Data Analytics and Mining

Hypothesis testing, Chi-square test, Correlation

Data Analytics and Mining, 2024

Majid Sohrabi
National Research University Higher School of Economics

November 1, 2024

Hypothesis Testing

Hypothesis test

A **hypothesis test** is a statistical method that uses sample data to evaluate a hypothesis about a population.

- **Step 1. Posing a hypothesis**

  First, **we state a hypothesis about a population**. Usually, the hypothesis concerns the value of a population parameter.

  For example, we might hypothesize that HSE students on average drink more coffee per week during the examination period compared to the rest of the academic year.

- **Step 2. Defining sample characteristics**

  Before we select a sample, we use the hypothesis to predict the characteristics that the sample should have.

  For example, if we predict that a population of HSE students on average drink 300 ml more coffee during the examination week. Then we would predict that our selected sample also should have a mean increase around 300 ml.

Remember:

The sample should be similar to the population, but you always expect a certain amount of error!

Hypothesis test

- **Step 3. Selecting a sample**

  Next, we obtain a random sample from the population. For example, we might select a sample of n = 200 HSE students and measure their regular coffee consumption and their coffee consumption during the examination week, finding the difference.

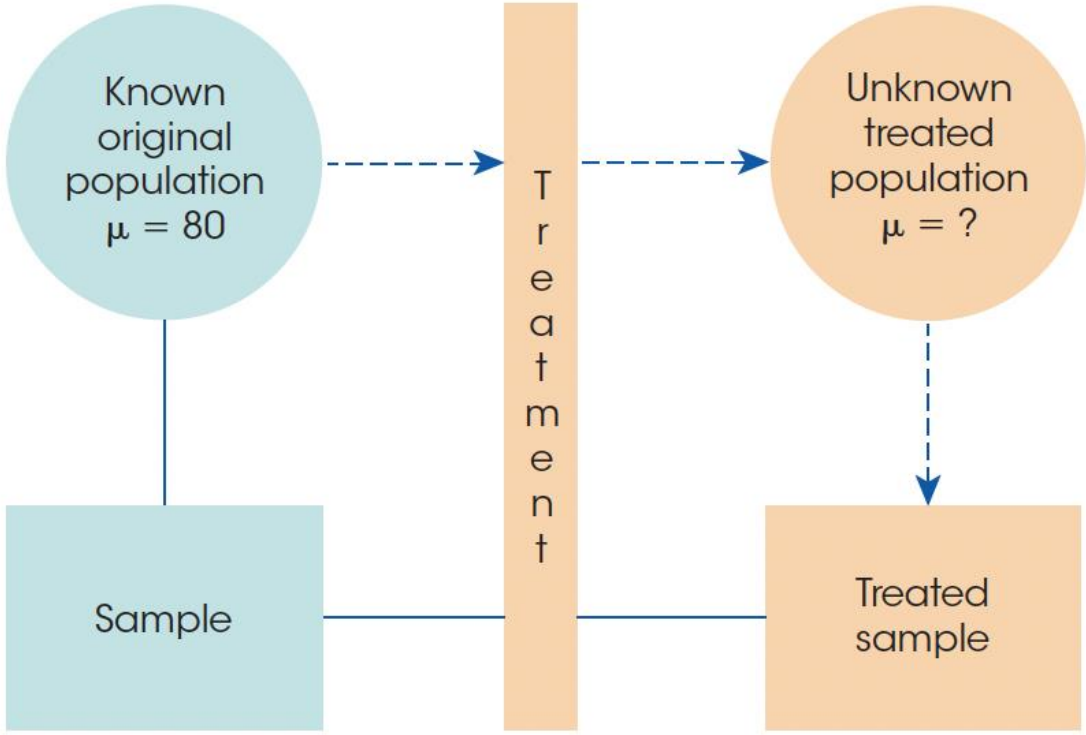
- **Step 4. Making a conclusion**

  Finally, we compare the obtained sample data with the prediction that was made from the hypothesis. If the sample mean is consistent with the prediction, we conclude that the hypothesis is reasonable.

  But if there is a big discrepancy between the data and the prediction, we decide that the hypothesis is wrong.

Why do we select a sample?

FIGURE 8.3
From the point of view of the hypothesis test, the entire population receives the treatment and then a sample is selected from the treated population. In the actual research study, however, a sample is selected from the original population and the treatment is administered to the sample. From either perspective, the result is a treated sample that represents the treated population.



Posing a hypothesis

The first and most important of the two hypotheses is called the null hypothesis. **The null hypothesis states that the treatment has no effect**.

In general, the null hypothesis states that there is no change, no effect, no difference—nothing happened, hence the name null.

The null hypothesis is identified by the symbol H_0 (The H stands for hypothesis, and the zero subscript indicates that this is the zero-effect hypothesis.)

$$H_0: \mu_{\text{examination week coffee consumption increase}} = 0$$

The alternative hypothesis (H_1) states that there is a change, a difference, or a relationship for the general population. H_1 predicts that the independent variable does have an effect on the dependent variable.

$$H_1: \mu_{\text{examination week coffee consumption increase}} = 300$$

Set criteria for decision

To formalize the decision process, we use the null hypothesis to predict the kind of sample mean that ought to be obtained.

Specifically, we determine exactly which sample means are consistent with the null hypothesis and which sample means are at odds with the null hypothesis.

To determine exactly which values are “near” and which values are “very different from”, we will examine all of the possible sample means that could be obtained if the null hypothesis is true.

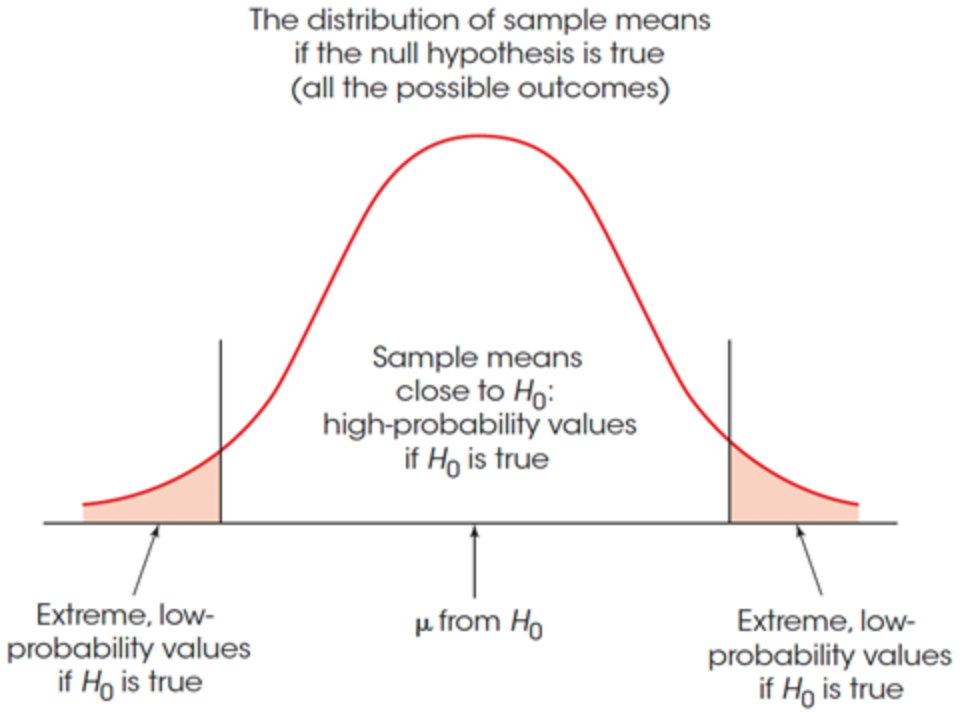


FIGURE 8.4
The set of potential samples is divided into those that are likely to be obtained and those that are very unlikely to be obtained if the null hypothesis is true.

