DSA 8010 - Exploratory Data Analysis Part 2

Exploring associations

Association

Two variables X and Y are associated if knowing the value of X gives some information about likely values of Y.

Association

Two variables X and Y are associated if knowing the value of X gives some information about likely values of Y.

Example 1: Undergraduate GPA is typically associated with whether a student graduates *cum laude*. There is probably association between the variables because *cum laude* status is determined by GPA.

Example 2: Is undergraduate GPA associated with SAT score? If so, students with high test scores tend to be more successful in college coursework.

Association

Two variables X and Y are associated if knowing the value of X gives some information about likely values of Y.

Associated variable are sometimes also called related or dependent.

Variables that are not associated are sometimes also called unrelated and independent.

Measures of association between numeric variables

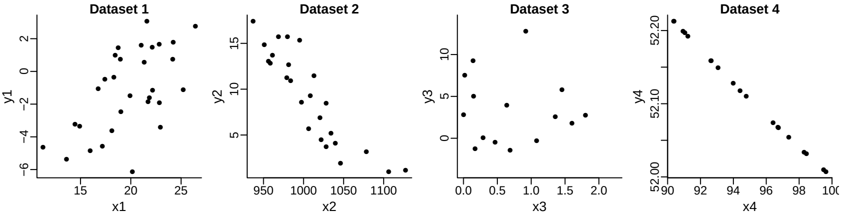
Scatterplots

A scatterplot is used to investigate association between two quantitative variables. The value of one variable is shown on the x-axis and the other on the y-axis.

Three main features of a scatterplot:

1. Strength - how closely points follow a pattern.
2. Pattern - linear, curved, etc.
3. Direction - positive: as one variable increases, the other tends to increase. Negative: as one variable decreases, the other tends to decrease.

Scatterplots



Pearson's correlation

- $-1 \leq r \leq 1$
- Positive $r \rightarrow$ positive relationship; negative $r \rightarrow$ negative relationship.
- large $|r| \rightarrow$ the relationship is strong.

  $0.75 \leq |r| \leq 1$
- small $|r| \rightarrow$ relationship is weak.

  $0 \leq |r| \leq 0.25$
- r does not have units and does not change if you change the units of measurement. (e.g., converting degrees Fahrenheit to Celsius or convert miles to kilometers.)

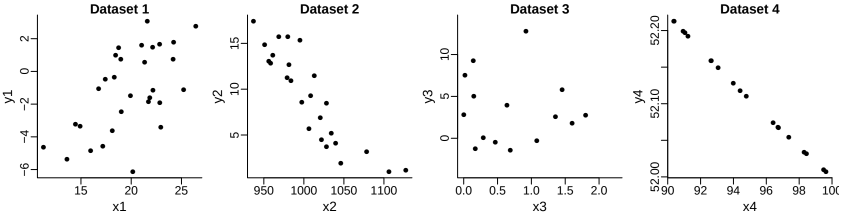
Pearson's correlation coefficient

Pearson's correlation coefficient (r) is calculated as

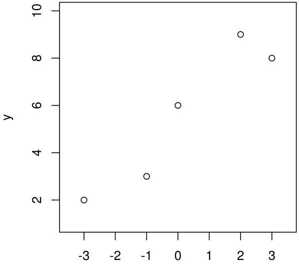
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson's correlation measures the strength and direction of the linear relationship between two quantitative variables.

Pearson's correlation coefficient



Find Pearson's correlation between X and Y.



X	Y
-1	3
3	8
2	9
0	6
-3	2

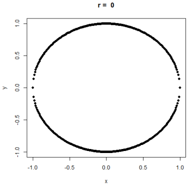
Find Pearson's correlation between X and Y.

X	Y
-1	3
3	8
2	9
0	6
-3	2

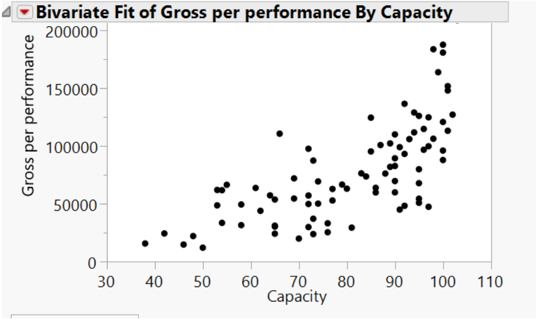
Cautions regarding r

Some cautions when using r:

- r is sensitive to outliers.
- r is only accurate when measuring the strength of *linear* relationships.
  Example:

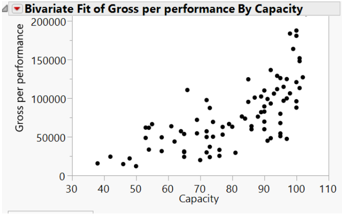


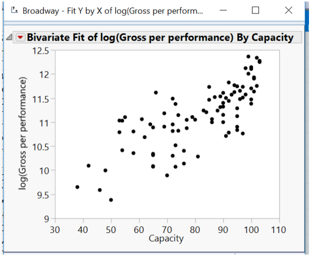
Nonlinear associations



Nonlinear associations: transformations

One way to address nonlinear associations is to transform one or more variable. To transform some variable x, apply a one-to-one function f to x. Then continue with the analysis using $f(x)$.





