Chapter Seven

Linear Correlation

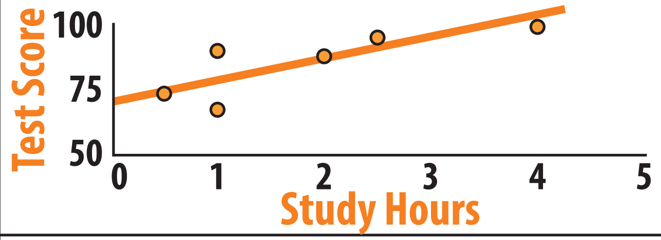
The Line of Best Fit

Statistically Significant Correlations

Causal Relationships

Linear Regression

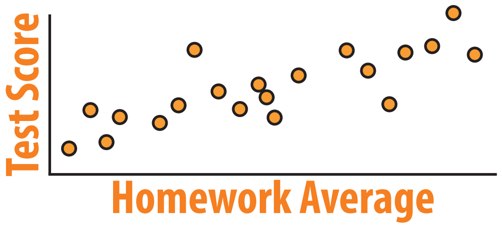
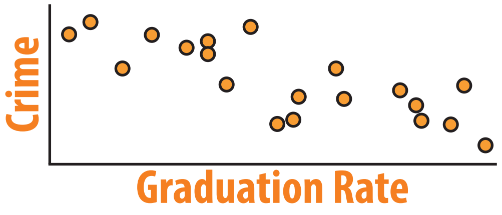
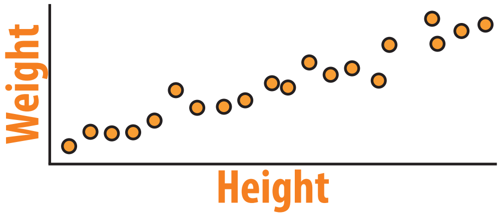
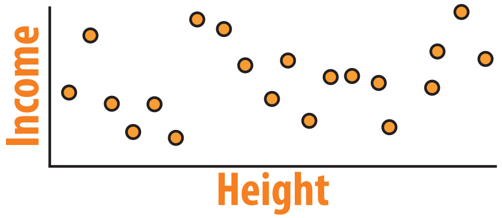
Linear regression is finding the line that best fits a data set. It is used to predict future data values.

Concept	Definition	Example
Paired Data	data values such that each data point has an *x* value and a *y* value	Study hours: 0.5 1.0 1.0 2.0 2.5 4.0 Test score: 74 68 90 88 95 99
Line of Best Fit	the line that best fits the paired data (which must be numerical) by having the smallest sum of squared residuals (see notes for directions on how to calculate it)	$\hat{y} = 7.46x + 72.0$ 
Interpolation	predicting a *y* value for an *x* value between the lowest and highest *x* value in the data set	The predicted test score of a student who studied three hours is $\hat{y} = 7.46(3) + 72.0 = 94$.
Extrapolation	predicting a *y* value for an *x* value below the lowest *x* value or above the highest *x* value in the data set	The predicted test score of a student who studied ten hours is $\hat{y} = 7.46(10) + 72.0 = 147$.*
Residual	the amount a *y* value is higher than would have been predicted by the line of best fit	The residual for the student who studied for two hours is $88 - (7.46(2) + 72.0) = 88 - 87 = 1$

* Extrapolation commonly leads to unrealistic predictions and should be used with caution.

The Correlation Coefficient

The **correlation coefficient** ***r*** is a value between -1 and 1 that summarizes the strength and direction of the relationship between the two variables in the sample.

Value of *r*	Correlation	Meaning	Example
Positive	positive	The higher *x* is, the higher *y* tends to be.	
Negative	negative	The higher *x* is, the lower *y* tends to be.	
Close to 1 or -1	strong	The two variables are closely related in the sample, so the line of best fit is a good predictor.	
Close to 0	weak	The two variables are not closely related in the sample, so the line of best fit is a poor predictor.	

Samples and Populations

Data are collected from **samples**, but researchers want to know about entire **populations**.

Term	Definition	Example
Population	the group being studied	Americans
Sample	the subset of the population from which data are actually collected	300 students in an introductory college chemistry class

Ideally, sample data fairly represent the overall population so that conclusions about the population can be made from the sample. Such conclusions may be limited, however, by both random and systematic error.

Cause of Error	Type	Definition	Example
Nonrandom Selection	systematic	the sample is not randomly drawn from the population	College students have more motivation, intelligence, and parental support than the average American.
Coincidence	random	the sample is small enough that a few values that do not fit the trend in the population lead to a misleading conclusion	Coincidentally, several people in the sample exercise a lot but have high blood pressure.

P Values

The ***P* value** of a sample is the probability that another random sample of the same size would, coincidentally, show at least as strong of a result in the hypothesized direction. Coincidence are common in small samples, so large *p* values are common for small samples. Some examples are shown below.

Hypothesis	**Result**	***P* value**	**Meaning**
Coins land on tails more than heads.	9 out of 15 coins land on tails.	$p = .15$	If coins really land on tails exactly half the time, another sample of 15 coins would still have a 15% chance of getting at least 9 tails.
People watch more than 20 hours of TV per week on average.	6 people watched an average of 24 hours of TV per week, with standard deviation 10 hours.	$p = .19$	If people really watch an average of only 20 hours of TV per week, another sample of 6 people would still have a 19% chance of averaging at least 24 hours.
Students with higher homework scores get higher test scores.	The correlation coefficient for the students in last year's class was $r = .69$.	$p < .01$	If homework scores really are not correlated with test scores, there is less than a 1% chance that another sample of 36 students would also get correlation coefficient as high as $r = .69$.

