Correlation analysis

Outline

- What is correlation?
- *Pearson product-moment correlation*, or *r*
- *Pointwise-biserial correlation*
- *Spearman rank-order correlation*, or ρ ("rho")
- *Goodman-Kruskal correlation*, or ɣ ("gamma")
- *Kendall correlation*, or τ ("tau")

Some tests we've seen so far (1/)

- The binomial test:
  - Samples are: booleans ("Bernoulli trials")
  - Null hypothesis: the sample is drawn from a population with some *P*(heads)
- The one-sample *t*-test:
  - Samples are: scalars
  - Null hypothesis: the sample is drawn from a population with a certain mean
- The paired *t*-test:
  - Samples are: pairs *a, b* where *a* and *b* are both scalars
  - Null hypothesis: *a* - *b* are drawn from a population with a certain mean

Some tests we've seen so far (2/)

- The two-sample *t*-test:
  - Samples are: pairs *c*, *d* where *c* is a scalar and *d* is a boolean indicating group membership
  - Null hypothesis: the samples where *d* = True and *d* = False have the same mean
- The Fisher exact test:
  - Samples are: pairs *e*, *f* where *e* and *f* are categorical
  - Null hypothesis: there is no association between values of *e* and *f*

Correlation

In correlation statistics and tests, our samples are pairs g, h where g and h are rank, interval, or ratio variables. ("Two measures per case")

Or equivalently, we have two rank (or better) vectors G and H of the same length, where g_i and h_i represent a single observation.

Correlation statistics quantify the association between two (rank or better) variables.

Correlation tests have a null hypothesis of no association between G and H.

Correlation

Correlation is a very general approach:

- It subsumes many other methods
- Variants exist for rank data and for interval/ratio data
- It is appropriate for both experimental and observational studies

Some examples:

- Is the time it takes to identify a word in a lexical decision task related to the number of letters in that word?
- Is there a relationship between time spent studying a language and the score on a proficiency test for that language?

