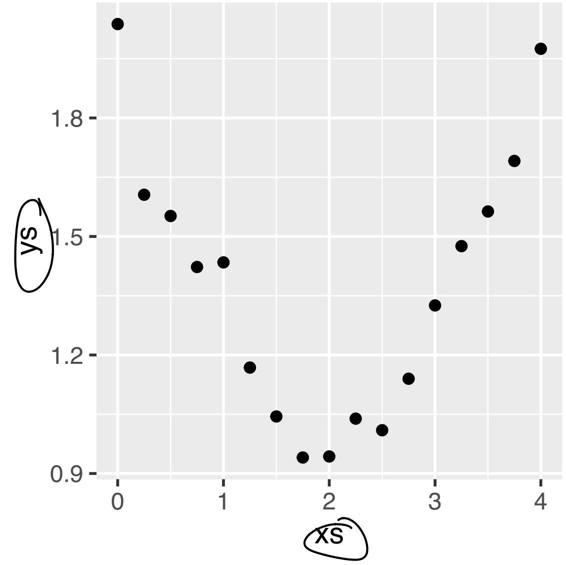
```
Stat 145, Wed 17-Feb-2021 -- Wed 17-Feb-2021
Biostatistics
Spring 2021


-------------------------------
Wednesday, February 17th 2021
-------------------------------
Note:: Ash Wednesday

-------------------------------
Wednesday, February 17th 2021
-------------------------------
Wk 3, We
Topic:: Correlation
```

The Correlation Coefficient

- It is used for (near) linear relationships between *quantitative* variables. The data involved must be true *bivariate data*—i.e., two quantities measured on the same subjects/units.
  - These are the same kind of scenarios (variable-wise) as those in which a scatterplot is possible.
  - You could not talk about the correlation coefficient between these two variables: *model of car* and *price of car*.
- It measures direction and strength of a *linear* relationship.
  - distinction between variables *having an association* and variables being *correlated*. The authors use the phrase "two variables are correlated" as synonomous with say "the two variables have an association", which seems to add only to the confusion.
  - Be careful! Data that has a strong association, can have a correlation coefficient near zero. Look at your data to see if a correlation coefficient makes sense.

    ```r
    xs = seq(0,4,.25)
    ys = (xs-2)^2 / 4 + 1 + rnorm(length(xs), 0, 0.1)
    gf_point(ys ~ xs)
    cor(ys ~ xs)
    ```

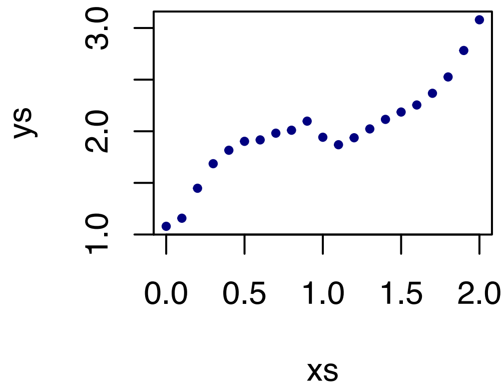


```
cor(ys ~ xs)

[1] -0.002206346
```

Very near zero, but there is an association, just not a linear one

- Similarly, data can produce a correlation coefficient close to (±1), even though the relationship is not linear:

```
xs = seq(-1,1,.1) + 1
ys = (xs-1)^3 + rnorm(length(xs), 0, 0.05) + 2
plot(xs, ys, col="navy", pch=19, cex=.5)
```



Not linear (cubic?)

```
cor(ys ~ xs)

[1] 0.9094978
```

- As with other quantities (the *mean*, for instance), there is a **population correlation** coefficient (denoted by ρ) and a **sample correlation** (denoted by r)
- Always a number between (-1) and 1.

At the lower extreme (-1), a scatterplot of the two variables will exactly lie on a straight line with negative slope.

At the upper extreme (1), a scatterplot of the two variables will exactly lie on a straight line with positive slope.

Correlation coefficients near zero indicate a weak or nonexistent linear association.

- The sample correlation coefficient is calculated using some of the same kinds of squared deviations from the mean as "sum of squares" calculations for ANOVA, or standard deviations/variances:

like a z-score

$$\boxed{r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\,\sqrt{\sum_i (y_i - \overline{y})^2}}} = \frac{1}{n-1}\sum_i \frac{(x_i - \overline{x})}{s_x}\boxed{\frac{(y_i - \overline{y})}{s_y}}.$$

That makes it a fairly complicated number to calculate by hand. Once again, we will get the number using software. In R, you type `cor(y ~ x)`, when `x` and `y` are vectors (with the same number of entries) whose correlation you seek.

- It is a dimensionless quantity—i.e., it has no units. It will not change if, say, your x-values are converted from inches to feet, or the like.
- It is fairly sensitive to outliers. See applet at

`http://www.stat.sc.edu/~west/javahtml/Regression.html`

Q: What is wrong with this statement? "There is a strong correlation between length of stay in a job and whether you are married or not."

2 vars:
married? : Cat.
length in job? : Quant.

Play the correlation game.

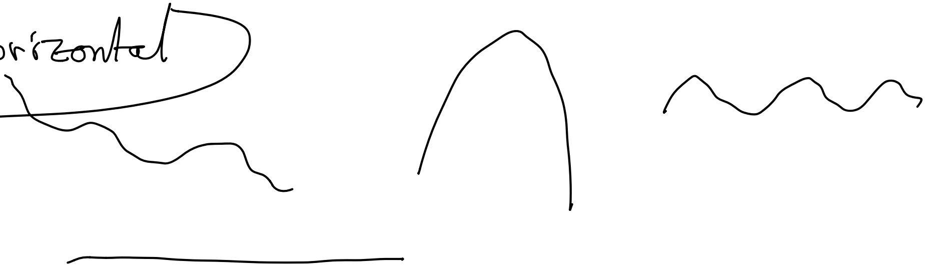
```
Q: ~~True or False~~. In the presence of two quantitative variables,
   is a 0 correlation the mark of no association?  ~~False~~
   Follow up: What is?
```

Yes or No

No

Q1: Write what you should look for on a scatter plot as a visual indication of an association

A1: Pattern, non-horizontal



Correlation not the same as slope of "regression line"

Correlation is positive / neg. matching the slope of regression line

Horizontal regression line ⟷ (near) zero correlation

Note: slope of regression line can be very close to zero (still not zero) and correlation be ± 1.

Regression line is one which best fits the points in the scatter plot (more later).

Remarks:

① Correlation ≠ slope

② Zero correlation does not mean no association

③ Correlation near ± 1 does not prove the association to be a linear one.