Distribution of sample means

Sample	Scores		Sample Mean
	First	Second	(*M*)
1	2	2	2
2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

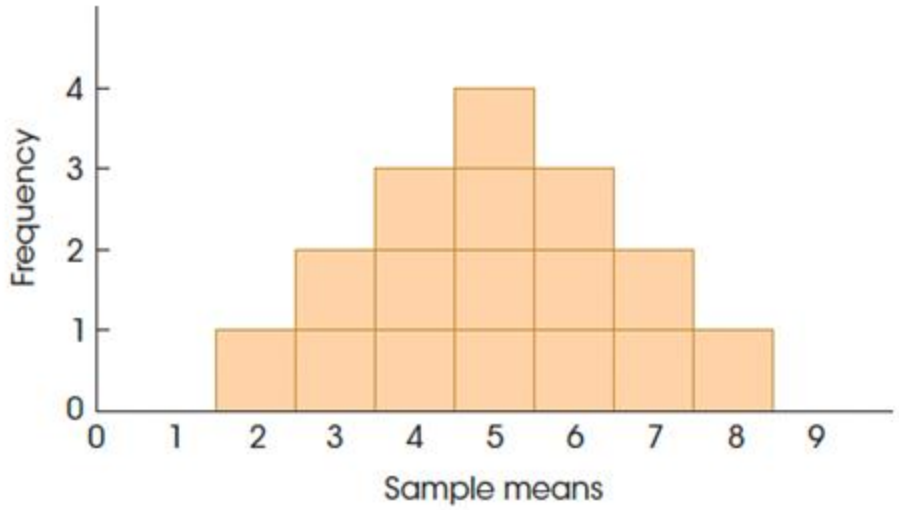
Consider a population that consists of only 4 scores: **2, 4, 6, 8**.

We are going to use this population as the basis for constructing the distribution of sample means for n = 2. For this example, there are 16 different samples.

Next, we compute the mean, M, for each of the 16 samples.

Thus, we simulate the situation of all the means we could get if we were to randomly obtain ONE sample from our population.

Distribution of sample means



Now, looking at sample means distribution, we can ask the questions about probability of obtaining a sample with a particular mean.

For example, if you take a sample of n = 2 scores from the original population, what is the probability of obtaining a sample with a mean greater than 7?

$$\boldsymbol{p(M > 7) = ?}$$

Since only one sample has a mean greater than 7, the probability is 1/16 (very unlikely!)

On the other hand, it is **VERY LIKELY** to obtain the sample mean that is equal or very close to the population mean.

$$\boldsymbol{p(M >= 4\ AND\ M <= 6) = 10/16}$$

The mean of the sample means distribution is the same as the population mean.

The Alpha Level

To find the boundaries that separate the high-probability samples from the low-probability samples, we must define exactly what is meant by “low” probability and “high” probability.

This is accomplished by selecting a specific probability value, which is known as the level of significance, or the alpha level, for the hypothesis test.

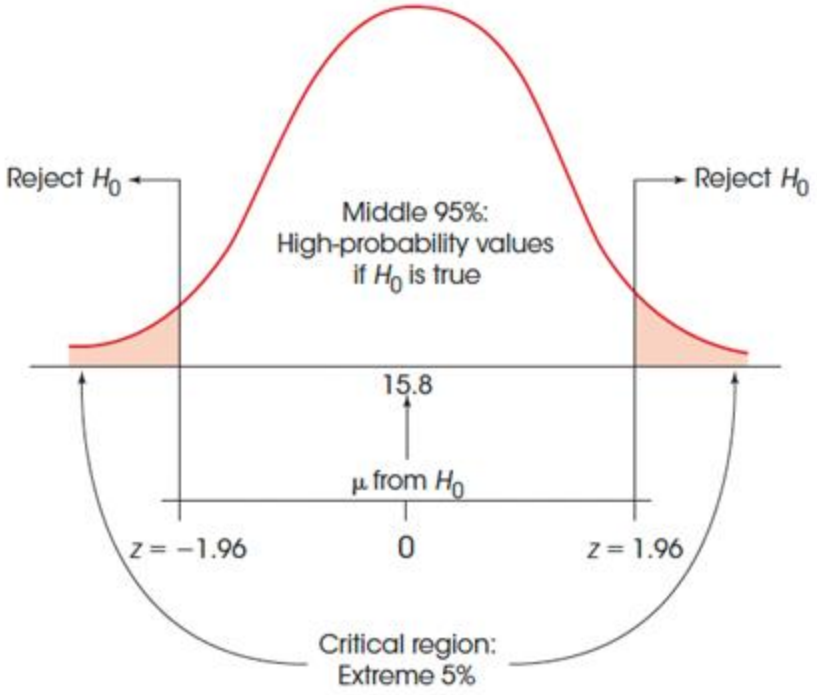
The alpha (α) value is a small probability that is used to identify the low-probability samples. By convention, commonly used alpha levels are α = .05 (5%), α = .01 (1%), and α = .001 (0.1%).

For example, with α = .05, we separate the most unlikely 5% of the sample means (the extreme values) from the most likely 95% of the sample means (the central values).

Critical Region

The **alpha level**, or the level of significance, is a probability value that is used to define the concept of "very unlikely" in a hypothesis test.

The **critical region** is composed of the extreme sample values that are very unlikely (as defined by the alpha level) to be obtained if the null hypothesis is true. The boundaries for the critical region are determined by the alpha level. If sample data fall in the critical region, the null hypothesis is rejected.



In different cases, different statistics and tests are used to find a boundary for a critical region.

For coffee example since we interested in average difference we can calculate a z-score for a found statistics and make a conclusion whether it belongs to a critical region or not.

We will look at another statistics called chi-square that allows us to test hypotheses about the distributions of categorical variables.

