Stat 145, Mon 15-Feb-2021 -- Mon 15-Feb-2021
Biostatistics
Spring 2021

Monday, February 15th 2021

Wk 3, Mo
Topic:: Correlation
Read:: Lock5 2.5

Notes from grader:

- Go over Lock5 Problem 1.20
  variables: time (quantitative) and sex (categorical)
  cases: 8

raw data set for 1.20

Sex	time
M	40
F	70
M	87
M	78
M	106
M	67
F	153
F	87

Answer to 1.20 (b)

- Go over Lock5 Problem ~~1.20~~ 1.22(d)
  variables:
  10 vars.
  "did meditation program?"
  "base brain wave reading"
  "brain wave reading after meditation program"
  "brain wave reading 4 months later"
  "response to vaccine after 1 month"
  "response to vaccine after 2 months"
  "before positive emotion score from survey"
  "after positive emotion score from survey"
  "before negative emotion score from survey"
  "after negative emotion score from survey"
  cases: 41
  ==> data frame would have (at least) 41 rows and 10 columns/variables

Entering into R and using a list of numbers

- c() command
- comparing a list to a data frame
- commands like favstats(), mean(), qdata(), ... on a list
  seem to work without tilde
  if it doesn't try adding it

Associations

- Requires *bivariate data*—i.e., two variables measured on the same subjects/units
- Usually come to think of one variable as explanatory and the other as response.
- Having an association means knowledge of the explanatory variable for a case makes you better informed (even just slightly) about the value of the response for that case.

  One of the main points of inferential statistics is to discern the real associations from the phantom ones.
- Pairings of variables can be
  - two categorical variables
  - one categorical variable, one quantitative

    In this case, it is usually the categorical one that serves as explanatory.
  - two quantitative variables

Q2: Write an R command.

1. If you had a (large) data frame whose variables included `ageCategory` (18 or younger, young adult 18–25, adult 25–65, senior) and `receivedShot` (Yes, No; indicates whether the person has received a Covid vaccine shot), what would a command that helped investigate an association between variables look like? Write one out.
2. If you wished to compare `waitTime` for individuals visiting the ER at one of the local hospitals (`hospital` variable has values Butterworth, Blodgett, and St. Mary's), write a command you could use to begin your investigation.

Tools when investigating associations between quantitative variables include

- scatter plots

  Any *real*, non-horizontal-line pattern is indicative of an association

  ```
  gf_point(weight ~ height, data=women)
  ```
- correlation

  A measure on how non-horizontal, linear the pattern is

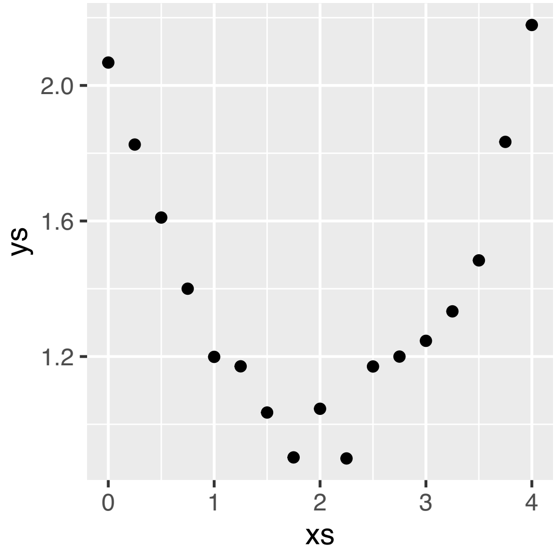
The Correlation Coefficient

- 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.

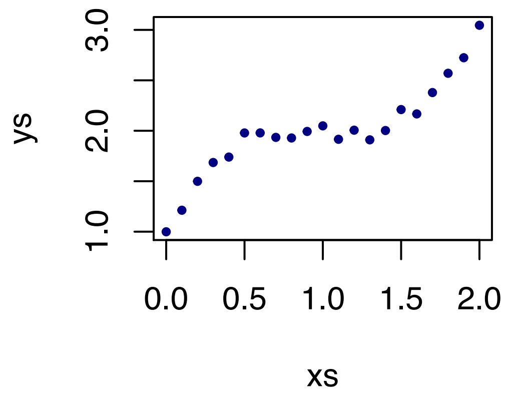
```
xs = seq(0,4,.25)
ys = (xs-2)^2 / 4 + 1 + rnorm(length(xs), 0, 0.1)
gf_point(ys ~ xs)
cor(ys ~ xs)
```



  - 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)
cor(ys ~ xs)

[1] 0.8995799
```



- 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:

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

  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`

Q3: 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."

Play the correlation game.

R Markdown

- feature in RStudio
- report-writing (can produce .pdf documents)
- suggestion: start new document using 145 H.W. template