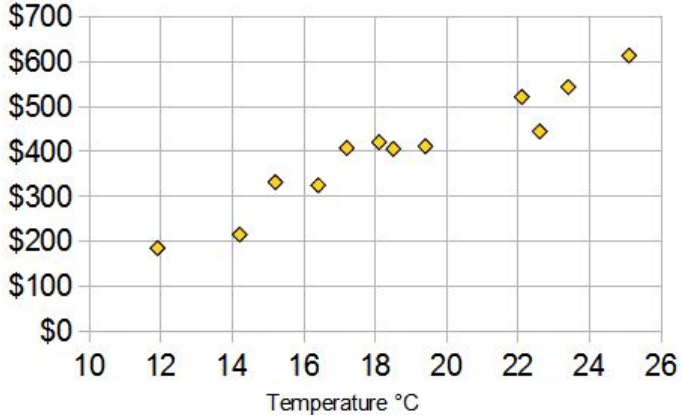
Relations between two variables

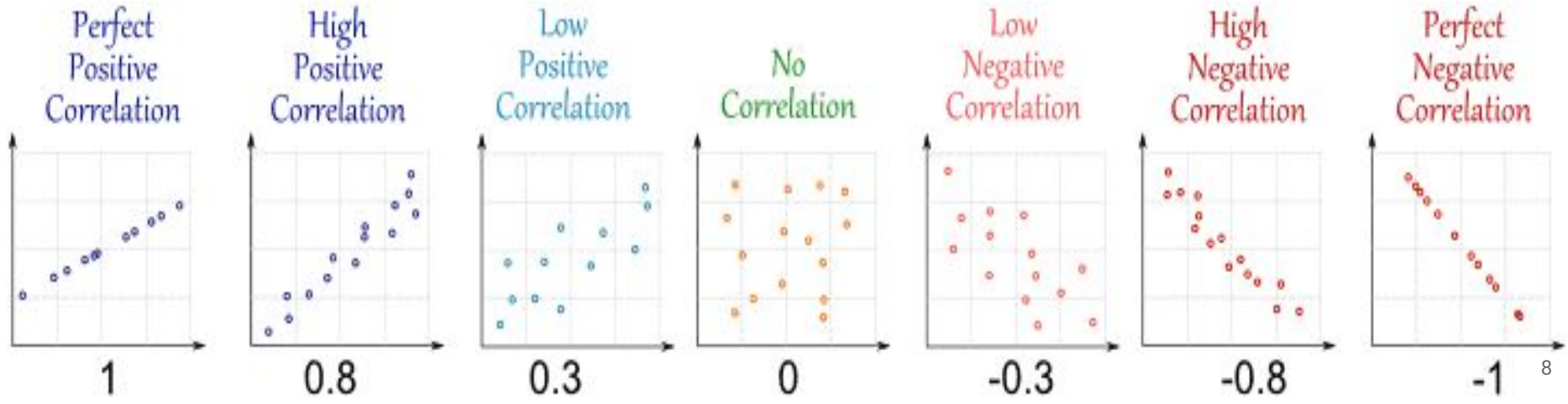
Let each sample be an *x*, *y* pair. Qualitative association patterns might include:

- Positive: *x* high and *y* high, *x* low and *y* low
- Negative: *x* high and *y* low, *x* low and *y* high
- None: *x* high and *y* high, *x* high and *y* low, *x* low and *y* low, *x* low and *y* high

These can be visualized using a *scatterplot*, in which each *x*, *y* pair is plotted on a Cartesian plane.

Ice cream sales as a function of temperature





Measuring association between interval variables...

...is done using Pearson's (1895) *product-moment correlation coefficient*, better known as *Pearson's r*.

First, let us define the *(sample) covariance* of two variables *X, Y*:

$$cov(X, Y) = \sum_i [(x_i - \bar{X})(y_i - \bar{Y})] / (n - 1)$$

i.e., a measure of how much *X* and *Y* are varying together, based on deviations from their respective means.

Relationship to variance

$$\text{cov}(X, Y) = \sum_i [(x_i - \bar{X})(y_i - \bar{Y})] / (n - 1)$$

$$\begin{aligned}\text{var}(X) &= \sum_i [(x_i - \bar{X})^2] / (n - 1) \\ &= \sum_i [(x_i - \bar{X})(x_i - \bar{X}) / (n - 1) \\ &= \text{cov}(X, X)\end{aligned}$$

Interpretation

Covariance is the sum of products of deviations.

- If $x > \bar{X}$ and $y > \bar{Y}$, the product will be positive
- If $x < \bar{X}$ and $y < \bar{Y}$, the product will be positive
- If $x > \bar{X}$ and $y < \bar{Y}$, the product is negative
- If $x < \bar{X}$ and $y > \bar{Y}$, the product is negative

Inconsistent products will cancel each other out, yielding a value near to zero.

Standardizing covariance

But covariance depends on the units of measure.

For examples the covariance of two variables measured in miles, and the covariance of those converted to kilometers, are not comparable.

Pearson's solution: standardize it by dividing by the standard deviations of both variables.

$$r = cov(X, Y) / (s_X s_Y)$$

Standardizing covariance

This has a range [−1, +1] where

- −1 is a *perfect negative correlation,*
- 0 is *no correlation*, and
- +1 is a *perfect positive correlation*.

The sign depends on how *X* and *Y* are defined. E.g.:

- Study time has a positive correlation with the # of items answered correctly.
- Study time has a negative correlation with the # of items answered incorrectly.

In some cases, we take the absolute value, a pure measure of association.

Sternberg's (1966) short-term memory scanning

Subjects are briefly shown 1-6 digits to remember (the "memory set").

After a short pause, the subject is shown a single digit (the "probe") and asked to indicate whether the digit was in the memory set or not.

