

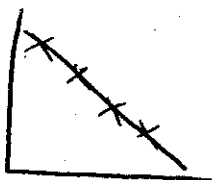
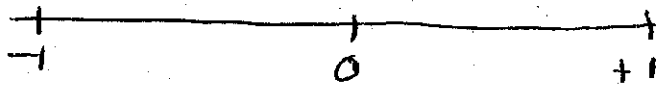
(1) Lecture #15 Correlation & Regression

26 Feb, 2009

$r$  = the correlation coefficient continued from  
lecture #4 24 Feb, 2009

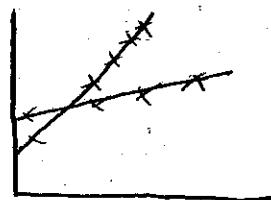
**Facts  
about  $r$**

- ①  $r$  is a pure # without units (equation cancels them all out)
- ②  $r$  is always between  $-1 \leq r \leq +1$



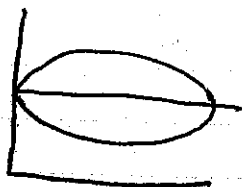
any negative slope  
(all pts on line)

(not necessarily -1)



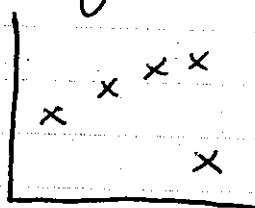
any positive slope  
(all points on line)  
(not necessarily +1)

0



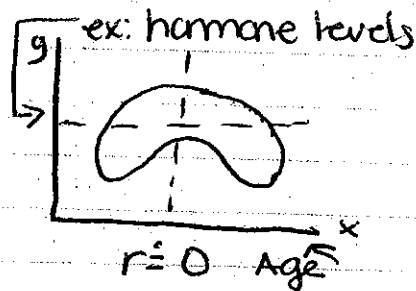
$r = 0$   
linear  
association

or



$r = 0$   
outliers

or



ex: hormone levels  
 $r = 0$   
ex: Quadratic or  
Parabolic  
relationship  
(nonlinearity)

→ Equation can be  
fooled with all of the  
conditions above which  
leads to  $r = 0$

⇒ ③  $r$  can be fooled by outliers  
(esp when  $n$  is small or nonlinear)

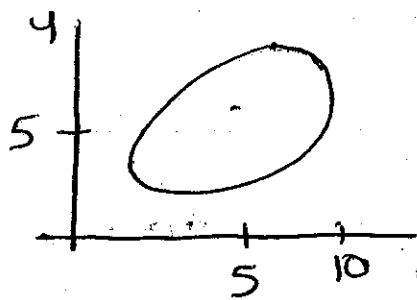
- Practice to train your eye to read correlation values  
is on page Reader 102.

(2) Lecture #15

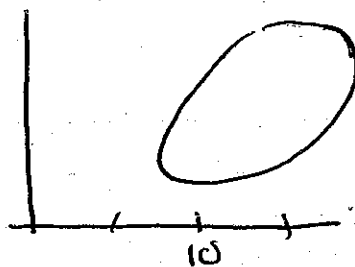
-JMP on pg 216 has the  $r$  coefficient for our sparrow's wing length from Lec #14 on 24/2/09 which is .8704 found underneath correlations

Here  $r = +0.87$ , a strong but not perfect linear association between wing length and tail length.

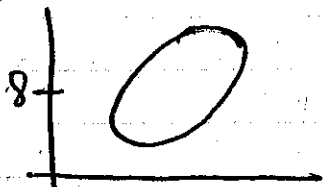
④ If you add a constant to all  $x$  or  $y$  values,  $r$  is unchanged (add  $\neq$  subtract)



Add 5 to  $x$

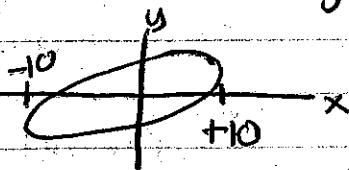


↓ (Add 3 in  $y$ )

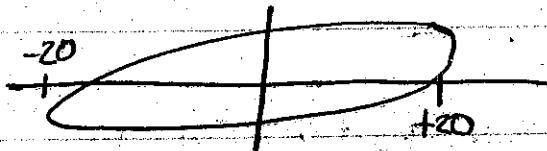


Same basic shape and relationship holds and  $r$  is the same

⑤ If multiply all  $x$  or  $y$  values by a positive constant,  $r$  also stays the same

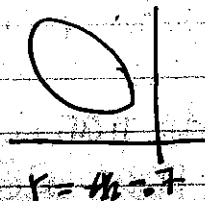
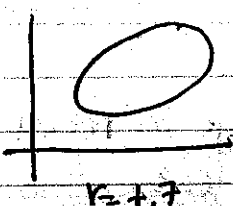


Multiply by 2  $\Rightarrow$



In the equation the constants cancel to keep the same  $r$

Also, if multiplied by negative:



(3) Lecture #15

Inference with A Correlation

Q: Is  $r = \pm 0.87$  Practsig?

