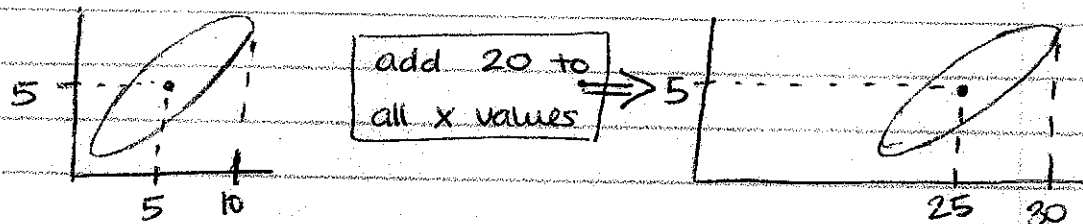


How CONSTANTS IMPACT r

ADD OR SUBTRACT a constant c to all the x or all the y :

- exactly the same scatterplot shifted so that the new \bar{x} or \bar{y} is found at (old value + constant)
- no difference in r (r is unchanged by addition/subtraction of a constant)



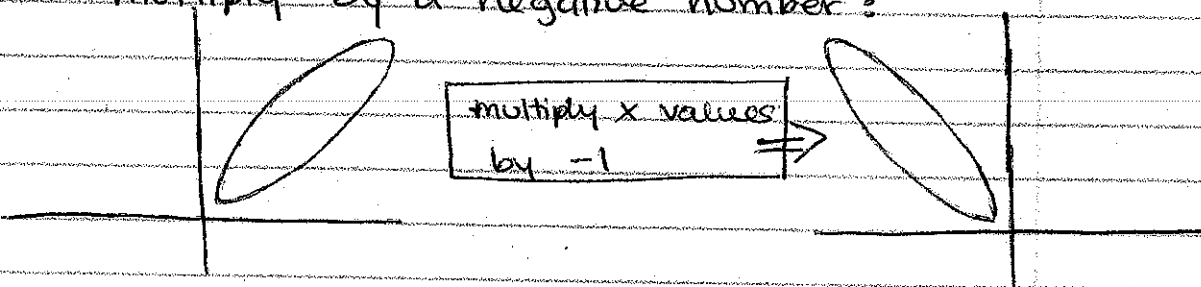
add/subtract c to $\sum x$ values: shift right/left respectively by c
 $\sum y$ values: shift up/down respectively by c

MULTIPLY all x values or all y values by a constant c

- the histogram stretches such that the mean \bar{x} or mean \bar{y} does not change
- correlation doesn't actually change, although it may seem that way.



- multiply by anything greater than 1 \rightarrow stretch
- multiply by a fraction (divide) \rightarrow compress
- multiply by a negative number:



Making an inference using a Correlation

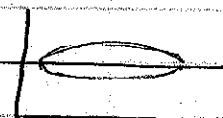
EX: sparrow winglength vs. tail length

$r = 0.87$

PRACTISIG?