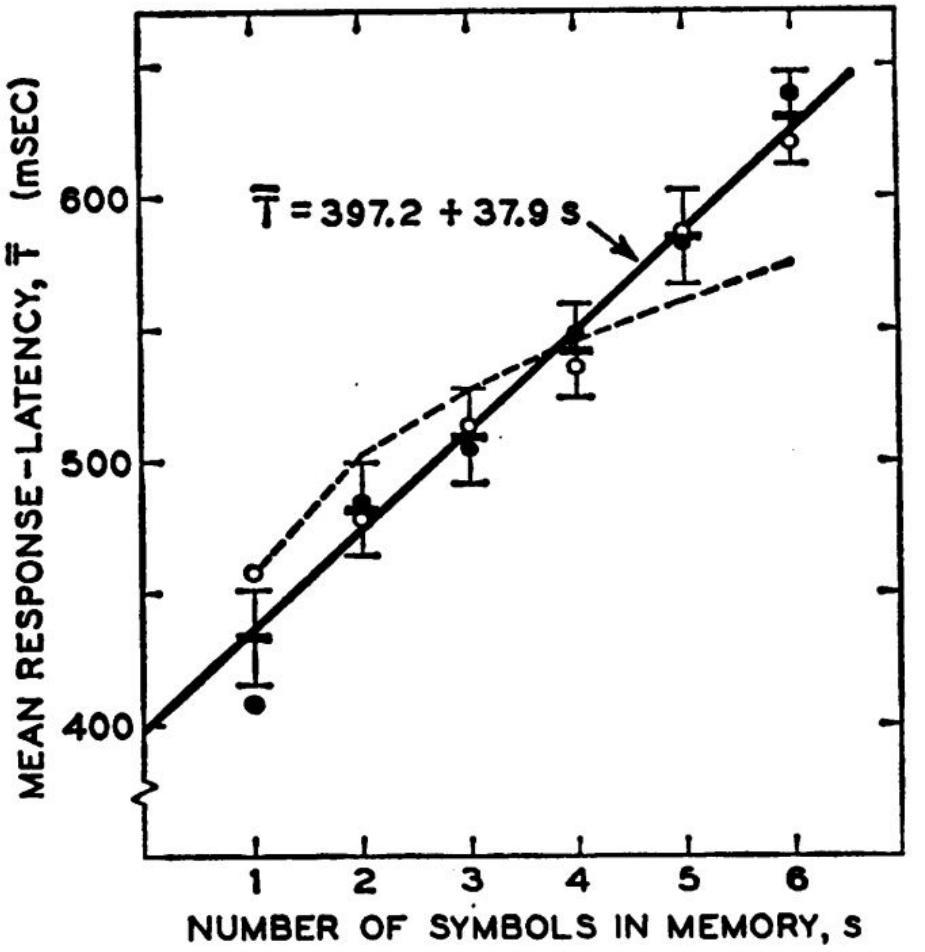
Independent variable: size of the memory set

Dependent variable: RT in seconds

Sample trial

Memory set:	3	5	2	7	9
Probe:	7				
Response:	Yes				

(Image credit: Sternberg 1966)



```
> x <- 1:6
> y <- c(.440, .515, .550, .575, .605, .615)
```

Let's do it by hand first:

```
> covariance <- cov(x, y)
> covariance / (sd(x) * sd(y))
[1] 0.9604371
```

Or, we can use other R built-ins:

```
> cor(x, y)
[1] 0.9604371
```

As a null hypothesis test

Let *R* be the correlation parameter (i.e., amount of correlation in the population)

Let *r* be the correlation statistic (i.e., amount of correlation in the sample).

Our null hypothesis is that $R = 0$.

r has a known distribution (i.e., there exist *r*-tables for any *n*), or we can convert *r* into a *t*-statistic (using a relatively simple, but opaque, formula) and use a *t*-test table.