A: Yes: There is a lot of info in wing length for prediction of tail length as shown by JMP on pg 16.

Q: Is  $r = +0.87$  Statsig? A: Let's see ↓

Inferential Summary (from model on pg 4)

unknown $\rho$ of interest	$\rho =$ pop. correlation between this wing/tail length in this species
estimate	$r = +0.87$
Give or Take for $r$ as est. of $\rho$	$\hat{SE}(r) = 0.16$
95% CI for $\rho$	approx: (0.55, 1.0) exact: (0.59, 0.96)

26 Feb, 2009

(4) Lecture #15

pop

All relevant birds of this species

tail ↑  
 wing ↑  
 ↓ y ↓  
 ↓ x ↓

r=?  
sig)

mean  $\mu_y$ ?  $\mu_x$ ?  
 SD  $\sigma_y$ ?  $\sigma_x$ ?  
 Corr  $\rho$  (rho) = ?

op. for ot

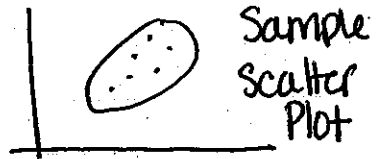


Sample  
The obs. Birds

(y) tail length  
 (x) wing length  
 $y_1, \dots, y_n$   
 $x_1, \dots, x_n$   
 mean  $\bar{y}$  = 7.6cm  
 SD  $\sigma_y$  = 0.35cm  
 mean  $\bar{x}$  = 10.7cm  
 SD  $\sigma_x$  = 0.40cm

actual (likeses) = 110

Correlation  $r = +0.87$



$\left[ \begin{matrix} \{ \\ \} \end{matrix} \right] n=18$   
 corr  $r = ?$   
 [ex: +0.82]

I.D.  
Possible values of r

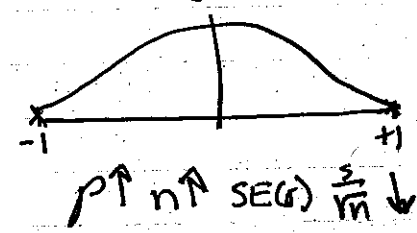
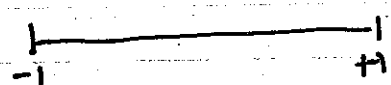
$\left[ \begin{matrix} +0.87 \\ +0.82 \\ \vdots \end{matrix} \right]$  ↑  
 ↓  $M = \infty$

long run mean  $E_{110}(r) = \rho$  fact

est. long run SD  $SE_{110}(r) = 0.16$

act

$SE_{110}(r) = ?$



A:  $SE_{110}(r) = \sqrt{\frac{1-\rho^2}{n-2}}$ ;  $SE_{110}(r) = \sqrt{\frac{1-r^2}{n-2}}$