Summary

1. We discussed the basic steps for hypothesis testing.

2. We formalized the rules of how do we know whether we should accept or reject a hypothesis.

3. Next time, we will explore a Chi-Square criteria that we can use to test some hypothesis about the distributions.

Chi-Square Criteria

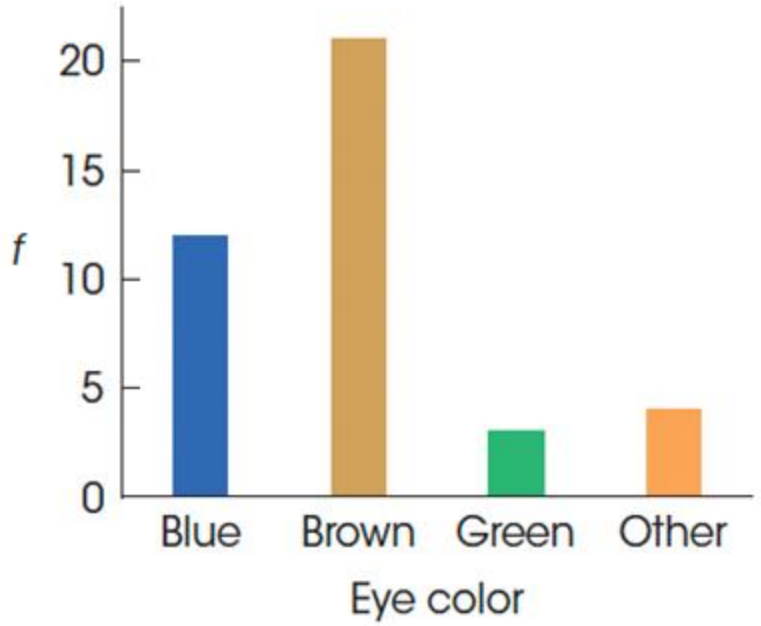
Chi-square criteria

Parameters such as the mean and the standard deviation are the most common way to describe a population, but there are situations in which a researcher has questions about the proportions or relative frequencies for a distribution. For example:

- How does the number of women lawyers compare with the number of men in the profession?
- Of the two leading brands of cola, which is preferred by most Japanese?

Note that each of the preceding examples asks a question about proportions in the population. The individuals are simply classified into categories and we want to know what proportion of the population is in each category.

Example of data for a Chi-square test



Eye color (X)	f
Blue	12
Brown	21
Green	3
Other	4

Blue	Brown	Green	Other
12	21	3	4

Chi-square test for Independence

The chi-square statistic may also be used to test whether there is a relationship between two variables.

The chi-square test for independence uses the frequency data from a sample to evaluate the relationship between two variables in the population. Each individual in the sample is classified on both of the two variables, creating a two-dimensional frequency distribution matrix. The frequency distribution for the sample is then used to test hypotheses about the corresponding frequency distribution in the population.

E.g. is the color preference independent of a personality type?

	Red	Yellow	Green	Blue	
Introvert	10	3	15	22	50
Extrovert	90	17	25	18	150
	100	20	40	40	$n = 200$

The Null Hypothesis for the Independence Test

The goal of the chi-square test is to evaluate the relationship between the two variables. For the example is to determine whether there is a consistent, predictable relationship between personality and color preference.

The null hypothesis states that there is no relationship. The alternative hypothesis, H_1, states that there is a relationship between the two variables.

H_0: For the general population of students, there is no relationship between color preference and personality.

or in other words:

H_0: The frequency distribution of color preference has the same shape (same proportions) for both categories of personality.

Q. When do we say that variables are independent?

Answer:

> **Two variables are independent** when there is no consistent, predictable relationship between them. In this case, the frequency distribution for one variable is not related to (or dependent on) the categories of the second variable.
>
> **As a result, when two variables are independent, the frequency distribution for one variable will have the same shape (same proportions) for all categories of the second variable.**

Degree of Freedom

Although the typical chi-square distribution is positively skewed, there is one other factor that plays a role in the exact shape of the chi-square distribution — **the number of categories**.

Recall that the chi-square formula requires that you add values from every category. The more categories you have, the more likely it is that you will obtain a large sum for the chi-square value.

Technically, each specific chi-square distribution is identified by degrees of freedom (**df**) rather than the number of categories.

Chi-square test for Independence and Degrees of Freedom:

the degrees of freedom for the chi-square test of independence are given by the formula

$$df = (R - 1)(C - 1)$$

where R — number of rows, and C — number of columns

Observed and Expected Frequencies for the Independence Test

Because the test for independence considers two variables, every individual is classified on both variables, and the resulting frequency distribution is presented as a two-dimensional matrix. The frequencies in the sample distribution are called observed frequencies and are identified by the symbol f_o.

The next step is to find the expected frequencies, or f_e values, for this chi-square test. The expected frequencies define an ideal hypothetical distribution that is in perfect agreement with the null hypothesis.

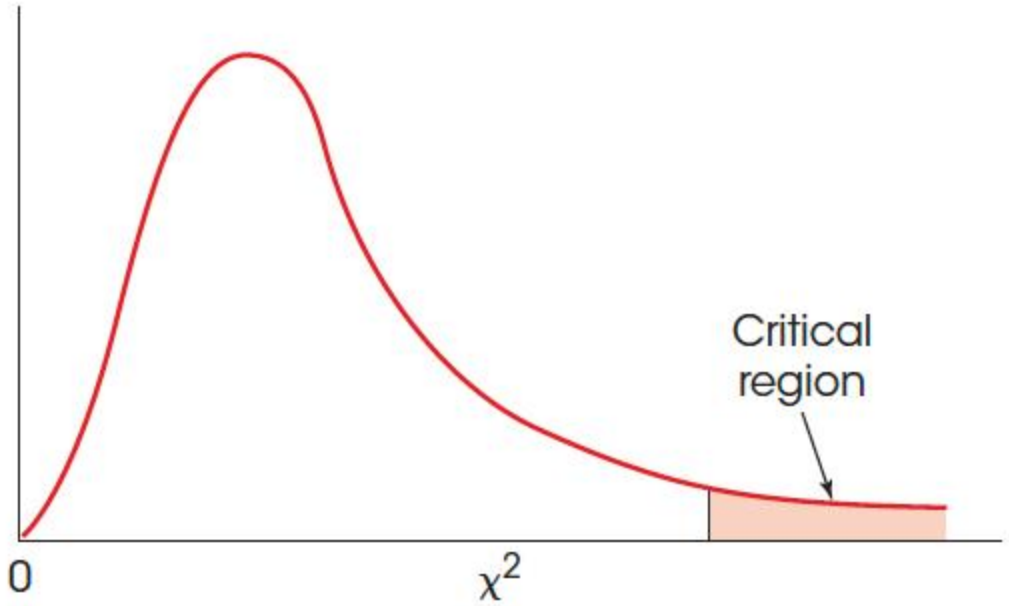
Once the expected frequencies are obtained, we compute a chi-square statistic to determine how well the data (observed frequencies) fit the null hypothesis (expected frequencies).

