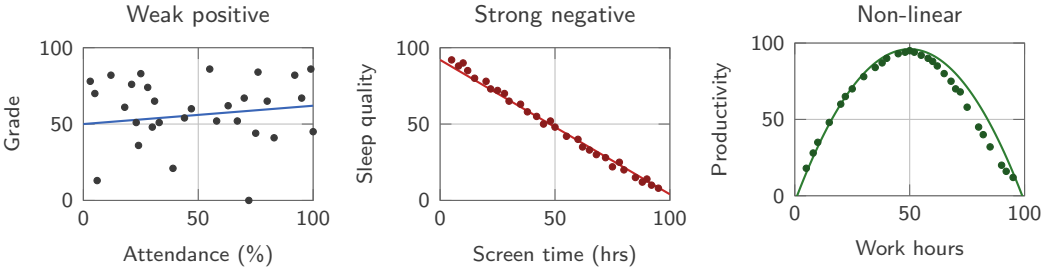
Introduction to Quantitative Research Methods: Week 8

Correlation, Crosstabs, and Seminar Exercises

James Rice

Thursday, March 12, 2026

Scatterplots



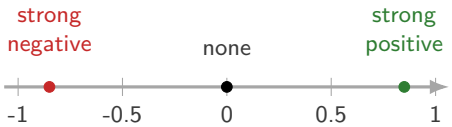
- A scatterplot is the quickest way to see whether two quantitative variables move together.
- **Upward cloud** $\rightarrow$ positive; **downward cloud** $\rightarrow$ negative; **curved cloud** $\rightarrow$ non-linear.
- The **tightness** of points around the trend tells us how strong the relationship is.
- Visual inspection matters: Pearson's r only captures *linear* patterns and can miss curves or outlier effects.

Visual cue

The left plot shows a weak linear trend with wide scatter ($r \approx 0.13$). The middle shows a tight downward pattern ($r \approx -0.99$). The right shows a clear relationship that Pearson's r would understate.

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y}$$

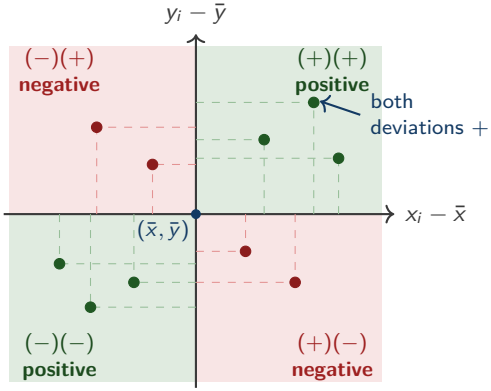
- **Direction**: positive r means high values of X tend to appear with high values of Y.
- **Strength**: the closer $|r|$ is to 1, the tighter the linear pattern.
- $r = 0$ means **no linear association**; it does *not* mean "no relationship at all."
- Where $s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ and $s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$ are the sample standard deviations of X and Y.



Interpretation

Pearson's r is a **standardised covariance**: it rescales joint movement so the statistic always lies between -1 and 1.

Covariance and deviations

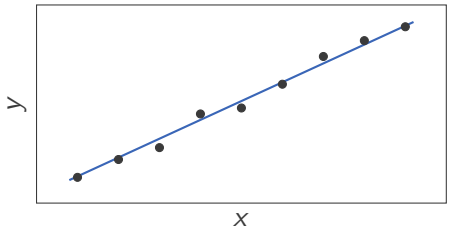


$$\mathrm{Cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- Each point's contribution is the **product of its two deviations** from the mean.
- Dashed lines show $(x_i - \bar{x})$ and $(y_i - \bar{y})$: when both are the same sign, the product is **positive**.
- When the signs differ, the product is **negative**, pulling the covariance down.
- Summing all products: if green points dominate $\Rightarrow \mathrm{Cov} > 0$; if red points dominate $\Rightarrow \mathrm{Cov} < 0$.
- Pearson's r divides by $s_X s_Y$ so the result is always in $[-1, 1]$.

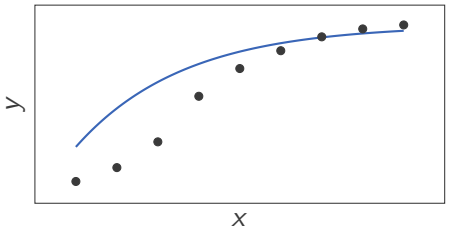
Pearson vs Spearman

Pearson: linear focus



$$r = \frac{\mathrm{Cov}(X, Y)}{s_X \, s_Y}$$

Spearman: non-linear focus



$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = \mathrm{rank}(x_i) - \mathrm{rank}(y_i)$

- Use **Pearson** when the relationship is approximately linear and both variables are quantitative.
- Use **Spearman** when the pattern is one-directional but not necessarily linear, or when one variable is ordinal.

Testing whether a correlation is real

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0$$

$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad \text{with df} = n - 2$$

- In `R`, use `cor(x, y)` for the coefficient and `cor.test(x, y)` for inference.
- A small ***p*-value** tells us that observing a sample correlation this extreme would be unlikely if the population correlation were zero.