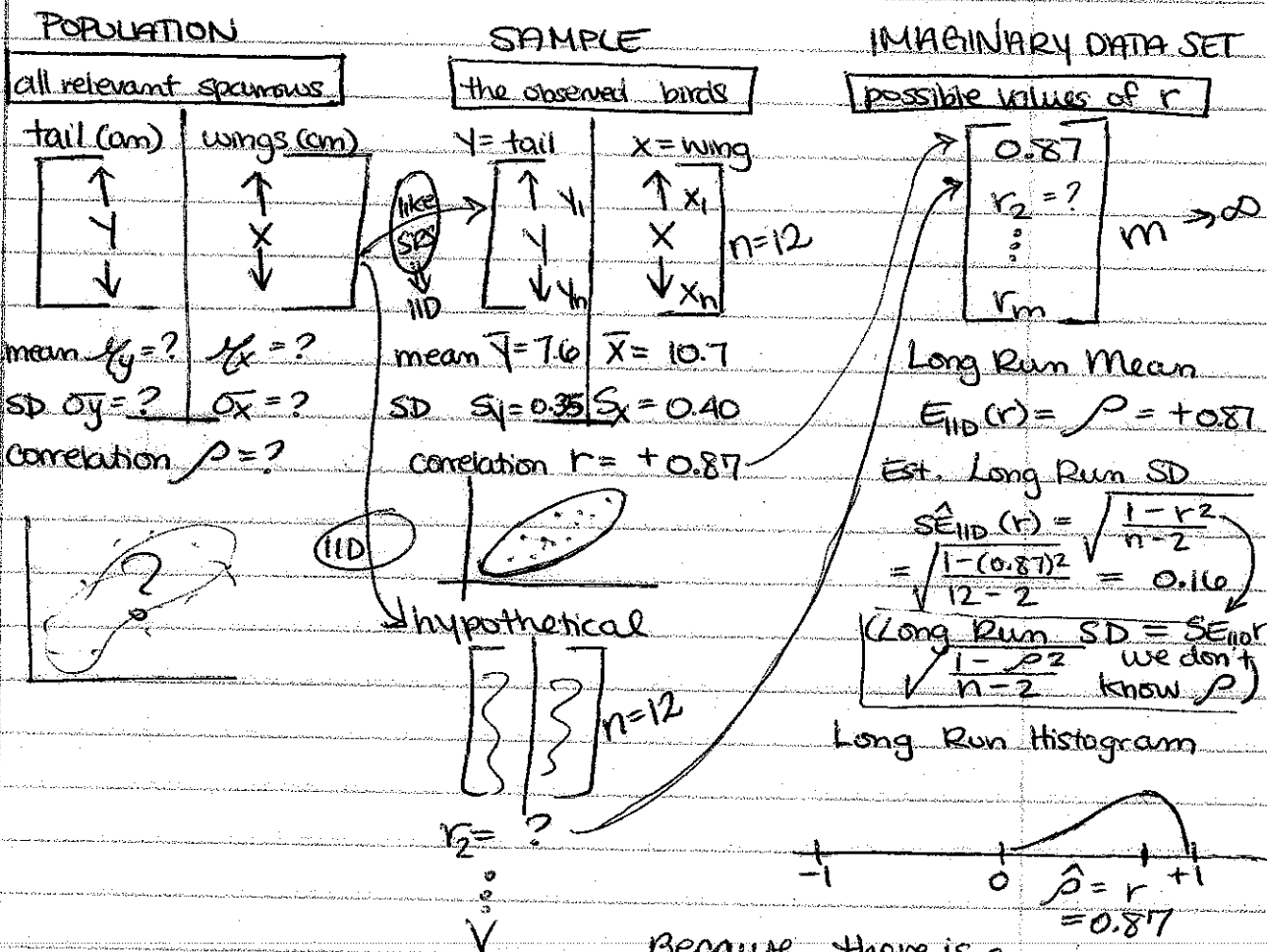
Yes \rightarrow correlation is large. Lots of info in winglength can be used to predict tail-length.

- boring null hypothesis: $r = 0$
 $0.87 \gg 0$



STATSIG?

Basic model: bivariate data set \rightarrow 2 column in both sample and population; histogram = scatterplot



Population symbol for correlation (r for sample) = ρ (rho)

Because there is a barrier @ +1, the histogram has a tendency to be skewed

Finding Long Run Mean, SE, & Histogram.

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

$$E_{IID}(r) = \rho = 0.87$$

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

We don't know ρ , so we use our estimate r

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

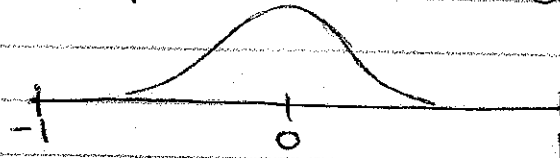
$$\hat{SE}_{IID}(r) = \sqrt{\frac{1-0.87^2}{12-2}} = 0.16$$

- This is a pretty wide uncertainty band in comparison to r .

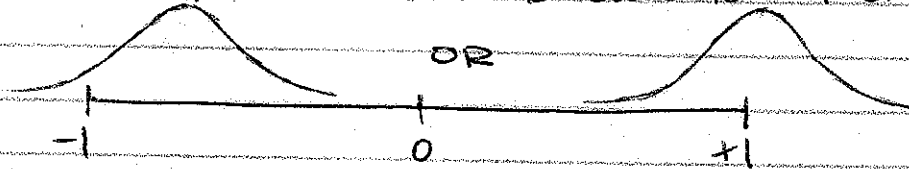
- Uncertainty bands around sample correlation are wide when $n = \text{small}$.