Here:  $SE(r) = \sqrt{\frac{1-(+0.87)^2}{12-2}} = 0.156 = 0.16$

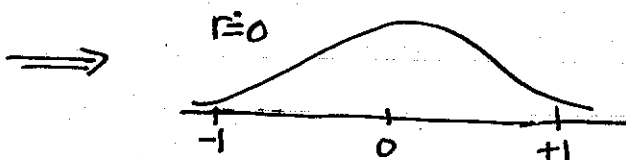
long run hist of r on next page =>

26 Feb, 2009

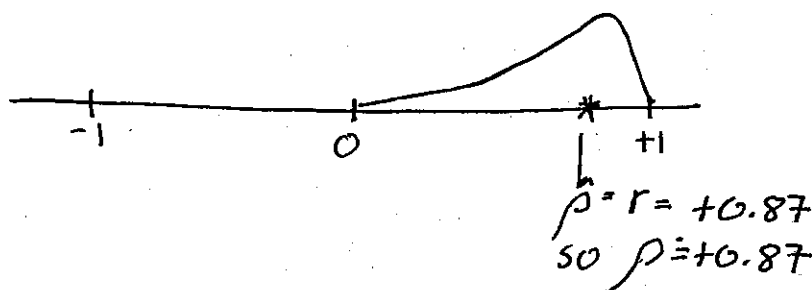
(5) Lecture #15

long run histogram of r

If r was  
 near zero

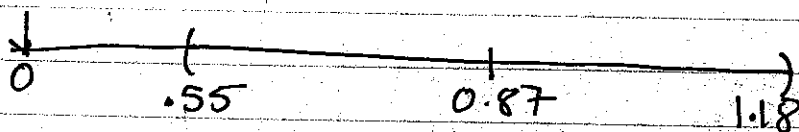


But  $r = 0.87$



95% Confidence Interval

$$\begin{aligned}
 & r \pm 2 \hat{SE}(r) \\
 & +0.87 \pm 2(0.16) \\
 & = 0.87 + 2(0.16) \doteq 1.18 \quad p \leq 1 \text{ so truncate at } 1 \\
 & = 0.87 - 2(0.16) \doteq .55
 \end{aligned}$$



The difference is statsig!

- Lecture Notes pages 231-244  
 were skipped - you do  
 not need to cover  
 that material.

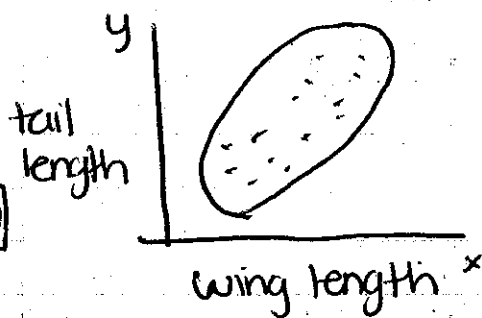
(u) Lecture Notes #15

(repeat) Approximate 95% CI for  $\rho$ :  $r \pm 1.96 \hat{SE}(r)$

$+0.87 \pm \underbrace{1.96(0.16)}_{\pm 0.32}$

$= (0.55, 1)$

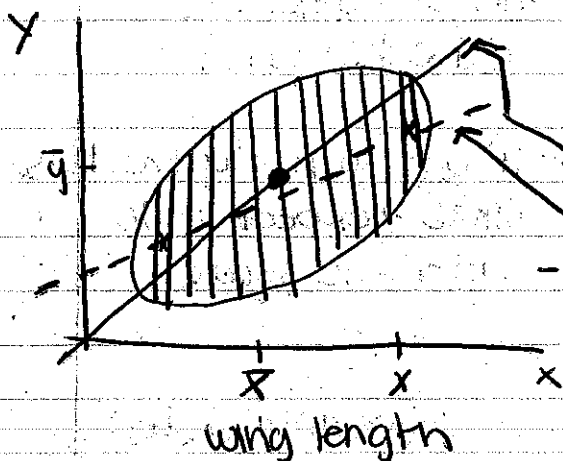
- Approximate produces safe intervals  $\Rightarrow$  ones that are wider than we need them to be
- D is not in the CI by any shape of the imagination so it is stat sig.



$r = +0.87$

Q<sub>1</sub>: How can you use x to predict y?

Q<sub>2</sub>: What is the equation of the best line for predicting y (tail length) from x (wing length)?



line: slope  $\frac{s_y}{s_x}$   
 (regression line for predicting y from x)  
 $\hookrightarrow$  slope  $r \cdot \frac{s_y}{s_x}$

Equation of predicting best line for y from x

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

Discussion  $\Rightarrow$

26 Feb, 2009

(7) Lecture #15

- ① Any good line has to go through  $(\bar{x}, \bar{y})$ .
- ② Slope? Many people guess best line is SD line, with slope  $s_y/s_x$ .

- Average  $y$  in each vertical strip: graph of Averages. Line which smooths out graph of averages - Galton called the regression line.

