2.3: Correlation

As before (in §§4 and 5), when we moved from describing histograms with words (like *symmetric*) to describing them with numbers (like the *mean*), we now will build a numeric measure of the strength and direction of a linear association in a scatterplot.

[def:corrcoeff] Given bivariate quantitative data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ the **[Pearson] correlation coefficient** of this dataset is

$$r = \frac{1}{n-1} \sum \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y} \tag{2.3.1}$$

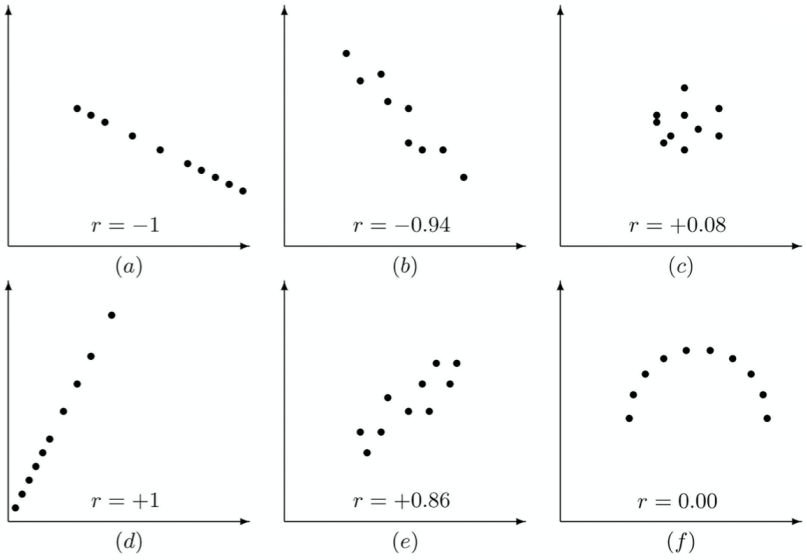
where s_x and s_y are the standard deviations of the x and y, respectively, datasets by themselves.

We collect some basic information about the correlation coefficient in the following

[fact:corrcoefff] For any bivariate quantitative dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with correlation coefficient r, we have

1. $-1 \leq r \leq 1$ is always true;
2. if $|r|$ is near 1 – meaning that r is near ± 1 – then the linear association between x and y is *strong*
3. if r is near 0 – meaning that r is positive or negative, but near 0 – then the linear association between x and y is *weak*
4. if $r > 0$ then the linear association between x and y is positive, while if $r < 0$ then the linear association between x and y is negative
5. r is the same no matter what units are used for the variables x and y – meaning that if we change the units in either variable, r will not change
6. r is the same no matter which variable is begin used as the explanatory and which as the response variable – meaning that if we switch the roles of the x and the y in our dataset, r will not change.

It is also nice to have some examples of correlation coefficients, such as



Many electronic tools which compute the correlation coefficient r of a dataset also report its square, r^2. There reason is explained in the following

[fact:rsquared] If r is the correlation coefficient between two variables x and y in some quantitative dataset, then its square r^2 it the fraction (often described as a percentage) of the variation of y which is associated with variation in x.

[eg:rsquared] If the square of the correlation coefficient between the independent variable *how many hours a week a student studies statistics* and the dependent variable *how many points the student gets on the statistics final exam* is .64, then 64% of the variation in scores for that class is cause by variation in how much the students study. The remaining 36% of the variation in scores is due to other random factors like whether a student was coming down with a cold on the day of the final, or happened to sleep poorly the night before the final because of neighbors having a party, or some other issues different just from studying time.