Long Run Histogram

- If sample correlation is close to 0



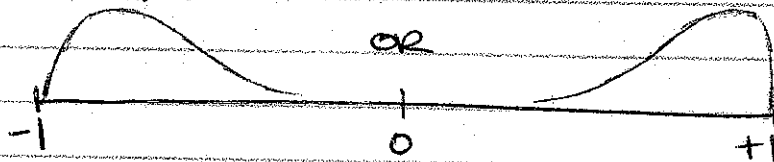
- If sample correlation is close to +1 or -1



Both of these situations are embarrassing because

-1 and +1 are the limits of sample correlation (r)

- We truncate



These histograms look nothing like the normal curve. Using the 95% CI can give us some unreasonable results.

- We will not go over how to use 95% CI w/o a normal curve

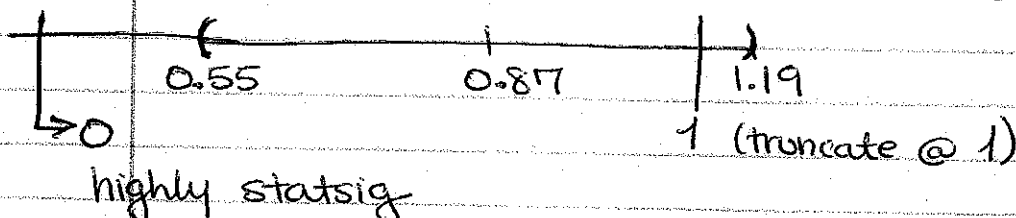
- Instead, approximate method for 95% CI \rightarrow we do have a normal curve

APPROXIMATE 95% CI FOR $\rho = r \pm 1.96(\widehat{SE}(r))$

Our method is more conservative than the actual method for 95% CI

- The intervals are wider

Sparrows: $r \pm 1.96(\widehat{SE}) = 0.87 \pm \underbrace{(1.96)(0.15)}_{0.32}$



Exact 95% interval = (0.59 → 0.96)

- narrower interval

- Although JMP can give us the exact answer by doing the exact method for any histogram, it CANNOT do correlation

- The squeaky clean version for finding the 95% CI w/o a Normal Curve is on pages 231-244

REGRESSION - the prediction of one variable from another

- here it does matter which variable correlates to each symbol

• Usually use x to predict y

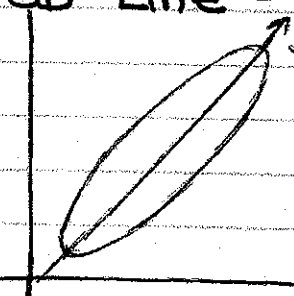
- usually x = independent variable, y = dependent var.

- we want to draw a line through the data that can be used to predict y from x

• must know the equation for the best line for predicting y from x .

• any good line should go through the point of averages.

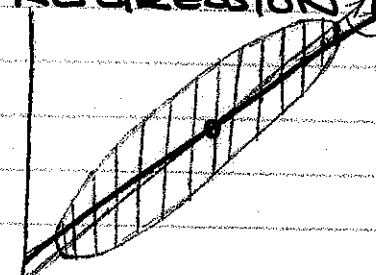
SD Line - line that goes through the major axis of the ellipse



$$\text{Slope} = m = \frac{S_y}{S_x}$$

Through trial and error, it turns out that this is not the best line for predicting y from x .

REGRESSION LINE: best line for predicting y from x



- When we split up the ellipse into sections, we see that the SD line does not go through the middle of each section.

- Regression line goes through the midpoint of each section.