Chi-square test for Independence formula:

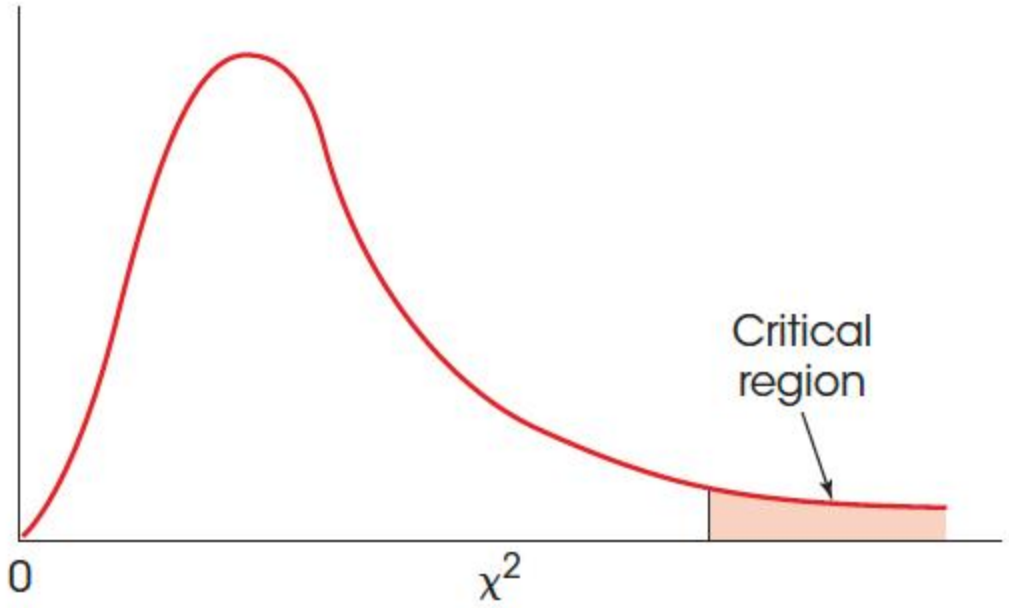
The chi-square test of independence uses the chi-square formula

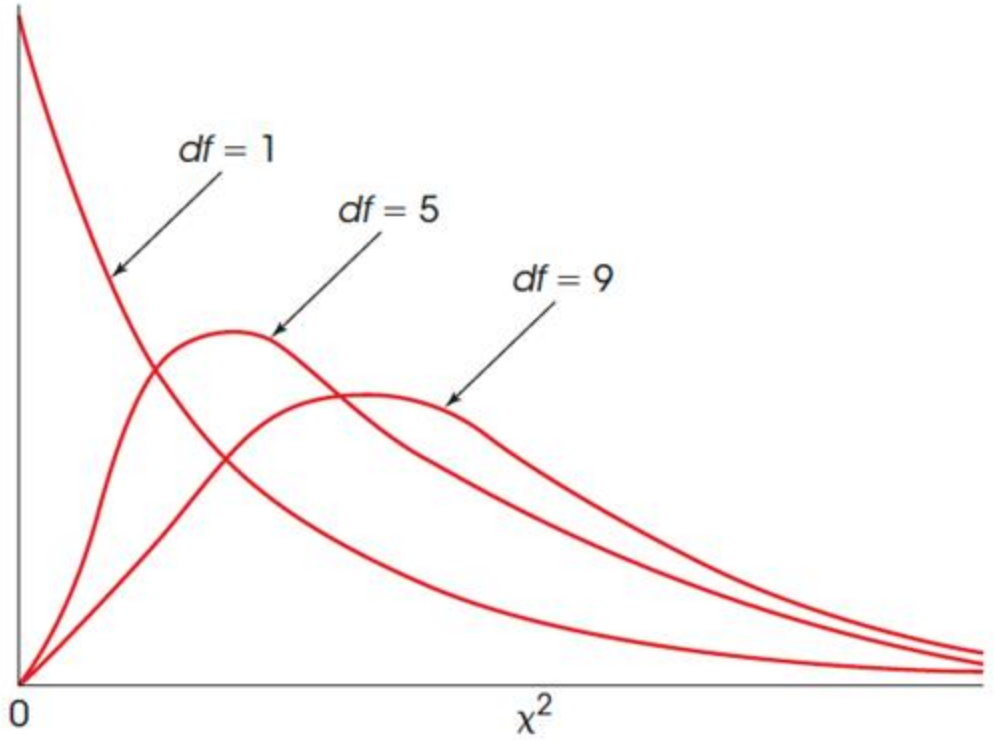
$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

Chi-square distribution



Chi-square distribution and df





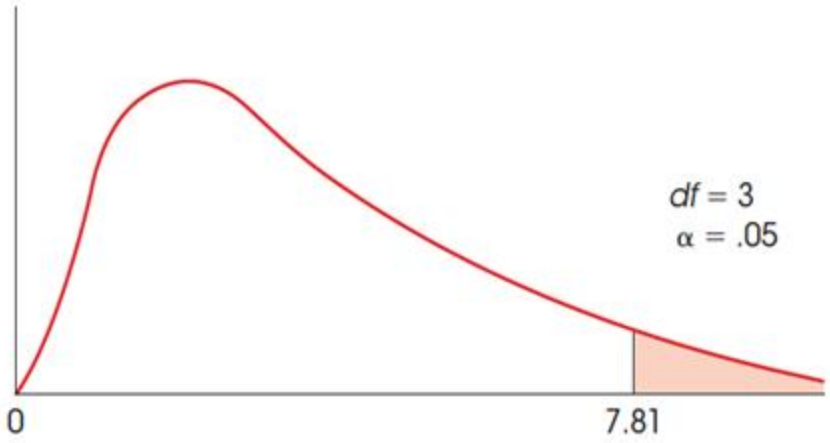
Chi-square statistics and alpha-level

	Proportion in Critical Region				
df	0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59

To determine whether a particular chi-square value is significantly large, you must consult The Chi-Square Distribution. A portion of the chi-square table is shown on the left.

The first column lists **df** values for the chi-square test, and the top row of the table lists proportions (alpha levels) in the extreme right-hand tail of the distribution. The numbers in the body of the table are the critical values of chi-square.

Chi-square statistics and alpha-level



df	Proportion in Critical Region				
	0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84

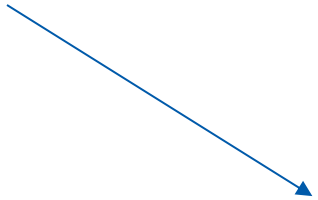
Chi-Square test: step-by-step:

1. State the hypotheses and alpha level.
2. Calculate degrees of freedom and locate a critical region.
3. Find observed and expected frequencies and calculate a chi-square statistic.
4. Is found statistic located in critical region? State a decision whether your null hypothesis should be rejected or not, and make a conclusion.

