Stat 202 - 2015S W6 - Friday

Review - Chapter 2 - Relationships between pairs of variables

Associated - two variables measured on same cases are associated if knowing values of one tells you something about values of other.

dep Response Variable - measures outcome of a study
indep Explanatory variable - explains or causes change in response variable

Scatterplot

Oasis - strong positive
Bioclocks - weak positive
Fidget - moderate negative

Form - eg line, also could be something else. A curve for example, sine wave. We will consider only lines

Strength - How closely do data follow form?

Direction Pos/neg depending on whether above average values of one occur with above average values of other (pos) or below average values of other (neg)

Correlation

A number which quantifies the direction and strength of the linear relationship between two variables.

↓

To the extent that the form is a line this is useful. Not necessarily as useful if form is something else.

Direction: {positive association
{ r is pos
{negative association
{ r is negative

Strength: {strong association
{ r is close to ±1
{weak association
{ r is close to 0

Defined

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$s_x = \sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$$

Standardized x × Standardized y

z-score

Properties

* r does not depend on units
  change data with a linear transformation
  to change unit cm to inches or lightyears
  doesn't change correlation
* r has no units
* $r > 0$ means pos association
* $r < 0$ means negative association
* no distinction between explanatory / response vars
* Both variables must be quantitative
* $-1 \leq r \leq 1$
* values near zero indicate a weak linear relationship.
* values close to ± 1 indicate data lie close to a straight line
* $r = \pm 1$ exactly means points lie exactly on a line
* measures only strength of linear relationship not curved relationship no matter how strong.

Oasis
Bioclocks
Fidget
] guess

Regression

Regression vs Correlation

* Correlation is a measure which quantifies the strength of the linear relationship between two variables. r

* Regression finds the best fitting line

$y = mx + b$ Finds m and b — Slope and intercept

Statisticians use other notation

$$y = b_0 + b_1 x$$

(m or b_1) $\text{Slope} = r \dfrac{S_y}{S_x}$ (b or b_0) $\text{intercept} = \bar{y} - b_1 \bar{x}$

Show StatCrunch

* Regression line passes through $(\bar{x}, \bar{y})$

$$y = \underbrace{(\bar{y} - b_1 \bar{x})}_{b_0} + b_1 x$$

plug $x = \bar{x}$ get $y = \bar{y}$

~~Correlation~~ Line on scatterplot intercept / slope

* Has different slope depending on what is explanatory / response even when plotted on same axes.
* must decide which is explanatory which is response.