Correlation

What it is

Correlation measures the strength of the relationship between two numeric variables. Specifically, Pearson's correlation coefficient measures the strength of the linear relationship/trend between the variables. It is important to remember this is not the slope of the linear relationship, but how closely the points fall to a straight line.

When to use it

Technically, correlation can be computed for any two numeric variables. However, is it important to remember Pearson's correlation coefficient only measures the strength of the *linear* trend. Other trends or patterns (exponential, cyclical, etc.) may be lost or obscured. It is always a good idea to plot the data prior to computing the correlation coefficient. In addition, in order for confidence intervals and hypothesis tests to work appropriately, both variables should be approximately normal.

How to use it

We can compute the Pearson correlation coefficient between physical and mental health in R using the `cor()` function:

```
cor(x=mydat$physhealth, y=mydat$menthealth, use='complete.obs')
```

```
## [1] 0.3232012
```

To run a hypothesis test or create a confidence interval, use the function `cor.test()`. The results R gives are for the hypothesis test $H_0 : \rho = 0$ against the given alternative[1]:

```
cor.test(x=mydat$physhealth, y=mydat$menthealth)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mydat$physhealth and mydat$menthealth
## t = 15.023, df = 1935, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2827329 0.3625210
## sample estimates:
##       cor
## 0.3232012
```

[1] Like with t-tests, the options for the `alternative` input are `"two.sided"`, `"less"`, and `"greater"`, with default set to "'two.sided"

Correlations range in value from -1 to 1, with ± 1 indicating perfect correlation (points fall in a perfect line) and 0 indicating the variables are uncorrelated (no linear trend). As a general rule of thumb, we can use the following cutoffs for strength:

- $|r| < 0.3$: weak
- $0.3 < |r| < 0.7$: moderate
- $|r| > 0.7$: strong

Why it works

The formula for computing Pearson's correlation coefficient is

$$r = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y}$$

The denominator is the product of the standard deviations of x and y, and the numerator is known as the **covariance** between x and y.

Let's focus on the numerator (covariance): It is helpful to consider where the means of the two variables in the scatterplot are located, as in Figure 1. For points in the upper right quadrant, $x_i > \bar{x}$ and $y_i > \bar{y}$, so that $x_i - \bar{x} > 0$ and $y_i - \bar{y} > 0$. Thus, the product $(x_i - \bar{x})(y_i - \bar{y})$ is greater than 0 and these points add a positive amount to the correlation numerator. Similarly, points in the lower left quadrant add a positive amount to the covariance[2]. So, points in the upper right and lower left quadrants increase the covariance (and hence the correlation) and the more points in these quadrants, the more positive the correlation. By similar argument, points in the upper left and lower right quadrants add negative values, decreasing the covariance (correlation)[3]. More points in these two quadrants will lead to a more negative correlation.

[2] $x_i < \bar{x}$ and $y_i < \bar{y}$, so $x_i - \bar{x}$ and $y_i - \bar{y}$ are both negative, making $(x_i - \bar{x})(y_i - \bar{y})$ positive.

[3] If, say, $x_i > \bar{x}$ but $y_i < \bar{y}$, then $x_i - \bar{x}$ will be positive but $y_i - \bar{y}$ will be negative, and their product will be negative.

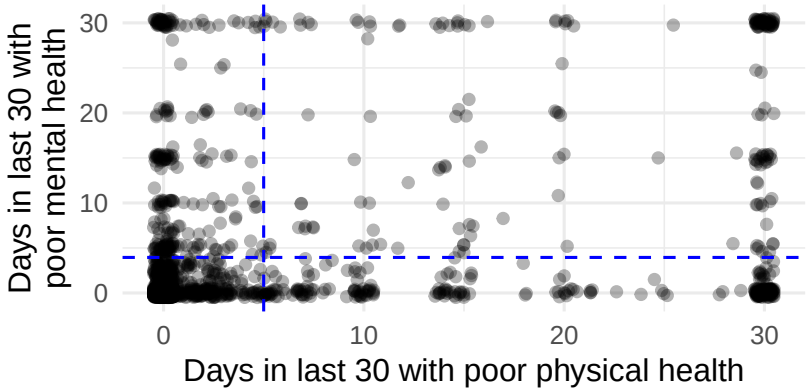

Figure 1: Scatterplot with quadrants defined by variable means. Most points fall in the bottom left and upper right quadrants, suggesting a positive correlation. However, the data are concentrated at the extremes and do not follow a normal distribution - this indicates the confidence interval and p-value above may not be accurate.