Statistical Signifigance

It is impossible to determine from a sample whether or not the result applies to the population as well, as opposed to being a coincidence. However, the lower the *p* value is, the less likely the results are coincidental. If **$p < .05$**, the results are considered **statistically significant**, and the researchers conclude that their hypothesis was correct not only for their sample (which they know) but also for the population overall (which they could be wrong about).

In the example below, researchers collect a sample of children to see if their grades tend to be higher the more they play outdoors, and although they find their hypothesis to be correct in their sample, it is possible that this is a coincidence there really is no correlation among children in general.

***p* value**	**Conclusion**	**Possible Error**
below 5%	Children tend to get higher grades the more they play outdoors.	This was true coincidentally for the children in their sample, but there is no such correlation for children in the population overall.
above 5%	There's not sufficient evidence to conclude that children tend to get higher grades the more they play outdoors.	In the population overall, children really do tend to get higher grades the more they play outdoors, but the sample was too small for the researchers to conclude this based on their data.

Data Snooping

Sample data can be used to form hypotheses or to test hypotheses, but when the data used to test a hypothesis were the same data used to come up with it, the researcher has gone in a circle and the *p* value will be meaningless. ***P* values only have meaning for hypotheses that were specifically stated prior to knowing the data.** Searching for any possible pattern within a data set is called **data snooping**, and it is a common fallacy among people unfamiliar with research methods to assume that any found pattern is likely to be legitimate rather than coincidental.

Correlation was predicted?	Conclusion	Action
yes	The correlation likely does exist in the population overall.	Claim that the hypothesis was correct.
no	The correlation exists in the sample, but there is no evidence that it exists in the population overall.	Do not make a claim. However, if there is a reasonable explanation for the discovered correlation, it can be tested in a new sample, and a claim can be made based on the new data.

Correlation and Causation

When a correlation between two variables is found, it is tempting to conclude that correlation is due to one variable affecting the other. However, in many cases, the correlation is partially or entirely due to outside variables affecting the independent and dependent variable simultaneously. A **confounding variable** is one that affects the dependent variable and is correlated with, but not affected by, the independent variable. Because confounding variables can provide alternative explanations for why one variable is correlated with another, **correlation does not imply causation**: Knowing that a correlation exists is not the same as knowing *why* it exists, such as in the examples below.

Correlation	Presumed reason	Confounding variables	Alternative explanation for correlation
years of education & annual salary	More schooling provides more skills, understandings, and opportunities for better jobs.	motivation, intelligence	People who are smart and motivated are more likely to get into college, but are also more likely to get a high paying job than people who are unintelligent and lazy, regardless of whether or not they go to college.
exercise & health	Exercising is good for your health.	caring about health	People who care about their health are more likely to exercise, but they are also more likely to do other healthy things such as have a good diet.

Causes of a sample correlation

There are four main categories of explanations for why a correlation exists in a sample. Multiple reasons may apply to a single correlation. For example, a professor may find that the closer students choose to sit to the front of the room in his calculus class, the better their grade tends to be.

Category	Definition	Possibility for seating example
Coincidence	The results in the sample coincidentally do not represent the population. Another sample would probably have much different results.	It is coincidentally the case that students sitting close to the front tend to have higher grades in this particular class, but these results would not occur in most other classes.
Causation	The independent variable affects the dependent variable, as assumed.	Sitting in front causes students to see and hear better and thus get a good grade.
Reverse Causation	The dependent variable affects the independent variable.	Doing well in the class causes students to be interested and want to sit in front.
Confounding Variables	Outside variables affect the dependent variable in the way that the independent variable was expected to affect the dependent variable.	Being motivated causes students to sit up front, and it also causes students to study, which is the actual reason why they have higher grades.

Usefulness of sample data

Four main factors determine how well findings in sample data can be applied to the population. For example, a professor may find that the closer students choose to sit to the front of the room in his calculus class, the better their grade tends to be.

Factor	Limitation without it	Possibility in seating example
Appropriate sampling	The population represented by the sample is only a subset of the entire population.	The trend is true for advanced classes like calculus, but not for most classes.
Strong correlation	The trend is real, but the size of the effect is small.	People who sit in front tend to score higher, but the difference is so small that it has no practical significance.
Statistical significance	The trend could easily be coincidental in the sample and not true for the population overall.	People sitting in front in this class coincidentally were the ones who scored higher.
Free from counfounding variables	The trend is true in the population overall, but not for the reason believed.	People who sit in front tend to score higher, but not because they sit up front.

Affect and Effect

Discussions of causation frequently use forms of the words *affect* and *effect*.

Word	Part of speech	Clarification	Examples
Affect(s)	verb	has a subject, which is usually one of the following: • an independent variable such as *age* • a confounding variable such as *socioeconomic status*	Smoking affects health. Childhood experiences affect adult personality.
Effect(s)	noun	usually preceded by one of the following: • the article *the* or *an* • an adjective such as *significant* or *two* • a possessive such as *religion's* or *its*	Alcohol has multiple effects. The data demonstrate music's effect on concentration.