

AMS 7
CORRELATION & REGRESSION

11/12/09

read lecture notes (L-214) → (L-268)
lab #5 due MON, NOV 16 by 5pm
HW#4 due 24 NOV

$$S_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

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

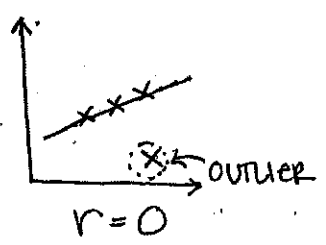
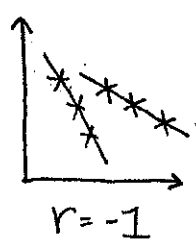
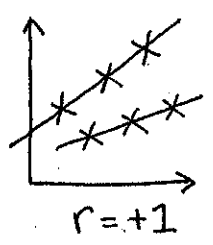
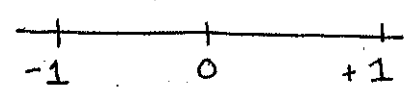
FACTS ABOUT r

① UNITS: $x \Rightarrow$ lbs $y \Rightarrow$ in
units cancel out $\Rightarrow r$ is a pure #: no units!

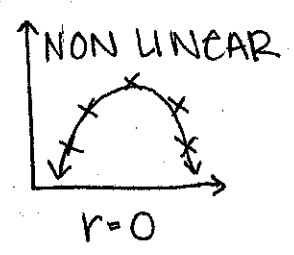
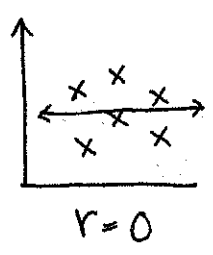
(P-25) FORMULA FOR r

②

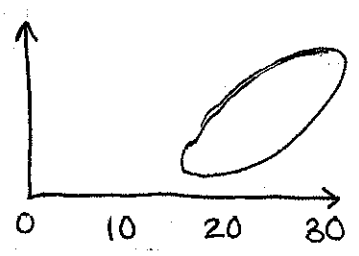
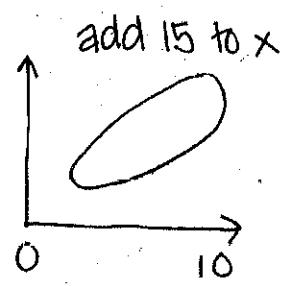
$$-1 \leq r \leq +1$$



• r -value based on normal curve;
 r can be fooled by outliers and
non-linearity (esp. w/ small n)



(3A) If I add a constant to all the x values:

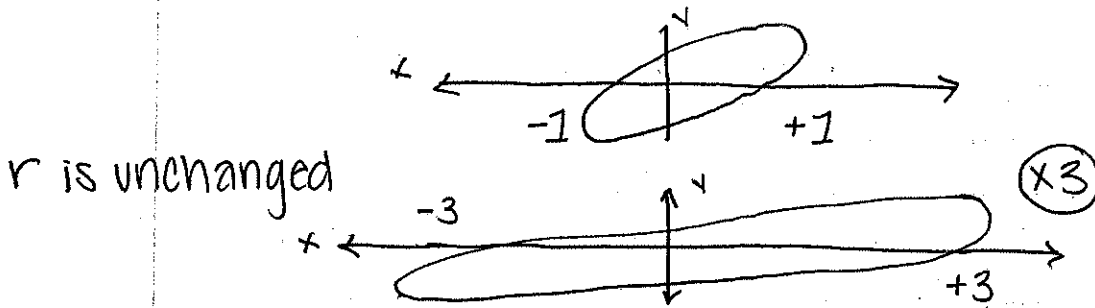


r is unchanged
SD is unchanged
same true for y

Note: $\hat{z} = i$
 → used to distinguish from 1 (1)

