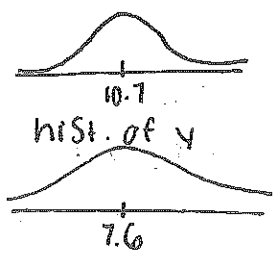
AMS 7
Lecture 12
7/15/16

Remember: Online course eval
HW 3 due monday

Simple correlation + regression (simple linear)

y = tail length (cm)
x = wing length (cm)

hist of x



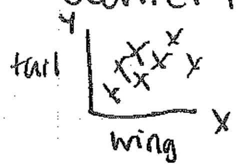
sample

Y	X
7.4	10.4
7.6	10.8
⋮	⋮
8.3	11.4

$n = 12$

mean $\bar{y} = 7.6$ $\bar{x} = 10.7$
$S_y = .36$ $S_x = 0.48$

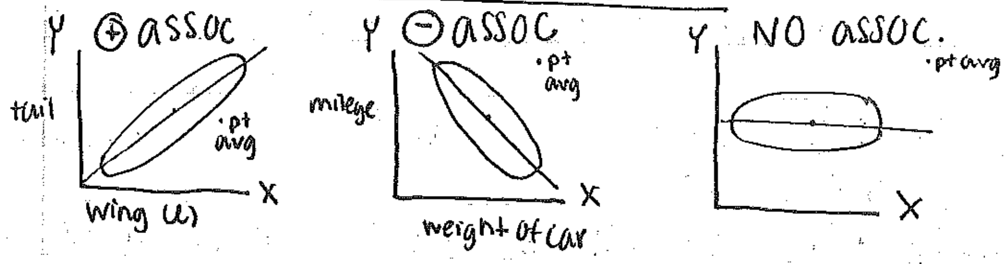
Scatter Plot



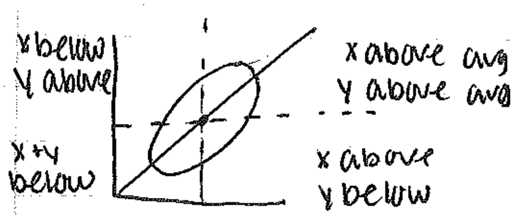
elliptical shape
bivariate normal dist

x normal
y normal

This is positive association x↑ y↑



r = strength of linear assoc. bw x + y.
r is the correlation coefficient



⊕ assoc
most of ellipse
in + and − quadrant
⊖ assoc
vice versa
no assoc
ellipse equal in 4 quadrants

$$\left.\begin{bmatrix} Y_1 & X_1 \\ Y_2 & Y_2 \end{bmatrix}\right| n \qquad \left(\frac{x_i - \bar{x}}{S_x}\right)\left(\frac{y_i - \bar{y}}{S_x}\right)$$

$\bar{Y}$ $\bar{X}$

convert

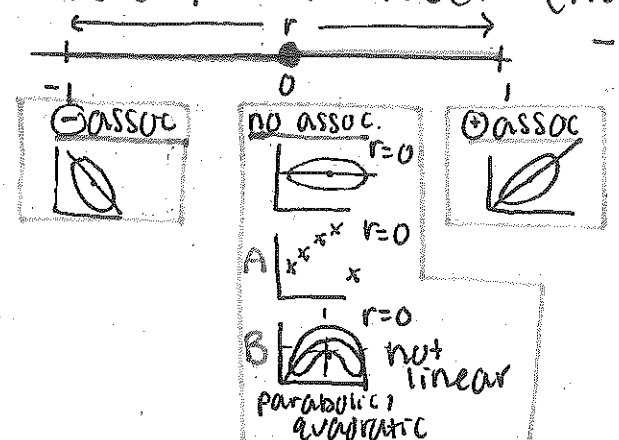
60

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

⊕ assoc.	$r > 0$
⊖ assoc.	$r < 0$
no assoc.	$r = 0$

r has no units

math facts about r:

1. r is a pure number (no units)
2. 

$-1 \leq r \leq +1$
3. r is unchanged when roles x + y reversed
4. correlation of any x with itself is +1
5. if $\pm c$ to all x values, r is unchanged (graph will shift)
   (same w/ y)
6. if $\cdot c$ to all x values, r is unchanged
   (same w/ y)

Using r inferentially:

Q_1 Is a corr. b/w wing ℓ (x) and tail ℓ (y) of r = +.87 large in practical terms?

