17. Correlation

In the last 2 chapters, we focused on the relationship between 2 variables. The variables were factors (i.e., categorical variables). We discussed measures of association:

- Odds ratio (perhaps averaging over multiple tables)
- Chi-square statistic (measures association in some rough sense)

Here we continue to investigate the relationship between 2 variables measured on the same unit, but now investigate measures for ordinal data and numerically discrete (e.g. counts) and continuous (e.g. height) data. We look at our data as arising from pairs of random variables (X_1, Y_1), …, (X_n, Y_n) each pair measured independently from the other (due to randomly sampling the units from which pairs are measured) We'll discuss 2 measures of correlation to see how the X_i values change with the Y_i values (or how the ranks of the X_i changes with the ranks of the Y_i):

- Pearson's Correlation coefficient (r): a measure of the linear association between two numerical variables, and
- Spearman's Rank Correlation (r_S): a measure of the linear association between two sets of ranks (from numerical or ordinal data pairs)

17.1. Two-Way Scatter Plot

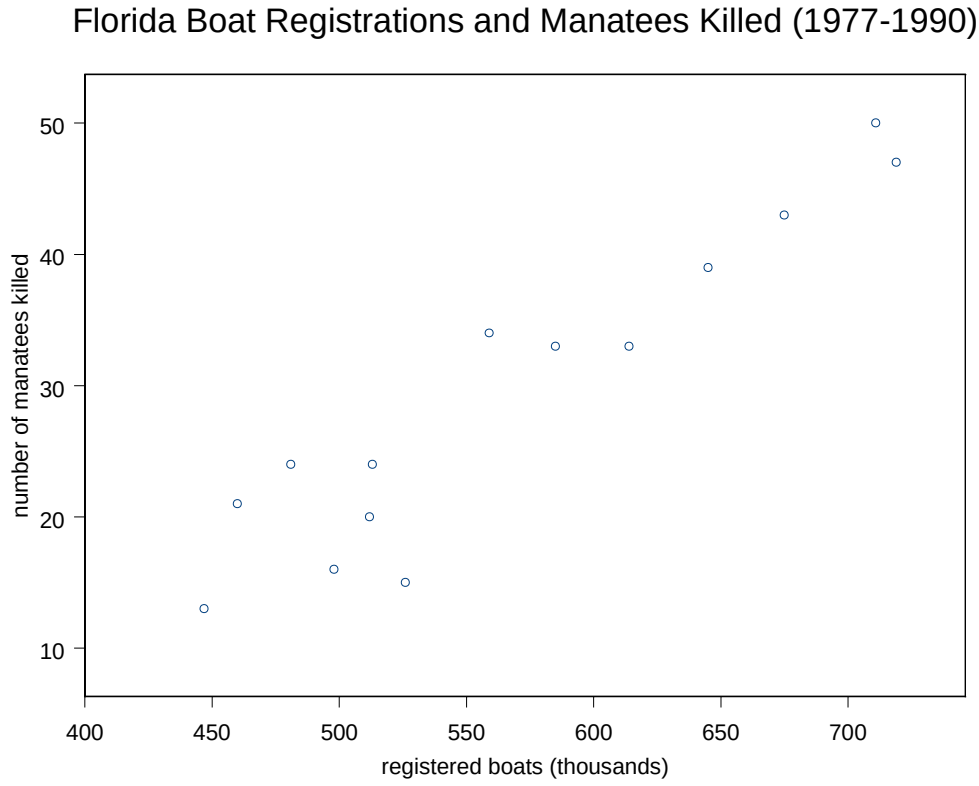
It's always a good idea to first plot your data before you begin your analysis. The 2-way scatter plot is particularly suited to

investigating the relationship between 2 numerical variables (or ranks).

S-Plus 2D Scatter plot

Graph>2D Plot…

then Plot Type: Scatter Plot (x, y1, y2, …)



What to look for:

- Direction—Is there an upward (positive) or downward (negative) trend?
- Form—Does the pattern seem straight or is it curved?
- Strength—Do the points appear to lie on or near a line or curve (strong relationship) or are they very scattered about a line or curve (weak relationship)?
- Outliers—Are there points that do not follow the general pattern in the data?

17.2. Pearson's Correlation Coefficient

Pearson's (sample) correlation coefficient:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

$$= \frac{1}{(n-1)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \right]}}$$

- Measures the linear association between two numerical variables
- r is used as an estimator of the population correlation coefficient:

$$\rho = \text{average} \left[\frac{(X - \mu_x)(Y - \mu_y)}{\sigma_x \sigma_y} \right] = E \left[\frac{(X - \mu_x)(Y - \mu_y)}{\sigma_x \sigma_y} \right]$$

- values of r and ρ are between -1 and +1
  - negative correlation means negative linear association
  - positive correlation means positive linear association
  - -1 means perfect negative linear association
  - +1 means perfect positive linear association
  - 0 means no *linear* association
- dimensionless
  - doesn't depend on the scale on which the data are measured (e.g., cm, ft, inches, meters)

- no change when multiply one or both variables by a constant or when shift (add a constant to) one or both variables