AC

(38) If I multiply all x (or y) values by a positive constant



$$r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}}{s_y} \right) \Rightarrow \frac{\cancel{3}x_i - \cancel{3}\bar{x}}{\cancel{3}s_x} \Rightarrow \frac{x_i - \bar{x}}{s_x}$$

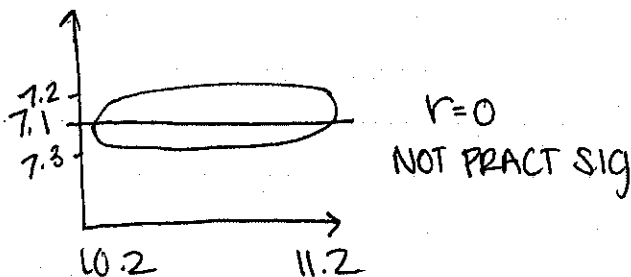
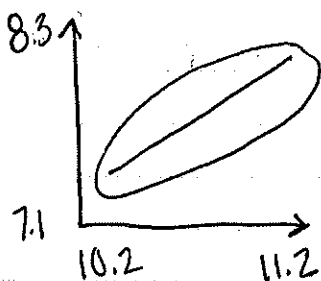
(P-98) TRAINING YOUR EYE TO READ CORRELATION VALUES (PRACTICE)

(4) The correlation between a variable & itself = +1
 • see wing & tail length data from 11110

(switch)
 If I interchange the role of x & y, r is unchanged
 • $x \cdot y = y \cdot x$

Q: WHEN IS A CORRELATION LARGE ENOUGH TO MATTER?
 PRACTICAL TERMS? STATISTICAL TERMS?

• PRACTICAL SIGNIFICANCE ⇒ the correlation between x & y is large when knowledge of x definitely helps you to predict y. Ex: sparrows (6-217)
 → when $x \approx 10.2$, $y \approx 7.1$ } $7.1 \approx 8.3$ differ by an amount
 when $x \approx 11.2$, $y \approx 8.3$ } that's pract sig ⇒ $r \approx .87$ PRACTICAL



note: spp. = species

$\rho = \rho_{HO}$

• Devil's Advocate: actual correlation = 0, +.87 due to bad sampling of 12 sparrows

$H_0: r = 0$ $H_A: r = +.87$

INFERENCEAL SUMMARY

UNKNOWN POP
quantity of interest

ρ = pop correlation between
wing & tail length

est. of ρ

$r = +.87$

give or take

$\hat{SE}(r)$

95% CI

POP

All relevant
birds of the spp.

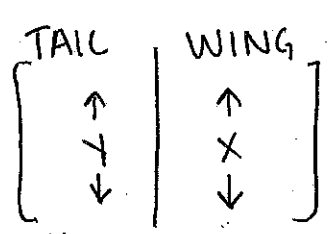
SAMPLE

observed birds

IMAGINARY

possible r values

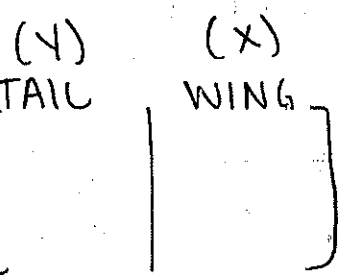
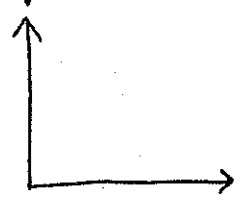
N=big