Finding expected frequencies

Step 1: Calculate rows and columns totals

	Red	Yellow	Green	Blue
Introvert	10	3	15	22
Extrovert	90	17	25	18



	Red	Yellow	Green	Blue	
Introvert					50
Extrovert					150
	100	20	40	40	

Finding expected frequencies

Step 2: Calculate expected preference proportion for each column

	Red	Yellow	Green	Blue	
Introvert	10	3	15	22	50
Extrovert	90	17	25	18	150
	100	20	40	40	$n = 200$

100 out of 200 = 50% prefer red

20 out of 200 = 10% prefer yellow

40 out of 200 = 20% prefer green

40 out of 200 = 20% prefer blue

Finding expected frequencies

Step 3: Calculate expected frequencies for both rows (example below is giving for introverts)

	Red	Yellow	Green	Blue	
Introvert	10	3	15	22	50
Extrovert	90	17	25	18	150
	100	20	40	40	$n = 200$

50% prefer red: $f_e = 50\%$ of $50 = 0.50(50) = 25$

10% prefer yellow: $f_e = 10\%$ of $50 = 0.10(50) = 5$

20% prefer green: $f_e = 20\%$ of $50 = 0.20(50) = 10$

20% prefer blue: $f_e = 20\%$ of $50 = 0.20(50) = 10$

Finding expected frequencies

Step 4: Fill the table with expected frequencies:

	Red	Yellow	Green	Blue	
Introvert	10	3	15	22	50
Extrovert	90	17	25	18	150
	100	20	40	40	$n = 200$

Observed frequencies

Expected frequencies consistent with the null hypothesis. Below is a simple formula for calculation f_e for each cell:

	Red	Yellow	Green	Blue	
Introvert	25	5	10	10	50
Extrovert	75	15	30	30	150
	100	20	40	40	

$$f_e = \frac{f_c f_r}{n}$$

Calculating chi-square statistics

As the formula indicates, the value of chi-square is computed by the following steps.

1. Find the difference between f_o (the data) and f_e (the hypothesis) for each category.

2. Square the difference. This ensures that all values are positive.

3. Next, divide the squared difference by f_e.

4. Finally, sum the values from all the categories.

$$\text{chi-square} = \chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

Interpreting the chi-square statistics

As usual, the sample data are not expected to provide a perfectly accurate representation of the population. In this case, the proportions or observed frequencies in the sample are not expected to be exactly equal to the proportions in the population.

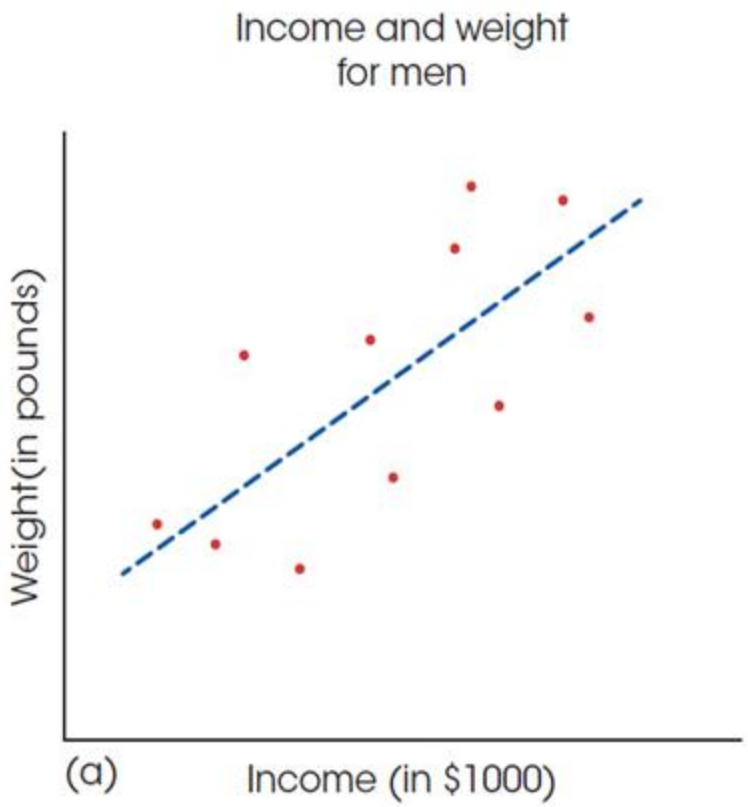
Thus, if there are small discrepancies between the f_o and f_e values, we obtain a small value for chi-square and we conclude that there is a good fit between the data and the hypothesis (fail to reject H_0).

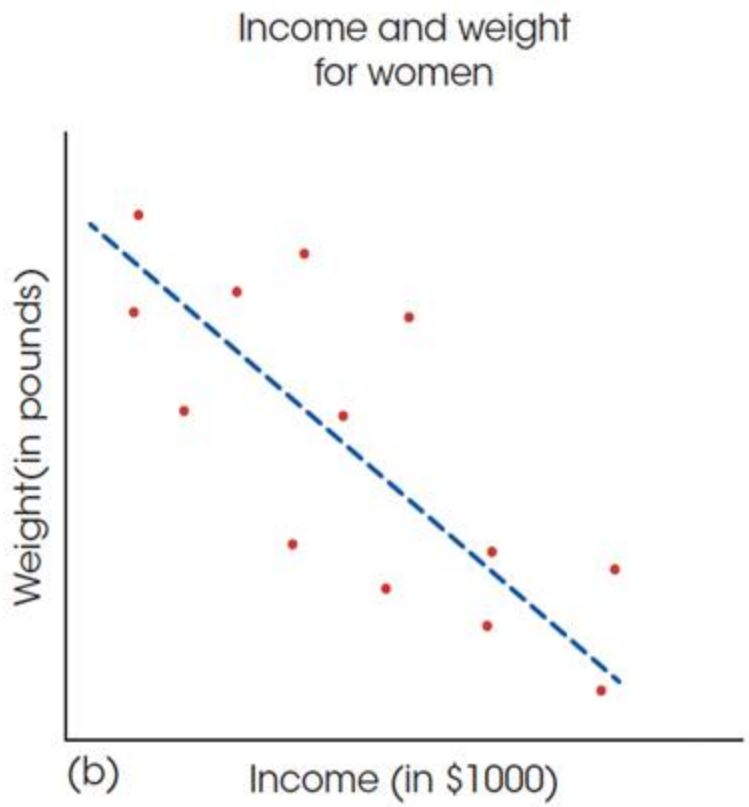
However, when there are large discrepancies between f_o and f_e, we obtain a large value for chi-square and conclude that the data do not fit the hypothesis (reject H_0).

Correlation

How does a relationship look in data?

FIGURE 15.1
Examples of positive and negative relationships. (a) Income is positively related to weight for men. (b) Income is negatively related to weight for women.