- r may be misleading if there are
  - outliers (strongly affect r), or
  - nonlinear associations
- does not necessarily imply cause and effect
  - may be other, lurking variables responsible for the apparent relationship

E.g. Manatee Data

Boats Registered (x_i)

```
447 460 481 498 513 512 526 559 585 614 645 675 711 719
```

Manatees Killed By Boats (y_i):

```
 13  21  24  16  24  20  15  34  33  33  39  43  50  47
```

Some Statistics

Statistics>Data Summaries>Summary Statistics…

```
      ***  Summary Statistics for data in:  ipsex2.39 ***

             boats   killed
    Mean: 567.50000 29.42857
 Total N:  14.00000 14.00000
Std Dev.:  91.90693 12.18899
```

Also, we need sum of cross-products:

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=$$

$$r = \frac{1}{(n-1)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} =$$

In S-Plus:
Statistics>Data Summaries>Correlations…

```
      ***  Correlations for data in:  ipsex2.39 ***

           boats    killed
 boats 1.0000000 0.9414773
killed 0.9414773 1.0000000
```

Although the correlation does not necessarily imply a cause and effect relationship, it often (not always) makes intuitive sense to think of one variable as the "cause" and the other as the "effect". In the above example, we like to think of the number of registered boats (x_i) as the "cause" and the number of manatees killed (y_i) as the "effect", although it should be clear that registering a boat does not cause boat to kill a manatee. In such a situation, we often call the variables the "explanatory variable" (x_i) and the "response variable" (y_i), respectively. Other terms used are "independent" and "dependent" and, less commonly, "exogenous" and "endogenous" (mostly economics?). But, it should be noted that correlation does not make any such distinction between the variables, so this terminology is somewhat premature at this point. But, correlation is a natural introduction to the material in the following chapter where will use this terminology more, and, there, it does matter which variable is dependent, etc.

The sample correlation in the manatee example seems large; is it statistically significantly different from zero? That is, is there a significant linear association between the number of registered boats and manatees killed?

H_0: $\rho = 0$
H_A: $\rho \quad 0$

If we assume the pairs, (x_i, y_i), i=1 to n, were obtained randomly (pairs are independent) and each variable X_i and Y_i is normally distributed, then we can use the Student's t distribution with (n-2) degrees of freedom (under the null $\rho = 0$)

Test statistic:

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} = r\sqrt{\frac{n - 2}{1 - r^2}} =$$

p-value:

$P(t_{(n-1)} \geq t) =$

17.3. Spearman's Rank Correlation Coefficient

If there are outliers in the data, or if the data is ordinal, then we may use an alternative measure of association between variables:

- The Spearman's rank correlation, r_S measures the association between variables by calculating the Pearson correlation coefficient for the pairs of ranks of the data (same formula, but use the ranks of the data instead)
- r_s is more resistant to outliers than r
- now linear refers to ranks, not the data, so does not necessarily measure linear association in the data

$$r_s = \frac{\sum_{i=1}^{n}(x_{ri} - \bar{x}_r)(y_{ri} - \bar{y}_r)}{\sqrt{\left[\sum_{i=1}^{n}(x_{ri} - \bar{x}_r)^2\right]\left[\sum_{i=1}^{n}(y_{ri} - \bar{y}_r)^2\right]}}$$

$$= \frac{1}{(n-1)} \frac{\sum_{i=1}^{n}(x_{ri} - \bar{x}_r)(y_{ri} - \bar{y}_r)}{\sigma_{rx}\sigma_{rx}}$$

For testing we again use a Student's t distribution, but now this is approximate

- assume pairs selected randomly (independent)
- large sample (n ≥10)

$$t_s = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

E.g. Manatees again. Just for the purpose of demonstrating the Spearman's correlation, we use the manatee data. (The data do not suggest any outliers, nor do they suggest any moderate departure from normality if we want to do a test.)

Ranks Corresponding to manatee data given previously:

Boats Rank (x_{ri})

```
1  2  3  4  6  5  7  8  9  10 11 12 13 14
```

Killed Rank (y_{ri})

```
1  7  4  6  2  3  5  9 10   8 11 12 14 13
```

Statistics>Data Summaries>Summary Statistics…

```
      ***  Summary Statistics for data in:  ipsex2.39 ***

          boats.rank killed.rank
    Mean:     7.5000      7.5000
 Total N:    14.0000     14.0000
Std Dev.:     4.1833      4.1833
```

Sum of cross-products:

$$\sum_{i=1}^{n}(x_{ri}-\bar{x}_r)(y_{ri}-\bar{y}_r)=$$

$$r_s=\frac{1}{(n-1)}\frac{\sum_{i=1}^{n}(x_{ri}-\bar{x}_r)(y_{ri}-\bar{y}_r)}{\sigma_{rx}\sigma_{rx}}=$$

In S-Plus:

Statistics>Data Summaries>Correlations…

```
            boats.rank killed.rank
 boats.rank  1.0000000   0.8637363
killed.rank  0.8637363   1.0000000
```

Could proceed to do test base on Student's t approximation to t_s, but instead, we move onto the next chapter.