A_1 Smallest sparrows (x = 10 cm y = 7 cm)
Largest (x = 11.6 cm y = 8.26 cm)

$$\frac{8.26cm - 7cm}{7cm} = \frac{1.26}{7} =$$

Q_2 Is corr. +.87 large in statistical terms?
(null boring value: 0) model P.-L-229

Corr value for population is RHO $= \rho$
sample is $= r$

Inf Summary

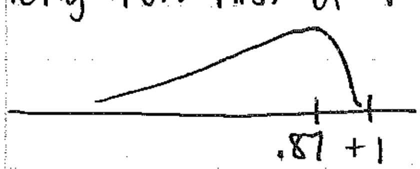
unknown	ρ pop corr
estimate	r = +.87
give/take	
95% CI	

math facts: ① $E_{IID}(r) = \rho$
② $SE_{IID}(r) = \sqrt{\frac{1-\rho^2}{n-2}}$

if we don't know ρ, use ③ $\hat{SE}_{IID}(r) = \sqrt{\frac{1-r^2}{n-2}}$

here $\hat{SE}(r) = \sqrt{\frac{1-(+.87)^2}{12-2}} = .16$

long run hist of r *



Let's try large-sample (large n) normal approx.
$\rho \rightarrow r \pm 1.96\, \hat{SE}(r)$

becomes $+.87 \pm \underbrace{1.96(.16)}_{.32}$

95% int ρ

(.55 | .87 | +1 |) ~~1.19~~ cannot be > 1

(.55, 1) truncate @ +1 (or −1)

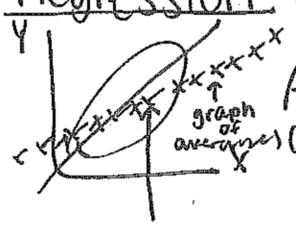
P. 231 optional
approx (+.55, +1)
exact (+.59, +.96)

95% CI

(.55 | .87) 1.0

This is stat sig from null (=0)
bc 0 is not in 95% CI

Regression Q: what is the equation of the line relating x + y



A: ① goal: use line to predict y from x

② goal: use line to predict x from y

③ goal: capture trend x-y relationships

Regression = linear / avgs

Slope of Regression Line $\boxed{\hat{\beta}_1 = r \cdot \frac{s_y}{s_x}}$ math fact

Y-int of Reg Line has to go thru $(\bar{x}, \bar{y})$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x \quad \text{so} \quad \bar{y} = \hat{\beta}_0 + \beta_1 \bar{x}$$

predicted y value — est intercept — est slope here $\hat{\beta}_1 = r \frac{S_y}{S_x} = +.8704 \frac{0.3499 \text{ cm tail } \ell}{0.3960 \text{ cm wing } \ell}$

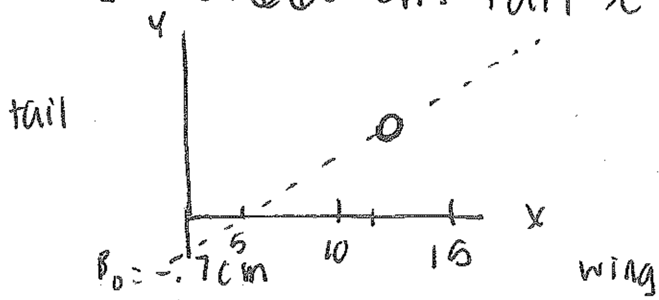
$= 0.771$ cm tail ℓ / cm wing ℓ

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

here y-int is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$

$= 7.66 - .771 \cdot 10.669$

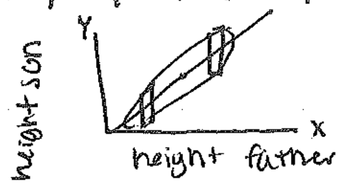
$= -0.669$ cm tail ℓ



↑ this is okay! wing and tail > 0 to exist

When scatterplot is centered in x far from origin, y-int = meaningless

Why Regression?



tall (short) fathers tend to have tall (short) sons, but not as tall (short) as father