Frequently used transformations include the (natural) log and square root.

Spearman's rank correlation

If X and Y have a nonlinear but monotonic relationship, Spearman's rank correlation may be a more appropriate way to measure the association.

To calculate Spearman's rank correlation, first record the rank of each observed X as an integer between 1 and n. Do the same for Y.

Then calculate Pearson's correlation between the ranks rather than the original variables.

Find Spearman's rank correlation between X and Y.

X	Y
-1	3
3	8
2	9
0	6
-3	2

Scatterplots with more than 2 variables

Consider using color, plotting symbol, or plotting size to display additional variables in a scatterplot.

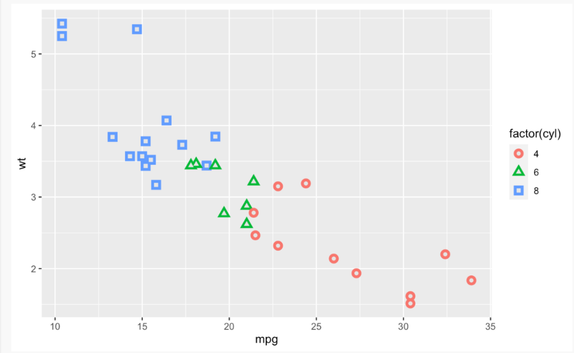
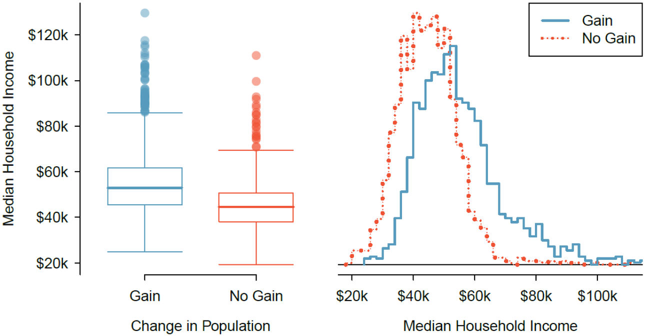


Image of the `mtcars` data set from ggplot2 documentation (`https://ggplot2.tidyverse.org/reference/geom_point.html`).

Measures of association with categorical variables

One categorical variable

When investigating association between one categorical variable and one numeric variable, compare plots and summary statistics across groups defined by the categorical variable.



Open Intro Statistics, 4th edition, Diez et al

One categorical variable

When investigating association between one categorical variable (X) and one numeric variable (Y), compare plots and summary statistics across groups defined by the categorical variable.

X and Y are associated if the distribution of Y differs across the groups determined by X.

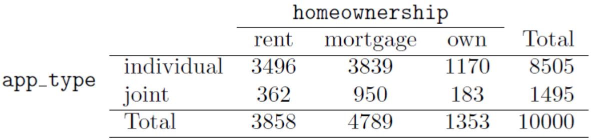
Contingency tables

marginal distribution. Distribution of one variable, ignoring all other variables. This distribution tells you the possible values of X and how often X takes those values.

joint distribution. Distribution of two (or more) variables combined. This distribution gives all possible combinations of X and Y and how often each combination occurs.

Contingency tables

The following table summarizes two variables recorded from a set of loan applications: the application type and the homeownership status of the applicant.

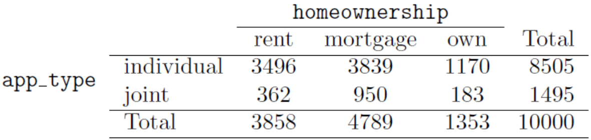


		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Example from *Open Intro Statistics, 4th edition, Diez et al*.

Contingency tables

Divide the counts in the frequency table by the overall total to obtain proportions summarizing the joint distribution.



		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Example from *Open Intro Statistics, 4th edition, Diez et al*.

Contingency tables

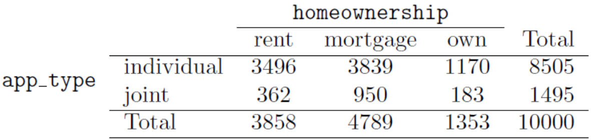
Joint distribution of `homeownership` and `app_type`.

	rent	mortgage	own	Total
individual	0.3496	0.3839	0.1170	0.8505
joint	0.0362	0.0950	0.0183	0.1495
Total	0.3858	0.4789	0.1353	1

What does the number 0.0362 represent?

Contingency tables

Divide the counts in the frequency table by the row total to display row proportions. (This is also referred to as a conditional distribution.)