Critical *r* values (two-tailed test)

d.f. = $n - 2$	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.284
90	.173	.205	.242	.267
100	.164	.195	.230	.254

But, R can run the whole thing for us too.

```
> cor.test(x, y)

    Pearson's product-moment correlation

data:  x and y
t = 6.8973, df = 4, p-value = 0.002317
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6750314 0.9958104
sample estimates:
      cor
0.9604371
```

```
> cor.test(x, y)

    Pearson's product-moment correlation

data:  x and y
t = 6.8973, df = 4, p-value = 0.002317
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6750314 0.9958104
sample estimates:
      cor
0.9604371
```

You can report this as simply: *r* = .960, *p* = .002.

r^2 ("r-squared")

The coefficient squared is the *coefficient of determination*, the

- proportion of variance shared by *X* and *Y*
- proportion of *Y*'s variance related to/accounted for/explained by *X*

If *r* = .1, 1% (.01) of the variance is shared.

If *r* = .9, 81% (.81) of the variance is shared.

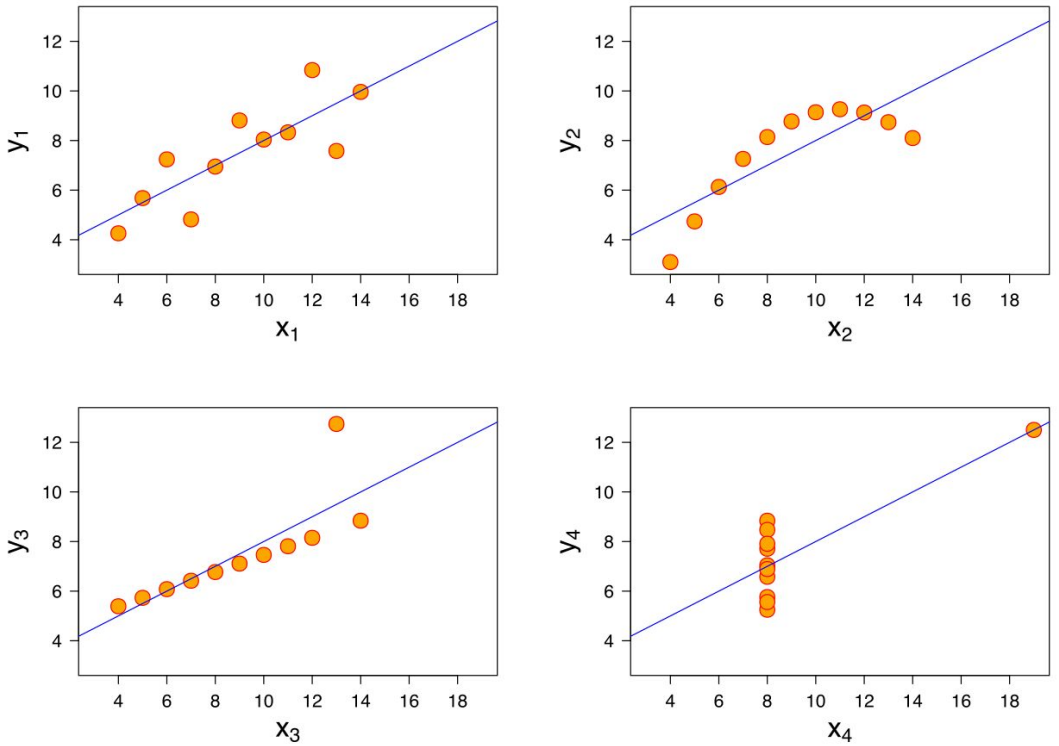
```
> r <- cor(x, y)
> round(r * r, 2)
[1] 0.92
```

Assumptions

- Both *X* and *Y* are interval or ratio measures
- Both *X* and *Y* are approximately normal
- There is a linear relationship between *X* and *Y*
- Both *X* and *Y* have the same variance
- Outliers are minimal/absent

Anscombe's (1973) quartet

$\overline{X}$:	9
$\overline{Y}$:	7.5
s_X:	11
s_y:	4.125
r:	+.816



(Image credit: Wikipedia)

What if *Y* is dichotomous?

Then we can measure the correlation using the *point-biserial correlation*.

