Agenda

1. Univariate Distribution
2. Bivariate Relationships
3. Correlation

Univariate Distribution Let's review the graphics and statistics we could use to describe one variable.

- Quantitative variables

- Categorical variables

Bivariate Relationships

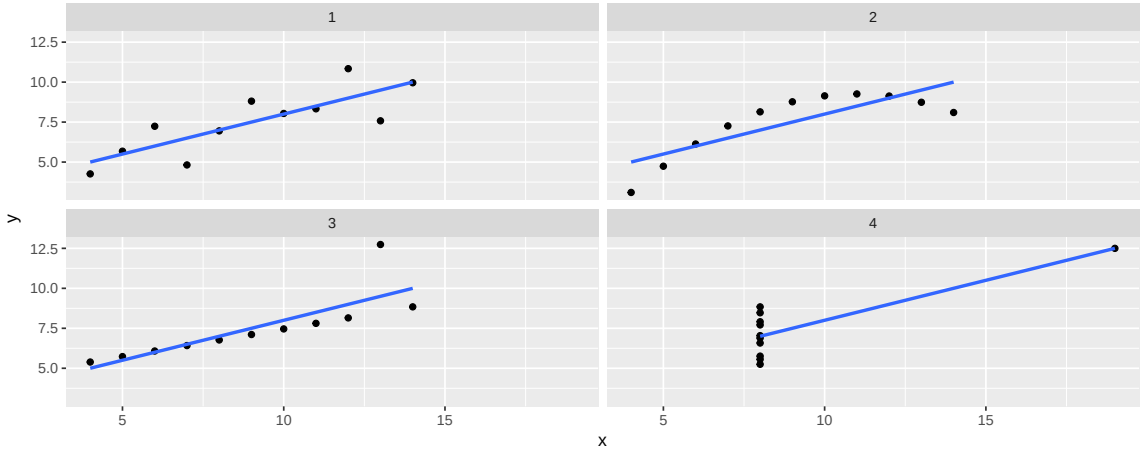
- Response variable (aka dependent variable): the variable that you are trying to understand
- Explanatory variable (aka independent variable, aka predictor): the variable that you can measure that you think might be related to the response variable
- Graphics: Put response variable on y-axis and explanatory variable on x-axis
  - Two quantitative variables: scatterplot [`qplot()` or `geom_point()`]
    - Overall patterns and deviations from those patterns
    - Form (e.g. linear, quadratic, etc.), direction (positive or negative), and strength (how much scatter?)
    - Outliers
  - Quantitative response and a categorical explanatory variable:
    - Side-by-side box plots [`geom_boxplot()`]
    - Multiple density plots [`geom_density()` with `color` aesthetic or *facets*]
  - Two categorical variables: mosaic plot [`mosaicplot()`]:
  - If a third categorical variable exists, use the `color` option or facets
- Correlation: numerical measure of direction and strength of a *linear* relationship!

```
require(mosaic)
qplot(data = KidsFeet, y = length, x = width)
qplot(data = KidsFeet, y = length, x = sex, geom = "boxplot")
qplot(data = KidsFeet, x = length, color = sex, geom = "density")
qplot(data = KidsFeet, x = length, facets = ~sex, geom = "density")
mosaicplot(domhand ~ sex, data = KidsFeet)
```

Correlation The (Pearson Product-Moment) correlation coefficient [`cor()`] is a measure of the strength and direction of the *linear* relationship between two numerical variables. It is usually denoted r and is measured on the scale of $[-1, 1]$.

```
## # A tibble: 4 x 5
##     set     N `mean(x)` `mean(y)` `cor(x, y)`
##   <chr> <int>     <dbl>     <dbl>       <dbl>
## 1     1    11         9  7.500909   0.8164205
## 2     2    11         9  7.500909   0.8162365
## 3     3    11         9  7.500000   0.8162867
## 4     4    11         9  7.500909   0.8165214
```

```
qplot(data = ds, x = x, y = y) +
  geom_smooth(method = "lm", se = 0) +
  facet_wrap(~set)
```



Note that correlation only measures the strength of a *linear* relationship. In each of the four very different (Anscombe) data sets shown above, the correlation coefficient is the same (up to three digits)!

Examples Get a feel for the value of the correlation coefficient in different scatterplots.

1. Do a Google Image search for "scatterplot" and describe the form, direction, and strength of three different-looking patterns. Sketch each plot.

   (a) :

   (b) :

   (c) :