What is correlation?

Correlation is a descriptive statistic. Just as a sample mean provides a concise description of an entire sample, a correlation provides a description of a relationship.

Correlation is a statistical technique that is used to measure and describe the relationship between two numerical variables.

A correlation requires **two scores** for each individual (one score from each of the two variables). These scores normally are identified as X and Y. The pairs of scores can be listed in a table, or they can be presented graphically in a scatter plot.

Data to compute a correlation:

Person	Family Income (in $1000)	Student's Average Grade
A	31	72
B	38	86
C	42	81
D	44	78
E	49	85
F	56	80
G	58	91
H	65	89
I	70	94
J	90	83
K	92	90
L	106	97
M	135	89
N	174	95

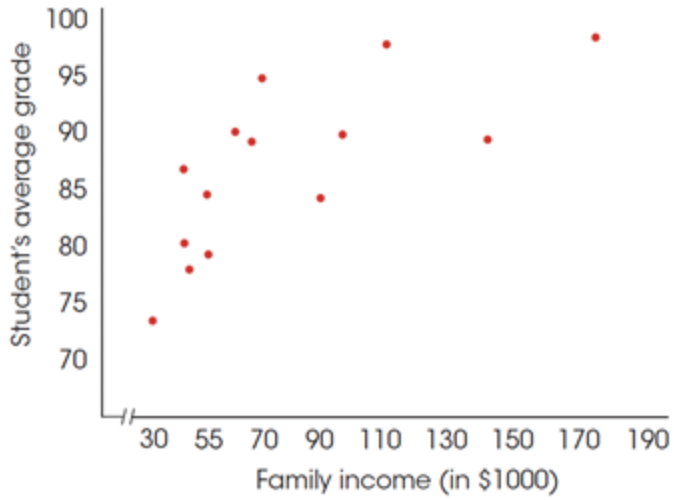


FIGURE 15.2
Correlational data showing the relationship between family income (X) and student grades (Y) for a sample of $n = 14$ high school students. The scores are listed in order from lowest to highest family income and are shown in a scatter plot.

Correlation and a relationship characteristics

1. **The Direction of the Relationship**

   The sign of the correlation, positive or negative, describes the direction of the relationship.

2. **The Form of the Relationship**

   The most common use of correlation is to measure straight-line relationships. However, other forms of relationships do exist and there are special correlations used to measure them. But we won't talk about those.

3. **The Strength or Consistency of the Relationship**

   For a linear relationship, for example, the data points could fit perfectly on a straight line. Every time X increases by one point, the value of Y also changes by a consistent and predictable amount. However, relationships are usually not perfect.

Correlation and direction of the relationship:

In a **positive correlation**, the two variables tend to change in the same direction: as the value of the X variable increases from one individual to another, the Y variable also tends to increase; when the X variable decreases, the Y variable also decreases.

In a **negative correlation**, the two variables tend to go in opposite directions. As the X variable increases, the Y variable decreases. That is, it is an inverse relationship.

Correlation and the form of a relationship

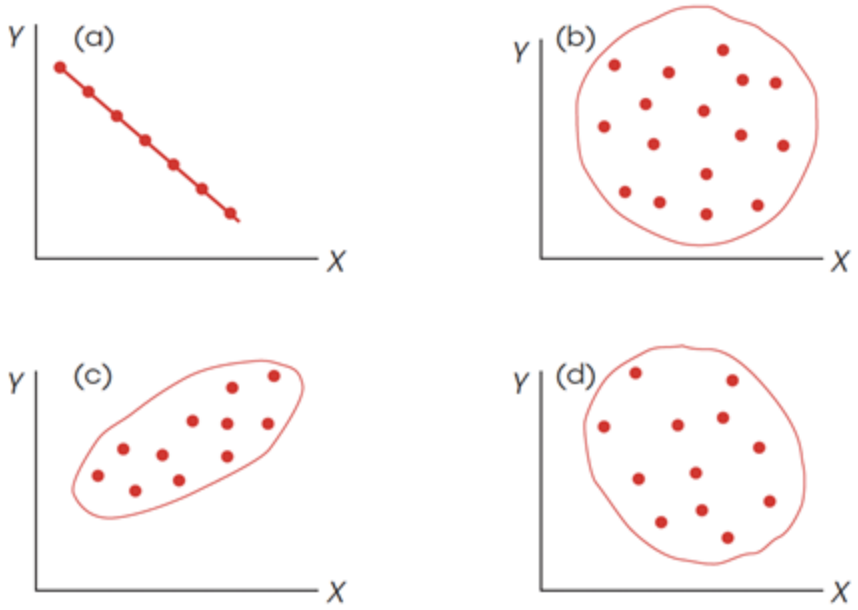
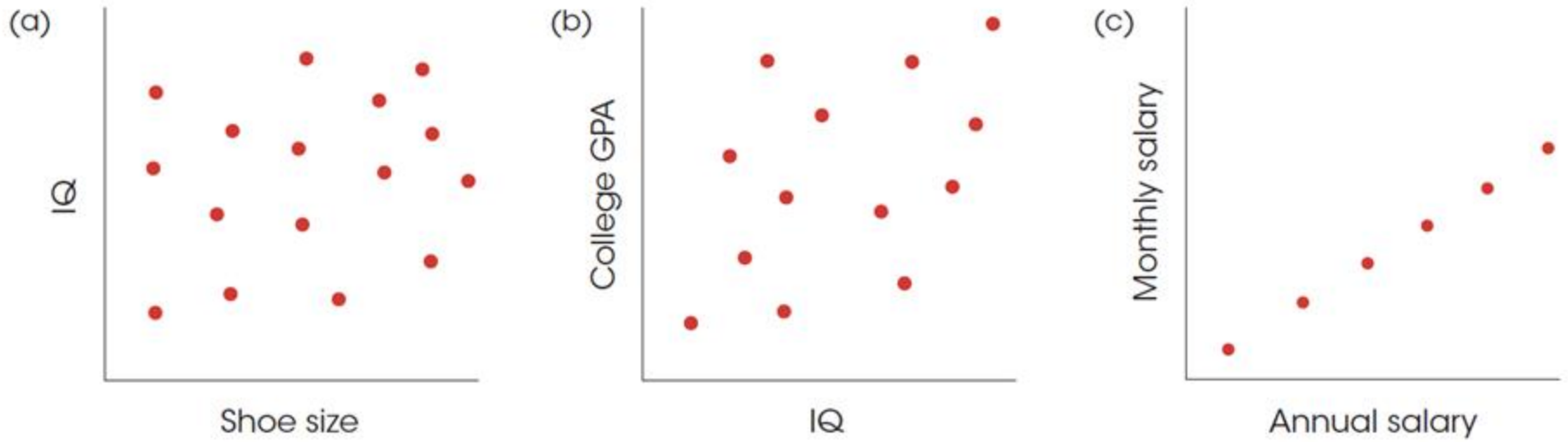


FIGURE 15.3
Examples of different values for linear correlations: (a) a perfect negative correlation, −1.00, (b) no linear trend, 0.00, (c) a strong positive relationship, approximately +.90, and (d) a relatively weak negative correlation, approximately −0.40.