This is mathematically equivalent to computing the Pearson correlation and re-coding one category/level of *Y* as 0 and another as 1.

Note that this makes the sign more or less arbitrary, so we often just take the absolute value.

```
> x <- runif(10)
```

Let's take ten coin flips.

```
> y <- sample(0:1, 10, replace = TRUE)
> round(abs(cor(x, y)), 2)
[1] 0.23
```

Measuring association between rank variables...

...can be done using Spearman's (1904) *rank order correlation coefficient*, better known as Spearman's ρ ("rho").

The ρ of two variables is equal to the Pearson correlation of the ranks of those variables.

While Pearson correlation measures *linearity*, Spearman measures *monotonicity*...

Rank

```
> x <- rnorm(10)
> x
 [1]  0.4403220  0.1351852  1.5537580 -1.4253280  0.3087166
1.3582023
 [7] -0.2841349 -0.1353629 -2.8563433  0.3699060
> rank(x)
 [1]  8  5 10  2  6  9  3  4  1  7
```

R has various ways to handle ties...

How to rank data

R has various ways to handle ties (controlled by the `ties.method` argument):

- Replace a series of ties with their `average` rank (the default)
- Use the rank of the `first` or `last` tied element (and don't skip ranks)
- Put them in `random` order
- Use the rank of the `min` or max tied element (and skip ranks)

Play along with this at home to get a sense...

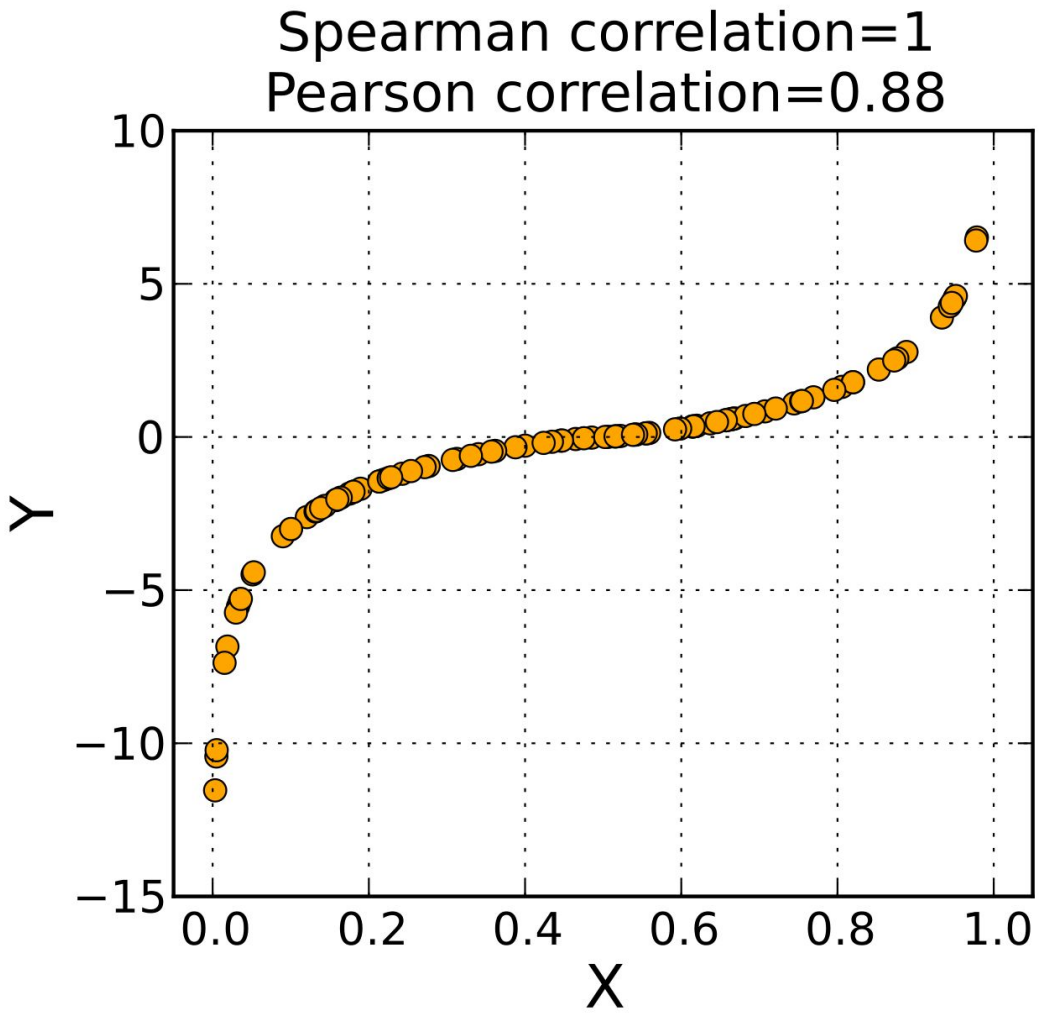
```
> x <- iris$Sepal.Length
> y <- iris$Petal.Length
```