		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

Contingency tables

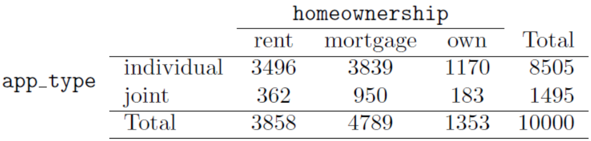
Contingency table with row proportions:

	rent	mortgage	own
individual	0.4111	0.4514	0.1376
joint	0.2421	0.6355	0.1224

What does the number 0.2421 represent?

Contingency tables

Divide the counts in the frequency table by the column total to display column proportions. (This is also referred to as a conditional distribution.)



		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

*Note: It is more conventional to display row proportions than column proportions.

Contingency tables

Contingency table with column proportions:

	rent	mortgage	own
individual	0.9062	0.8016	0.8647
joint	0.0938	0.1984	0.1353

What does the number 0.0938 represent?

Association in contingency tables

The row variable (X) is associated with the column variable (Y) if the rows have different row proportions. Equivalently, if the columns have different column proportions.

Association in contingency tables

Do the data suggest that home ownership is associated with loan type?

Table with row proportions:

	rent	mortgage	own
individual	0.4111	0.4514	0.1376
joint	0.2421	0.6355	0.1224

Association in contingency tables

Do the data suggest that home ownership is associated with loan type?

Table with row proportions:

	rent	mortgage	own
individual	0.4111	0.4514	0.1376
joint	0.2421	0.6355	0.1224

It appears that there is some association between the variables because home_ownership has a different distribution between the individual and joint applicants. Applicants who file as individuals are more likely to rent their homes (41.11%) than those who file jointly (24.21%), while joint applicants are more likely (63.55%) to mortgage their homes than individual applicants (45.14%). The proportion of applicants who own their homes is similar across joint and individual applicants.

Interpreting associations

Interpretation of associations

- Correlation is not causation.
- Relationships among unobserved variables may be more meaningful than observed association.
- Make sure to correctly interpret associations among aggregate variables.
- In real data, relationships are observed with noise.

Example: sunscreen

Suppose an observational study tracked sunscreen use and skin cancer. It was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean that sunscreen causes skin cancer?

Open Intro Statistics, 4th edition, Diez et al

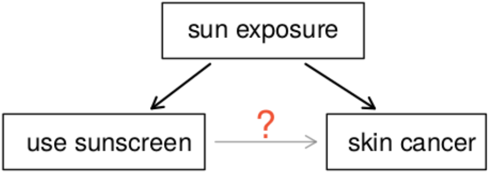
Example: sunscreen

Suppose an observational study tracked sunscreen use and skin cancer. It was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean that sunscreen causes skin cancer?

Open Intro Statistics, 4th edition, Diez et al

Example: sunscreen

Suppose an observational study tracked sunscreen use and skin cancer. It was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean that sunscreen causes skin cancer?



Open Intro Statistics, 4th edition, Diez et al

Confounding (lurking) variables affect the response but are not measured. Confounding can lead us to mistakenly attribute a causal effect to one of the measured variables.

Example (1930 census)

Data from the 1930 census show a negative correlation ($r = -0.53$) between states' literacy rates and nativity rates (percentage of residents who were born in the U.S.).

From *Ecological correlations and the behavior of individuals* (Robinson, 1950).

Example (1930 census)

Data from the 1930 census show a negative correlation ($r = -0.53$) between states' literacy rates and nativity rates (percentage of residents who were born in the U.S.).

At the individual level, there was a weak positive correlation ($r = 0.12$) between literacy and nativity.

From *Ecological correlations and the behavior of individuals* (Robinson, 1950).

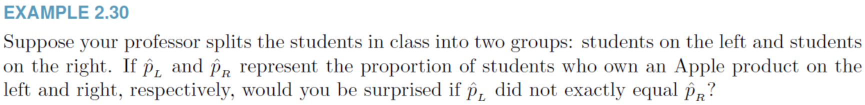
Example (1930 census)

Data from the 1930 census show a negative correlation ($r = -0.53$) between states' literacy rates and nativity rates (percentage of residents who were born in the U.S.).

At the individual level, there was a weak positive correlation ($r = 0.12$) between literacy and nativity.

Associations among aggregate variables might be absent, or even have a different direction, at the individual elvel.

This is very similar to a phenomenon called Simpson's paradox.



EXAMPLE 2.30

Suppose your professor splits the students in class into two groups: students on the left and students on the right. If $\hat{p}_L$ and $\hat{p}_R$ represent the proportion of students who own an Apple product on the left and right, respectively, would you be surprised if $\hat{p}_L$ did not exactly equal $\hat{p}_R$?

Intro Statistics, 4th edition, Diez et al.

Interpretation of associations

- Correlation is not causation. Carefully-designed experiments are the gold standard for inferring a causal effect of some treatment.
- Relationships among unobserved variables may be more meaningful than observed association. Consider ”lurking variables” and confounding variables that might be affecting the response variable.
- Make sure to correctly interpret associations among aggregate variables. Associations at the individual level need not match associations at an aggregate level.
- In real data, associations are observed with noise.