Correlation and consistency of the relationship

The consistency of the relationship is measured by the numerical value of the correlation. A **perfect correlation** always is identified by a correlation of **1.00** and indicates a perfectly consistent relationship. For a correlation of 1.00 (or –1.00), each change in X is accompanied by a perfectly predictable change in Y.

At the other extreme, a correlation of **0** indicates no consistency at all. For a correlation of 0, the data points are scattered randomly with no clear trend.

Intermediate values between 0 and 1 indicate the degree of consistency.

Correlations coefficients

Coefficient	Interpretation
from 0 to 0.3	very weak / small — not paying attention
from 0.3 to 0.5	weak / small — not paying attention
from 0.5 to 0.7	moderate — some attention
from 0.7 to 0.9	high / large — something interesting is going on
from 0.9 to 1	very high / very large — perfect relationship

For negative correlation coefficients, it is all the same but with the minus sign.

Also, note that those interpretations are conventions. There could be some discrepancies in different areas of research.

Learning check

1. A negative value for a correlation indicates _____.

a. a much stronger relationship than if the correlation were positive

b. a much weaker relationship than if the correlation were positive

c. increases in X tend to be accompanied by increases in Y

d. increases in X tend to be accompanied by decreases in Y

2. Which of the following is the correct order, from strongest and most consistent to weakest and least consistent, for the following correlations?

a. −1.00, 0.85, −0.43, 0.02

b. 0.85, 0.02, −0.43, −1.00

c. −1.00, −0.43, 0.02. 0.85

d. 0.85, −0.43, 0.02, −1.00

The Pearson correlation & formula

The **Pearson correlation** measures the degree and the direction of the linear relationship between two variables.

$$r = \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}}$$

$$= \frac{\text{covariability of } X \text{ and } Y}{\text{variability of } X \text{ and } Y \text{ separately}}$$

Formula:

$$r = \frac{SP}{\sqrt{(SS_X)(SS_Y)}}$$

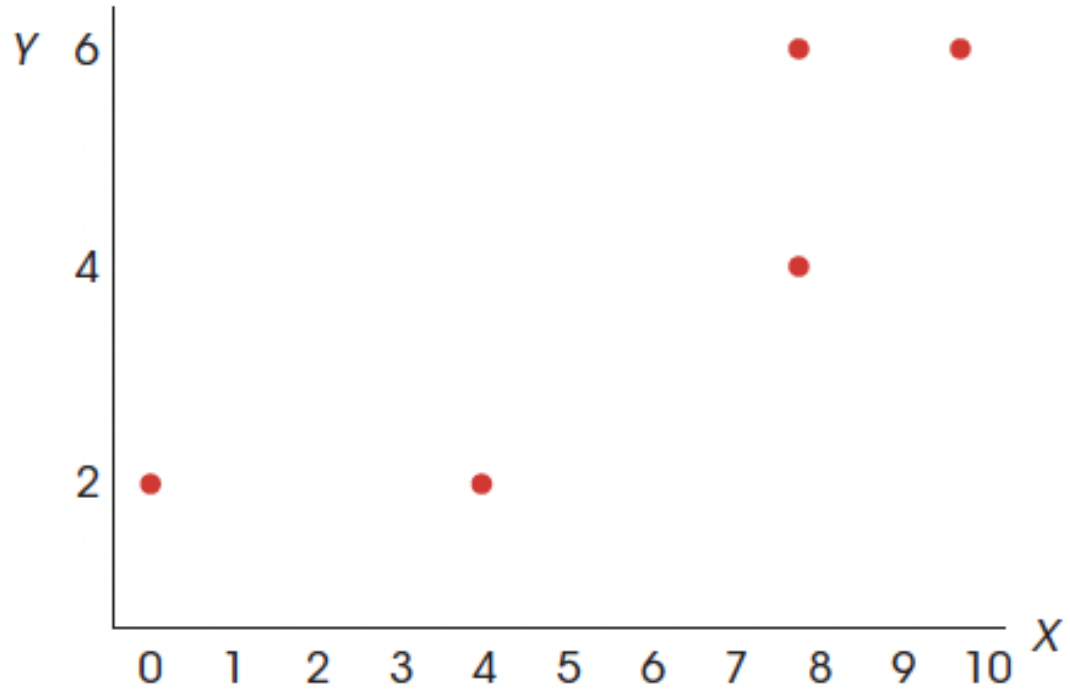
where SP is the sum of products of deviations and SS is the sum of squares of deviations.

$$SP = \Sigma(X - M_X)(Y - M_Y)$$

$$SS = \Sigma(X - M)^2$$

Calculating the Pearson correlation

Scores	
X	Y
0	2
10	6
4	2
8	4
8	6



Calculating the Pearson correlation

Scores		Deviations		Squared Deviations		Products
X	Y	$X - M_X$	$Y - M_Y$	$(X - M_X)^2$	$(Y - M_Y)^2$	$(X - M_X)(Y - M_Y)$
0	2	−6	−2	36	4	+12
10	6	+4	+2	16	4	+8
4	2	−2	−2	4	4	+4
8	4	+2	0	4	0	0
8	6	+2	+2	4	4	+4
				$SS_X = 64$	$SS_Y = 16$	$SP = +28$

$$r = \frac{SP}{\sqrt{(SS_X)(SS_Y)}} = \frac{28}{\sqrt{(64)(16)}} = \frac{28}{32} = +0.875$$

Learning check

1. What is the sum of products (SP) for the following data?

a. 6

b. −5

c. 43

d. None of the other 3 choices is correct.

X	Y
2	4
5	2
3	5
2	5

2. A set of $n = 5$ pairs of X and Y values has $SS_X = 5$, $SS_Y = 20$ and $SP = 8$. For these data, the Pearson correlation is _____.

a. $r = \frac{8}{100} = 0.08$

b. $r = \frac{8}{10} = 0.80$

c. $r = \frac{8}{25} = 0.32$

d. $r = \frac{8}{20} = 0.40$

When do we use correlation?