Let's do it by hand first:

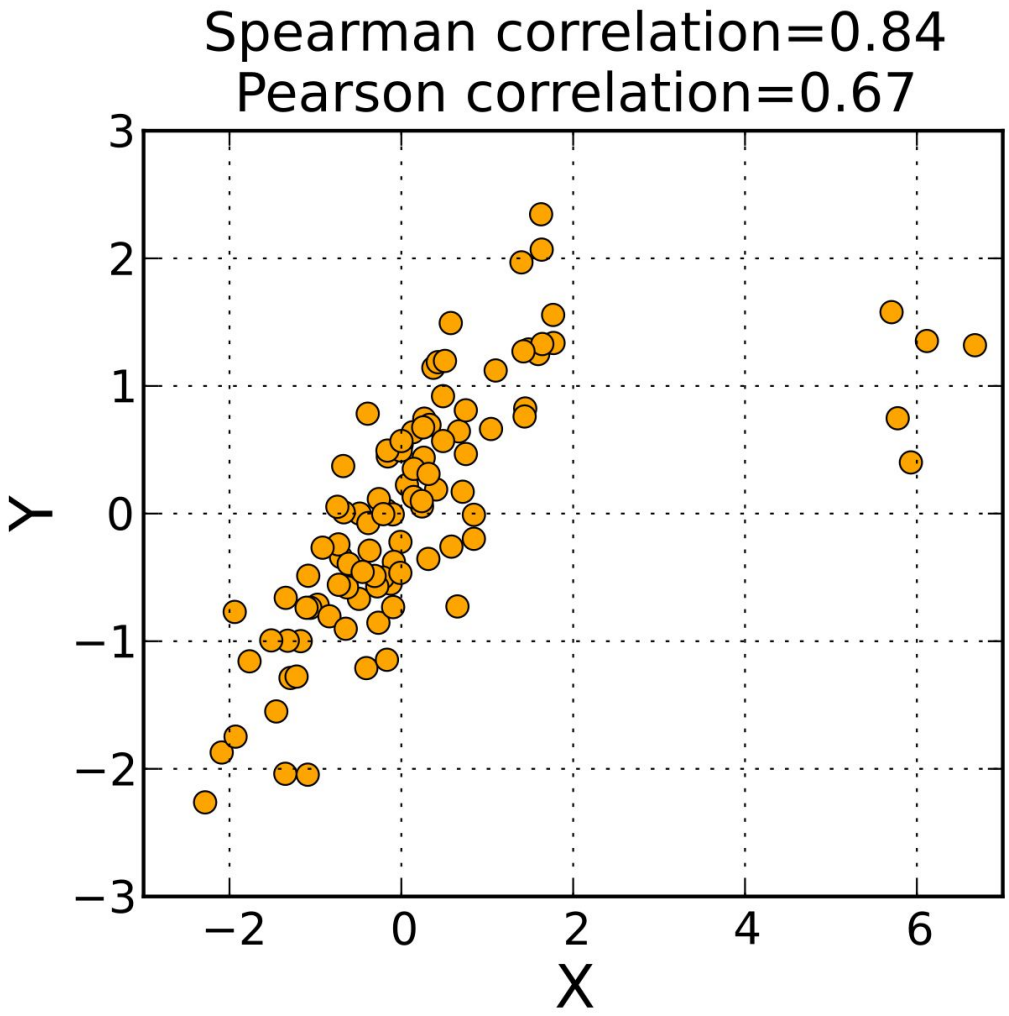
```
> x.r <- rank(x)
> y.r <- rank(y)
> cor(x.r, y.r, method = "pearson")
[1] 0.8818981
```