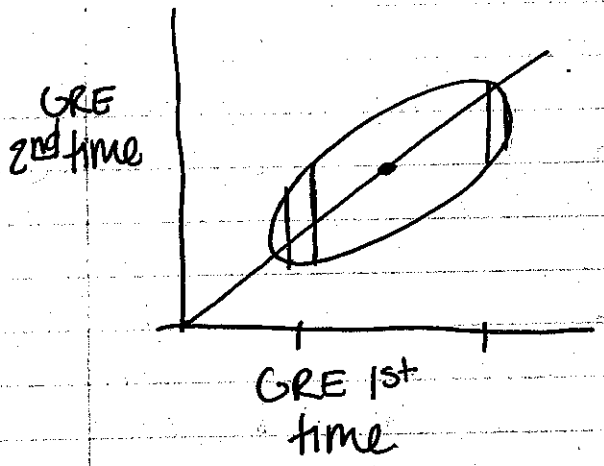
Fact: (Galton, 1880's) Slope of regression line

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

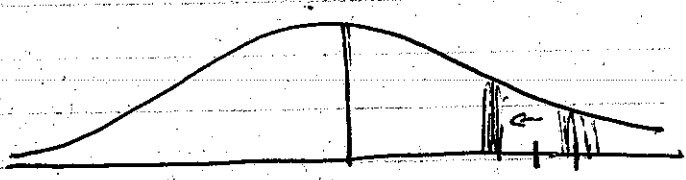
line goes through  $(\bar{x}, \bar{y})$  -  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

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

ex: GRE Test



- If you do great the first time  $\rightarrow$  expect to do worse the second
- If you do poorly the first time  $\rightarrow$  expect to do better the second.



More area under lower score (lucky w/ high) = regression line

26 Feb, 2009

(8) Lecture #15

ex: tail L | wing L.  
 $\begin{bmatrix} y & x \end{bmatrix} n=12$   
 mean  $\bar{y} = 7.567$  cm of t.L. |  $\bar{x} = 10.668$  cm of W.L.  
 SD  $s_y = 0.3499$  cm |  $s_x = 0.3950$  cm  
 $r = +0.8704$

$$\hat{\beta}_1 = (0.8704) \frac{0.3499 \text{ cm of T.L.}}{0.3950 \text{ cm of W.L.}}$$

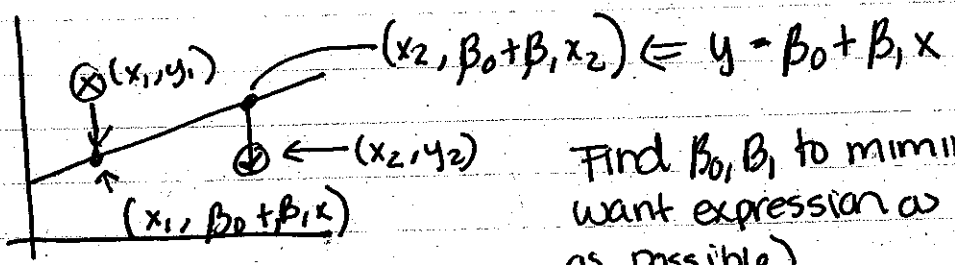
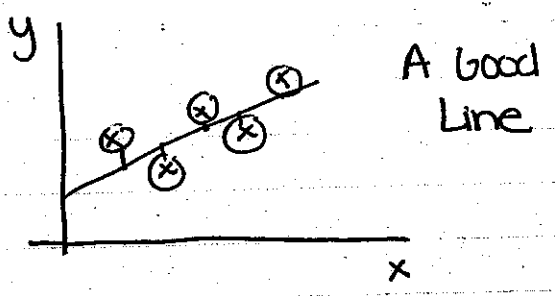
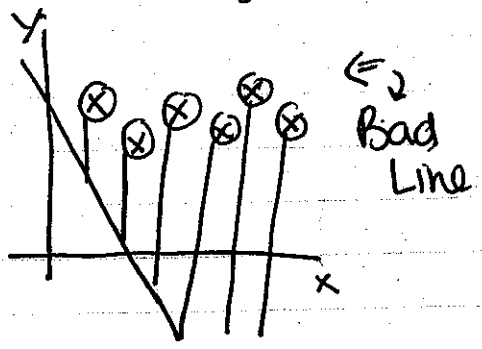
$$= 0.771 \frac{\text{cm of T.L.}}{\text{cm of W.L.}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= 7.567 \text{ cm of T.L.} - \left( 0.771 \frac{\text{cm of T.L.}}{\text{cm of W.L.}} \right) \cdot (10.668 \text{ cm of W.L.})$$

$$= -0.669 \text{ cm of T.L.}$$

Another way to think about this: (Gauss)



Find  $\beta_0, \beta_1$  to minimize (want expression as small as possible)

\* least squares line \*

$$\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

**FACT** \* Regression Line = Least Squares Line \*