Prediction

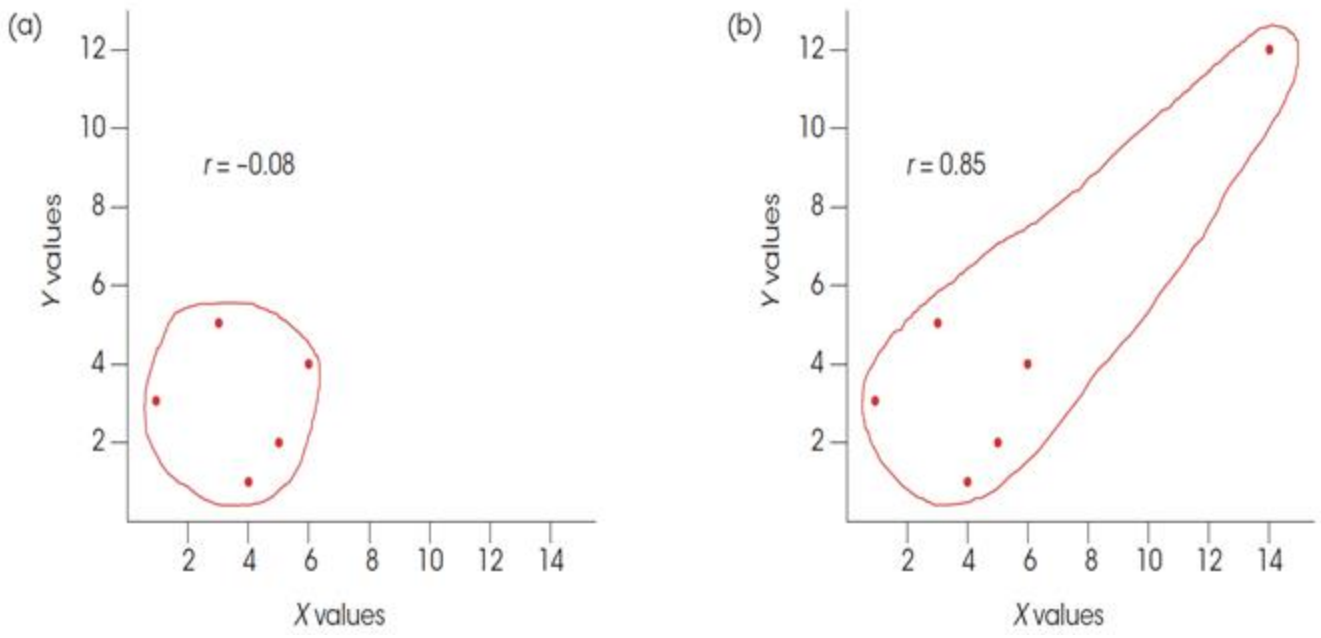
If two variables are known to be related in some systematic way, it is possible to use one of the variables to make accurate predictions about the other.

For example, when you applied for college admission, you were required to submit a great deal of personal information, including your exam scores. College officials want this information so they can predict your chances of success in college. It has been demonstrated over several years that school exam scores and college grade point averages are correlated. Students who do well on the SAT tend to do well in college.

Interpreting correlation: (Cautions!)

1. Correlation simply describes a relationship between two variables. It does not explain why the two variables are related. Specifically, a correlation should not and **cannot be interpreted as proof of a cause-and-effect relationship between the two variables.**
2. The value of a correlation **can be affected** greatly by the **range of scores** represented in the data.
3. One or two extreme data points, often called **outliers**, can have a **dramatic effect** on the value of a correlation.

Correlation and outliers

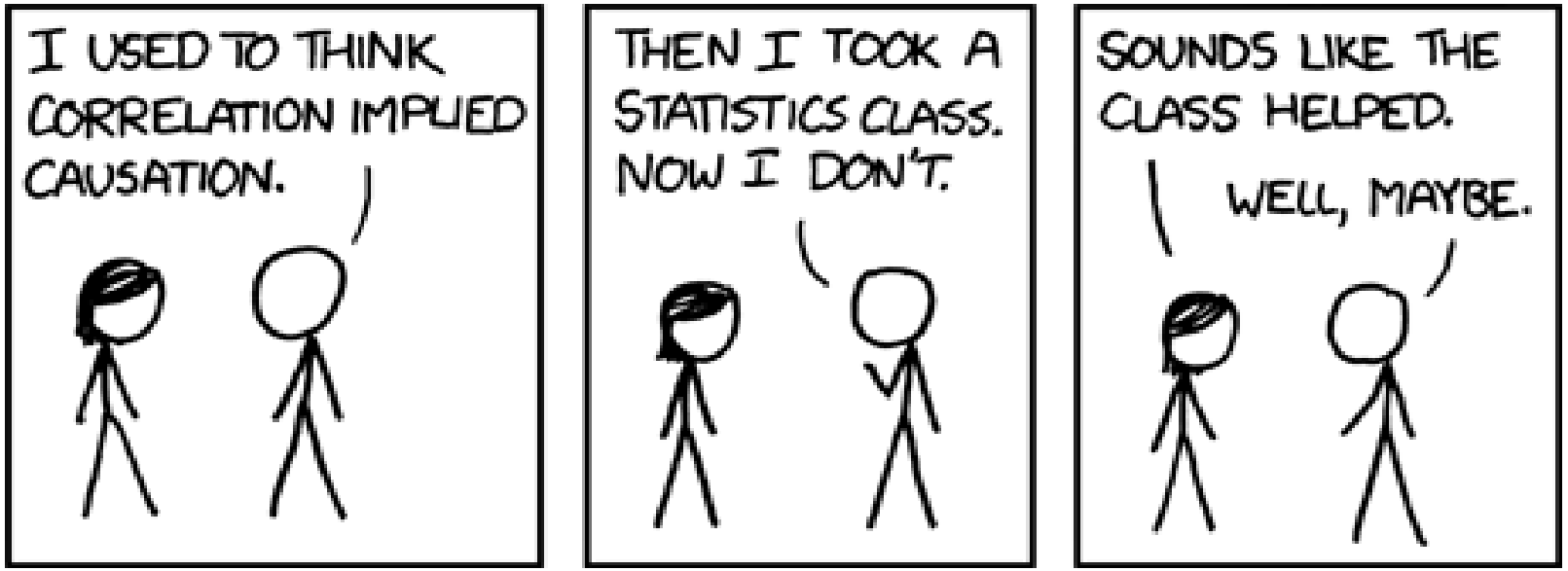


…Interpreting correlation: (Cautions!)

4. When judging how "good" a relationship is, it is tempting to focus on the numerical value of the correlation.

However, a correlation should not be interpreted as a proportion. Although a correlation of 1.00 does mean that there is a 100% perfectly predictable relationship between X and Y, a correlation of .5 does not mean that you can make predictions with 50% accuracy. To describe how accurately one variable predicts the other, you must square the correlation. Thus, a correlation of r = .5 means that one variable partially predicts the other, but the predictable portion is only r2 = .52 = 0.25 (or 25%) of the total variability.

Correlation does not imply causation



Although there may be a causal relationship, the simple existence of a correlation does not prove it.

To establish a cause-and-effect relationship, it is necessary to conduct a true experiment in which one variable is manipulated by a researcher and other variables are rigorously controlled. Or to make a rigorous study with some theoretical background.

Thank you!

Majid Sohrabi

msohrabi@hse.ru

@MSOHRABI_CS