- both lines go through point of averages

$$\text{Slope} = m = r \left(\frac{S_y}{S_x} \right)$$

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

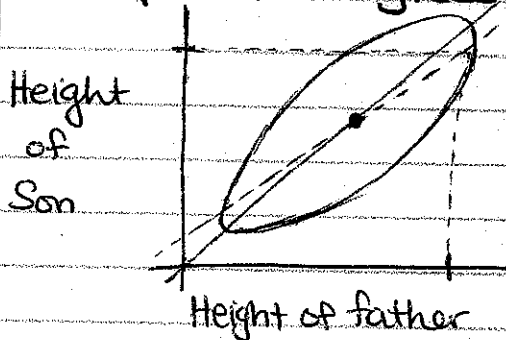
Y-intercept
slope

estimate for y
(predicted value)

→ We know that the regression line goes through the point of averages (\bar{x}, \bar{y}) . We can rearrange the equation of the line to find the y -intercept:

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

Example of Regression Line:



Say the father's height is 2 SDs above average.

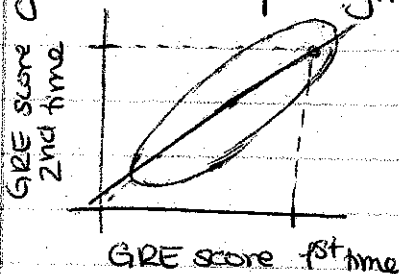
- tall fathers tend to have tall sons who are shorter than their fathers

• Son is tall, but not 2 SDs above average.

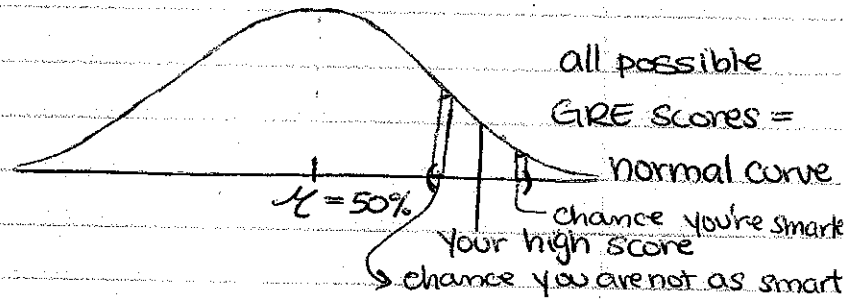
- Similarly, short fathers tend to have short sons who are taller than their fathers.

Example: GRE

Suppose that the first time you take the GRE, you get a really high score.



The second time you take the GRE, you will likely get a lower score.

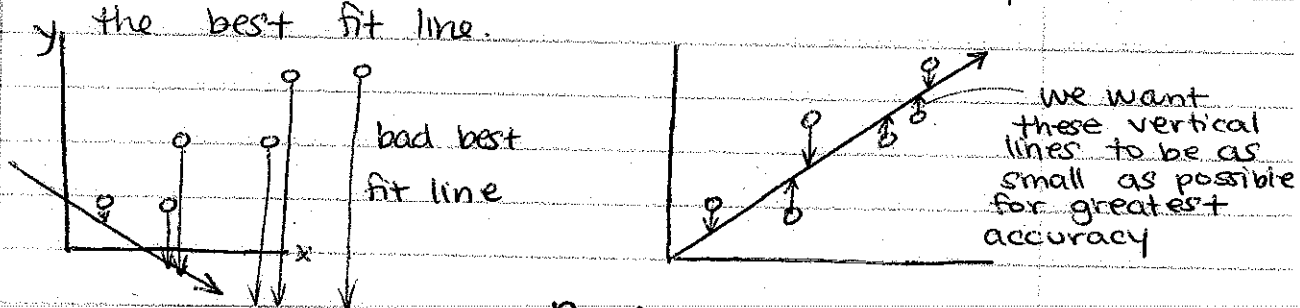


On the histogram, there is a little bubble within which your true knowledge is found. Because the slope of the histogram at that point is negative, there is a better chance that your actual knowledge is a little lower than you performed than that you are actually smarter.

- Similarly, if you perform really badly, there is a good chance that you will do a little better the second time.

Another way to look at this line. (Gauss)

- We want to make a line through the data points such that all the points are as close as possible to the best fit line.



LEAST SQUARES LINE:
$$\sum_{i=1}^n (Y_i - (B_0 + B_1 X_i))^2$$

Regression Line = Least Squares Line.