mean $\mu_Y = ?$, $\mu_X = ?$

SD $\sigma_Y = ?$, $\sigma_X = ?$

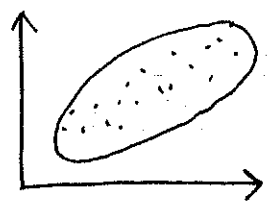
corr ρ



$\bar{Y} = 7.60\text{cm}$ $\bar{X} = 10.7\text{cm}$

$S_Y = 0.35\text{cm}$ $S_X = 0.40\text{cm}$

corr $r = +0.87$

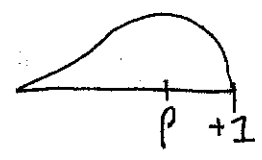


$$\begin{bmatrix} +0.87 \\ +0.82 \\ \vdots \end{bmatrix} N = \infty$$

1 $E_{IID}(r) = \rho$

2 $\hat{SE}_{IID}(r) = 0.14$

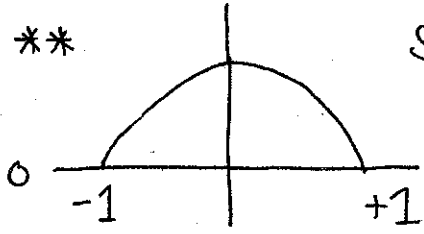
3



MATH FACT 1 $E_{IID}(r) = \rho$

MATH FACT 2 $\hat{SE}_{IID}(r) : \uparrow n, SE(r) \downarrow$, also depends on ρ ***

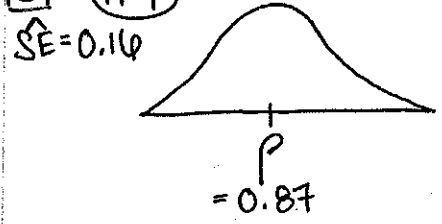
*** $SE_{IID}(r) = \sqrt{\frac{1-\rho^2}{n-2}} \Rightarrow$ don't have $\rho : SE$



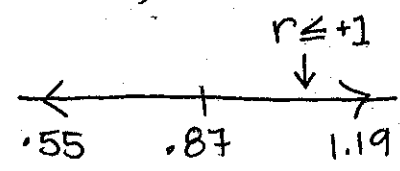
$\hat{SE}_{IID}(r) = \sqrt{\frac{1-r^2}{n-2}}$ ON FORMULA SHEET

$\sqrt{\frac{1-(+.87)^2}{12-2}} = 0.16$

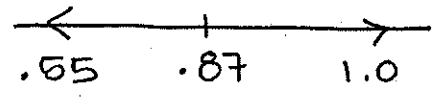
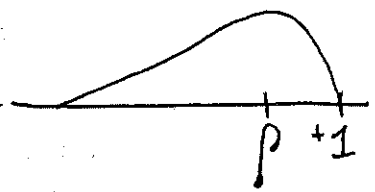
3 (TRY) APPROX W/NORMAL



$r \pm 1.96 \hat{SE}(r) \Rightarrow 0.87 \pm 2(0.16)$



LONG RUN HISTOGRAM OF r



95% APPROX CI (.55, 1.0)

2 SOLUTIONS:

↑
○ VERY UNLIKELY (H_0): STAT SIG (EC)

HW#4
PROB#3,
(a) = EC
(b) = expected

- (a) EXACT : (L-23) \rightarrow (L-244)
- (b) APPROX (CLT) = 1.96

optional for extra credit

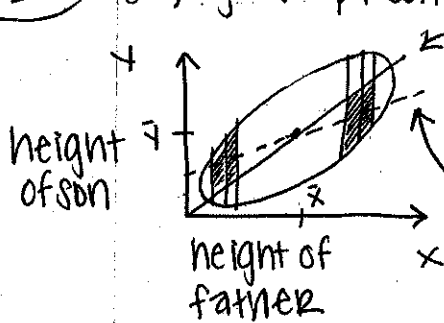
○ REGRESSION \Rightarrow what's the equation of the line capturing trend of ellipse? THREE possible lines:

- * (a) predict y from x
- (b) predict x from y
- (c) capturing overall trend

L-245

(a) goal: predict y from x

Galton (1880's - 90's)



- ① line must pass through \bar{x}, \bar{y} (pt of origin)
- ② slope: worked out averages of
VERTICAL STRIPS = REGRESSION LINE