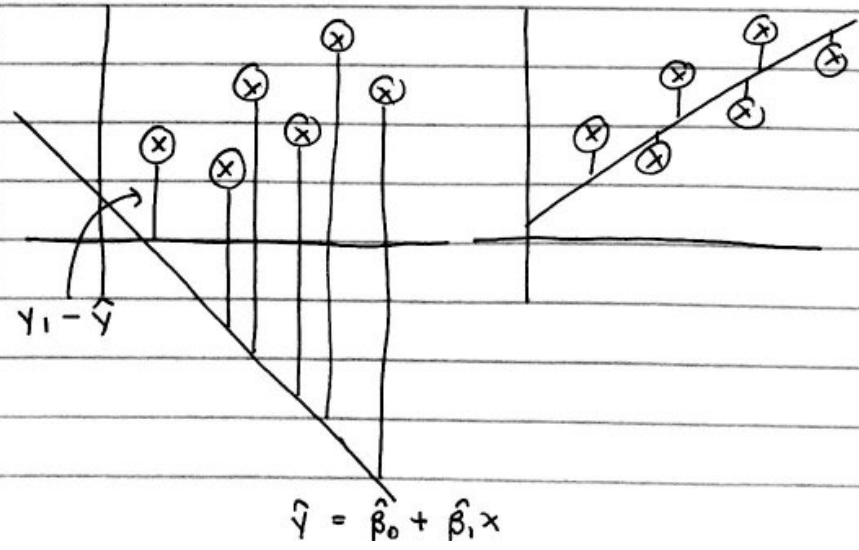
$$\hat{\beta}_1 = r \frac{s_x}{s_y}$$

$$\hat{\beta}_1 = (0.8704) \frac{0.3499 \leftarrow \text{cm of TL}}{0.3950 \leftarrow \text{cm of WL}} = 0.771 \frac{\text{cm of TL}}{\text{cm of WL}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.567 \text{ cm of TL} - (0.771 \frac{\text{cm of TL}}{\text{cm of WL}})(10.68 \frac{\text{cm}}{\text{of WL}})$$

$$= -0.669 \text{ cm of TL}$$

Gauss
1800



find $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize

$$\sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x))^2$$

↳ regression line
= $(\hat{\beta}_0, \hat{\beta}_1)$
(least squares line)