Or, we can use other R built-ins:

```
> cor(x, y, method = "spearman")
[1] 0.8818981
```



(Image credit: Wikipedia)



(Image credit: Wikipedia)

Compared to Pearson's *R*...

Spearman's ρ is

- less sensitive to strict linearity,
- less sensitive to extreme outliers,
- same range and sign-based interpretation, and
- uses the same statistical test (`cor.test` with `method = "spearman"`).

It is also appropriate if only one of the two variables is rank (cf. homework 04).

I generally prefer it unless there's some reason to assert a strong "interval" interpretation of the data.

Kendall (1938) and Goodman & Kruskal (1954-1972)...

...propose an alternative family of rank correlation statistics based on the notion of *concordance*.

Let us first define a function:

$$\operatorname{sgn}(x) = \begin{cases} +1 & \text{if } x > 0 \\ \text{(undefined)} & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

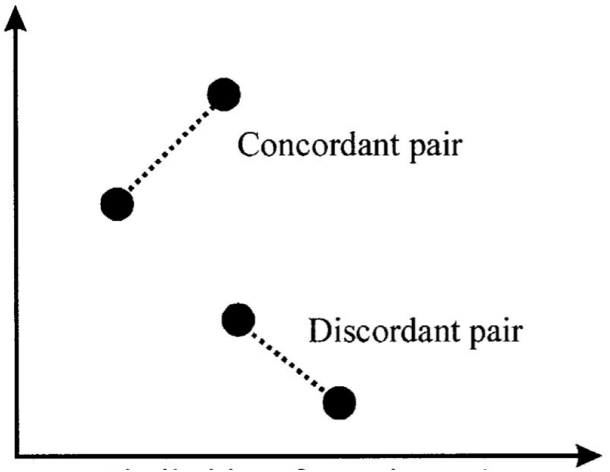
Concordance

Consider two pairs of observations x_0, y_0 and x_1, y_1. We say that this pair is *concordant* if

$$\text{sgn}(x_1 - x_0) = \text{sgn}(y_1 - y_0)$$

and *discordant* otherwise.

Concordant pairs define a line segment with a positive slope; discordant pairs define a line segment with a negative slope.



Goodman-Kruskal ɣ

For each ordered x_0, y_0 and x_1, y_1 pair, compute whether it is concordant, discordant, or neither (a "tie").

Let C be the count of concordant pairs, and let D be the count of discordant pairs.

Then,

$$ɣ = (C - D) / (C + D)$$

This also has a range [−1, +1], and a known reference distribution.

Kendall's τ_b

The Goodman-Kruskal statistic simply discards ties; τ_b applies a "correction" to the denominator which penalizes them.

Because of tie correction, for any a given X and Y, $\gamma \geq \tau_b$.

```
> x <- iris$Sepal.Length
> y <- iris$Petal.Length
> round(cor(x, y, method = "kendall"), 2)
[1] 0.72
```

Compared to Spearman's ρ...

The Kendall-Goodman-Kruskal statistics are

- computationally more expensive to compute,
- have somewhat opaque behavior with ties, and
- tend to have smaller magnitudes than the related Spearman coefficient.

But the interpretation is a bit clearer.

Questions? Please take them to email, or Slack.

This is for remote presentation…please stay safe and don't share.