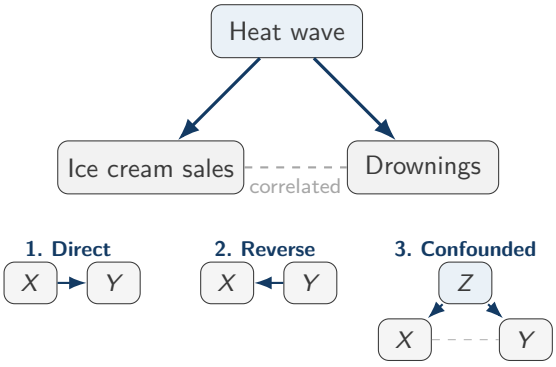
Decision rule

- If $p < 0.05$, reject H_0.
- If $p \geq 0.05$, fail to reject H_0.
- Statistical significance does not tell us whether the relationship is substantively large.

R commands

```
cor(attendance, grade)
cor.test(attendance, grade)
```

Correlation is not causation



- A correlation can arise because:
  - X causes Y (direct),
  - Y causes X (reverse), or
  - a **third variable** causes both (confounded).
- The ice cream example shows scenario 3: **heat** drives both variables, creating a spurious link.
- Correlation is a useful **first diagnostic**, but causal claims need theory, design, and additional evidence.

Key takeaway

Seeing $r \neq 0$ tells you the variables are associated. It does *not* tell you *why*. Always ask: is there a plausible confounder, or could the direction run the other way?

Seminar tasks: descriptive work and crosstabs

- Use the British Crime Survey data to examine age, `yrsarea`, `rural2`, `walkdark`, and `tcarea`.
- Report the main summary statistics and briefly describe each distribution.
- Decide whether a cross-tabulation of `rural2` and `tcarea` is sensible.
- Create a cross-tab of `rural2` (urban/rural) by `walkdark` (perceived safety walking alone after dark), including row percentages.
- Identify the **response variable** and the **explanatory variable**, then interpret the table substantively.

Core R command

```
with(british_crime, CrossTable(rural2, walkdark,
  prop.chisq=FALSE, prop.c=FALSE,
  prop.t=FALSE, format=c("SPSS")))
```

This produces a cross-tabulation with **row percentages only**, suppressing chi-square contributions, column proportions, and table proportions for a cleaner output.

Seminar tasks: hypothesis test and follow-up

- State the hypotheses for a **chi-square test of independence**.
- Report the chi-square statistic, degrees of freedom, and p-value.
- Test the **strength** of the relationship using Cramer's V.
- Subset the sample to respondents aged 90 or over and rerun the cross-tab.
- Check expected cell counts; if they are too small, switch from chi-square to **Fisher's exact test**.

Follow-up R commands

```
assocstats(table(british_crime$rural2,
  british_crime$walkdark))
british_aux <- subset(british_crime, age >= 90)
fisher.test(table(british_aux$rural2,
  british_aux$walkdark))
```

Seminar answers: Q1 summary statistics and variable choice

R code

```
summary(british_crime$age)
summary(british_crime$tcarea)
prop.table(table(
 british_crime$walkdark))
prop.table(table(
 british_crime$rural2))
prop.table(table(
 british_crime$yrsarea))
```

- **Age**: median = 49, mean = 50.42, range 16 to 101.
- **tcarea**: mean close to 0 and many distinct values, so it behaves like a **scale variable**.
- **walkdark**: about 66% report feeling *fairly safe* or *very safe*.
- **rural2**: about 25.5% rural and 74.5% urban.

Conclusion

It is **not** wise to cross-tabulate `rural2` with `tcarea` because `tcarea` is essentially continuous. Crosstabs are designed for variables with a manageable number of categories.

Seminar answers: Q2 crosstab, response variable, and hypotheses

R code

```
with(british_crime, CrossTable(
 rural2, walkdark,
 prop.chisq=FALSE, prop.c=FALSE,
 prop.t=FALSE, format=c("SPSS")))
```

- **Response variable**: `walkdark`.
- **Explanatory variable**: `rural2`.
- We compare how the **distribution of perceived safety** changes across rural and urban settings.

Hypotheses

H_0: `rural2` and `walkdark` are independent.

H_A: the distribution of `walkdark` differs between rural and urban respondents.

Interpretation

The crosstab asks whether feeling safe after dark is distributed similarly in rural and urban areas. If the row percentages differ noticeably, that is evidence against independence.

Seminar answers: Q3–Q4 chi-square test and Cramer's V

R code

```
assoc <- assocstats(table(
 british_crime$rural2,
 british_crime$walkdark))
```

Key output

Pearson $\chi^2 = 629.30$
$df = 3$
$p < 4.479 \times 10^{-136}$
Cramer's $V = 0.233$

- Because the p-value is far below 0.05, we **reject** H_0.
- There is a statistically significant association between area type and perceived safety after dark.
- Cramer's $V = 0.233$ suggests a **moderate** relationship: the association is real, but not overwhelmingly large.
- Substantively, rurality appears related to how safe respondents feel, but it is only one part of the story.

Seminar answers: Q5 respondents aged 90+ and Fisher's exact test

R code

```
british_aux <- subset(british_crime,
age >= 90)
contingency_table <-
table(british_aux$rural2,
              british_aux$walkdark)
fisher.test(contingency_table)
```

- In the 90+ subsample, **two expected counts fall below 5**.
- That violates a key assumption for the chi-square approximation.
- So the better choice is **Fisher's exact test**.

Result and interpretation

Fisher's exact test gives $p = 0.0001776$. We therefore conclude that, even among respondents aged 90 or over, there is still a statistically significant association between whether the area is rural or urban and how safe people feel walking alone after dark.